Journal of Machine Learning Research 24 (2023) 1-97 Submitted 7/21; Revised 4/23; Published 4/23

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

Hui Jin HUIJIN@UCLA.EDU
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095, USA

Guido Montúfar MONTUFAR@MATH.UCLA.EDU
Department of Mathematics and Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095, USA; and
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Editor: Christoph Lampert

Abstract

We investigate gradient descent training of wide neural networks and the corresponding 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.

Keywords: implicit bias, overparametrized neural network, cubic spline interpolation, smoothing spline, effective capacity.

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 (Zhang et al., 2017). A series of works have observed that this phenomenon must involve some form of capacity control beyond the network size (Neyshabur et al., 2015) and, specifically, an implicit bias resulting from the parameter optimization procedures (Neyshabur et al., 2017). 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 work 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, Oymak and Soltanolkotabi (2019) 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, Zhang et al. (2020) showed that gradient flow optimization converges to the global minimum which is closest to the initialization in parameter space. Although neural networks have a non-linear parametrization, Jacot et al. (2018), Lee et al. (2019) and Lai et al. (2023) 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, Chizat et al. (2019) 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. Savarese et al. (2019) 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. Ongie et al. (2020) 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 (Keskar et al., 2017; Wu et al., 2017; Dinh et al., 2017). Gunasekar et al. (2018a) 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. Soudry et al. (2018) showed that in classification problems with separable data, gradient descent with linear networks converges to a max-margin solution. Gunasekar et al. (2018b) presented a result on implicit bias for deep linear convolutional networks, and Ji and Telgarsky (2019) studied non-separable data. Chizat and Bach (2020) 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, Maennel et al. (2018) 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. Baratin et al. (2021) showed the evolution of the tangent features during training which can be interpreted as feature selection and compression. Williams

et al. (2019) 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 article is organized as follows. In Section 2 we provide settings and notation. We present our main results in Section 3, along with a discussion. The main techniques pertaining wide networks and the infinite width limit are presented in Sections 4 and 5. In Sections 6 and 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. Notation 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

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

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:

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

In the analysis of Jacot et al. (2018); Lee et al. (2019), $\mathcal{W}$ and $\mathcal{B}$ are Gaussian $\mathcal{N}(0, \sigma^2)$. In the default initialization of PyTorch (Paszke et al., 2019), $\mathcal{W}$ 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:

$$\begin{aligned} (\mathbf{W}_i^{(1)}, b_i^{(1)}) &\overset{d}{=} (\boldsymbol{\mathcal{W}}, \mathcal{B}) \\ W_i^{(2)} &\overset{d}{=} \sqrt{1/n}\, \mathcal{W}^{(2)}, \quad b^{(2)} \overset{d}{=} \sqrt{1/n}\, \mathcal{B}^{(2)}. \end{aligned} \tag{3}$$

The setting (1) is known as the standard parametrization. Some works (Jacot et al., 2018; Lee et al., 2019) use the so-called NTK parametrization, where the factor $\sqrt{1/n}$ is carried outside of the trainable parameter (for details see Appendix 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 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

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta) = \theta_t - \eta \nabla_\theta f(\mathcal{X}, \theta_t)^T \nabla_{f(\mathcal{X},\theta_t)} L, \tag{4}$$

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 (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]

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.

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 feedforward network with a single input unit, a hidden layer of n rectified linear units, and a single linear output unit. Assume standard parametrization* (1) *and parameter initialization* (3)*, which means for each hidden unit the input weight and bias are initialized from a sub-Gaussian $(\mathcal{W}, \mathcal{B})$ with continuous 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 $S = \operatorname{supp}(\zeta) \cap [\min_j x_j, \max_j x_j]$, 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*

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 $\mathbb{P}(|X_n/a_n| > M_\epsilon) < \epsilon, \forall n > N_\epsilon$.

following variational problem.[2]

$$\begin{aligned} \min_{g \in C^2(S)} \quad & \int_S \frac{1}{\zeta(x)} (g''(x) - f''(x, \theta_0))^2 \, \mathrm{d}x \\ \text{subject to} \quad & g(x_j) = y_j - u x_j - v, \quad j = 1, \ldots, M. \end{aligned} \tag{5}$$

The proof is provided in Appendix C. Our main theorem also holds when the network parameters are trained by stochastic gradient descent. We provide details in Theorem 24 and Remark 25 in Appendix D. In Appendix L we also present a corresponding result for networks with skip connections, which does not need a linear adjustment of the data. We will give an interpretation of the result in Section 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 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* (5) *in Theorem 1 corresponds to cubic spline interpolation with natural boundary conditions (see, e.g., Ahlberg et al., 1967). 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 O for details).*

For different activation functions, we have the following corollary, proved in Appendix 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* (5) *changed to*

$$\begin{aligned} \min_{g \in C^2(S)} \quad & \int_S \frac{1}{\zeta(x)} [\mathrm{L}(g(x) - f(x, \theta_0))]^2 \, \mathrm{d}x \\ \text{subject to} \quad & g(x_j) = y_j - p(x_j), \quad j = 1, \ldots, M, \end{aligned} \tag{6}$$

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

2. The existence of the minimum of the variational problem is not obvious. We prove that the minimum exists and the solution of the variational problem is g^*.

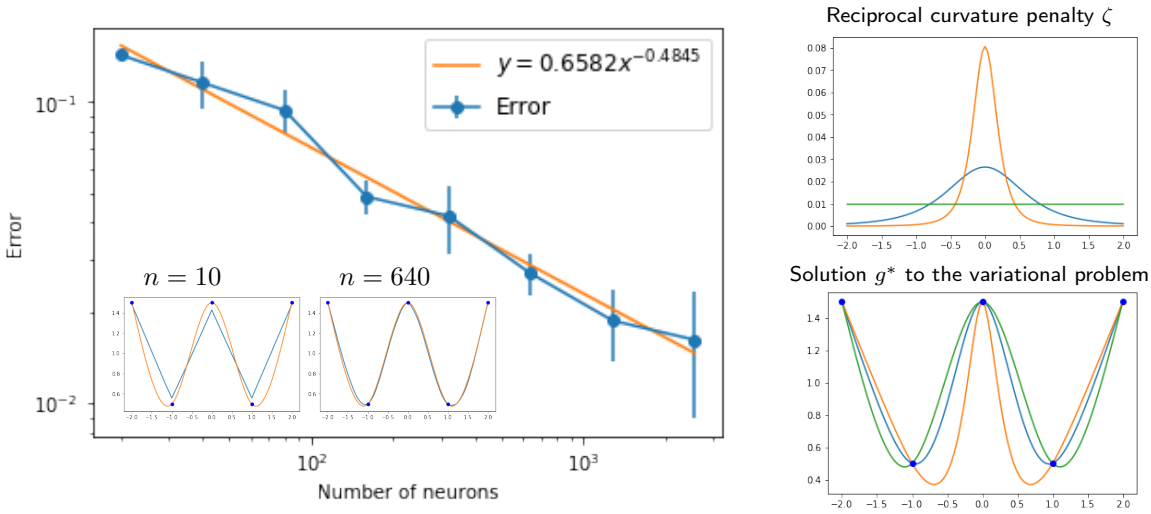

Figure 1: Illustration of Theorem 1. Left: Uniform error between the solution g^* to the variational problem and the functions $f(\cdot, \theta^*)$ obtained by gradient descent training with uniform initialization $\mathcal{W} \sim \text{Unif}(-1, 1)$, $\mathcal{B} \sim \text{Unif}(-2, 2)$, against the 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.

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 (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 Abramovich and Steinberg, 1996; Pintore et al., 2006). This problem can be solved explicitly and can be shown to approximate early stopping. In Appendix 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:*

$$\min_{g \in C^2(S)} \quad \sum_{j=1}^{M} [g(x_j) - y_j]^2 + \frac{1}{\bar{\eta} t} \int_S \frac{1}{\zeta(x)} (g''(x) - f''(x, \theta_0))^2 \, \mathrm{d}x. \tag{7}$$

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 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 continuous 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:*

$$\begin{aligned} \min_{g\in \mathrm{Lip}(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0))\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)} \, \mathrm{d}\boldsymbol{V}\mathrm{d}c \\ \text{subject to} \quad & g(\mathbf{x}_j) = y_j - \langle \mathbf{u}, \mathbf{x}_j\rangle - v, \quad j = 1, \dots, M \\ & \mathcal{R}\{(-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0))\}(\boldsymbol{V}, c) = 0, \quad (\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta) \\ & (-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0)) \in L^p(\mathbb{R}^d), \ 1 \le p < d/(d-1). \end{aligned} \tag{8}$$

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 C. In Appendix L we also present a corresponding result for networks with skip connections, which does not need a linear adjustment of the data. 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 (8). We prove the following theorem in Appendix 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 $a_b \ge \max_i \|\mathbf{x}_i\|_2$. Then the variational problem (8) is equivalent to*

$$\begin{aligned} \min_{h\in \mathrm{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\mathbb{R}^d} \left((-\Delta)^{(d+3)/4}(h(\mathbf{x}) - f(\mathbf{x}, \theta_0))\right)^2 \, \mathrm{d}\mathbf{x} \\ \text{subject to} \quad & h(\mathbf{x}_j) = y_j, \quad j = 1, \dots, M \\ & (-\Delta)^{(d+1)/2}(h(\cdot) - f(\cdot, \theta_0)) \in L^p(\mathbb{R}^d), \ 1 \le p < d/(d-1). \end{aligned} \tag{9}$$

We can solve the simplified variational problem (9) explicitly. We prove the following theorem in Appendix I.3.

Theorem 8 (Closed form solution) *Suppose $h(\mathbf{x})$ solves the variational problem (9). Then $h(\mathbf{x})$ is given by*

$$h(\mathbf{x}) - f(\mathbf{x}, \theta_0) = \sum_{j=1}^M \lambda_j \|\mathbf{x} - \mathbf{x}_j\|^3 + \langle \mathbf{u}, \mathbf{x}_i\rangle + v, \tag{10}$$

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

$$\begin{cases} \sum_{j=1}^{M} \lambda_j \|\mathbf{x}_i - \mathbf{x}_j\|^3 + \langle \mathbf{u}, \mathbf{x}_i \rangle + v = y_i - f(\mathbf{x}_i, \theta_0), \quad i = 1, \dots, M \\ \sum_{j=1}^{M} \lambda_j = 0 \\ \sum_{j=1}^{M} \lambda_j \mathbf{x}_j = \mathbf{0}\,. \end{cases} \tag{11}$$

Remark 9 *A function of the form* (10)–(11) *is referred to as a polyharmonic spline (see Potter, 1981), which is a special type of radial basis function interpolation (Du Toit, 2008). 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 (Segeth, 2019) with decreasing regularization parameter, similar to the description* (7) *for the univariate case.*

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 1 and more extensively in Appendix A.[3] 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 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

3. The code of our experiments and the plots can be found in our GitHub repository: `https://github.com/huijin12/Implicit_Bias_Wide_Neural_Networks`

distributions, Theorem 8 shows that the network output after training is a polyharmonic spline. We illustrate Theorem 6 numerically in Figure 2 and more extensively in Appendix 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 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 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 $y_j = \langle \mathbf{u}, \mathbf{x}_j \rangle + v + \epsilon_j$, $j = 1, \dots, M$, and set the adjusted data as $\{(\mathbf{x}_j, \epsilon_j)\}_{j=1}^M$. Furthermore, if we consider a network architecture with skip connections from the inputs to the outputs, our result holds for the original training data without any adjustments. We present the details to this result in Appendix L.

Generalization results Theorem 1 allows us to show how gradient descent on wide neural networks learns a target function. In the following paragraphs, we show how the solution to the variational problems (5), (7) and (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 (5) is the cubic spline interpolation of the training data. Hall and Meyer (1976) 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 (7). If ζ is constant on $[a, b]$, the solution g^* of (7) is the cubic smoothing spline of the training data. Ragozin (1983, 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 $\mathbb{E}\|g^* - g_0\|_2^2 \leq C \left((1/t + (1/M)^4) \|g_0''\|^2 + t^{1/4}/M \right)$, 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. Ragozin (1983) also gives an error bound for g^* in the case of non-uniform training inputs. Eggermont and LaRiccia (2006) 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 (7) is called a spatially adaptive smoothing spline of the training data. Wang et al. (2013, 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 $q \in C^3([a,b])$, then $|g^*(x) - g_0(x)| = O_p(M^{-4/9})$. 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. Wang et al. (2013, Corollary 1) showed that the 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 a suitable ζ for the training data. Abramovich and Steinberg (1996) and Storlie et al. (2010) proposed to choose ζ based on an estimation of the second derivative of g_0. Liu and Guo (2010) used a piecewise constant ζ and proposed a search algorithm to find such ζ. Proposition 15 shows a way to choose the joint distribution of weight and bias parameters in order to have that ζ is proportional to a given function. Once we find an appropriate ζ according to the training data using the methods in the 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, this parameter initialization should perform better than uniform or Gaussian initialization.

For multivariate noiseless models, if ζ is constant over its support, then the solution g^* of variational problem (8) is the polyharmonic spline. For this setting, Potter (1981, Theorem 3.2) gave an error bound between g^* and the target function g_0.

Strategy of the proof In Section 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 Zhang et al., 2020). 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 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. A similar result has been previously obtained by Daniely (2017). Under these settings, the implicit bias of gradient descent amounts to minimizing the distance from the initial parameter, subject to fitting the training data. In Section 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 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 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 (Lai et al., 2023, Proposition 3.2) bounding the difference in function space of the solutions obtained from training a wide network and its linearization to conclude the proof.

Related works Zhang et al. (2020) 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 M. Cao and Gu (2019) derived 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.

Savarese et al. (2019) 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 B.4 for more details). A recent work by Parhi and Nowak (2019) 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 (Heiss et al., 2019) 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 Sahs et al. (2020b). Schmidt-Hieber (2020) showed that a shallow ReLU network with one input and one output node approximately converges to the natural cubic spline interpolant under SGD training. Williams et al. (2019) 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 that 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 description given by Williams et al. (2019) 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, Ongie et al. (2020) 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 Parhi and Nowak (2021) 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.

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.

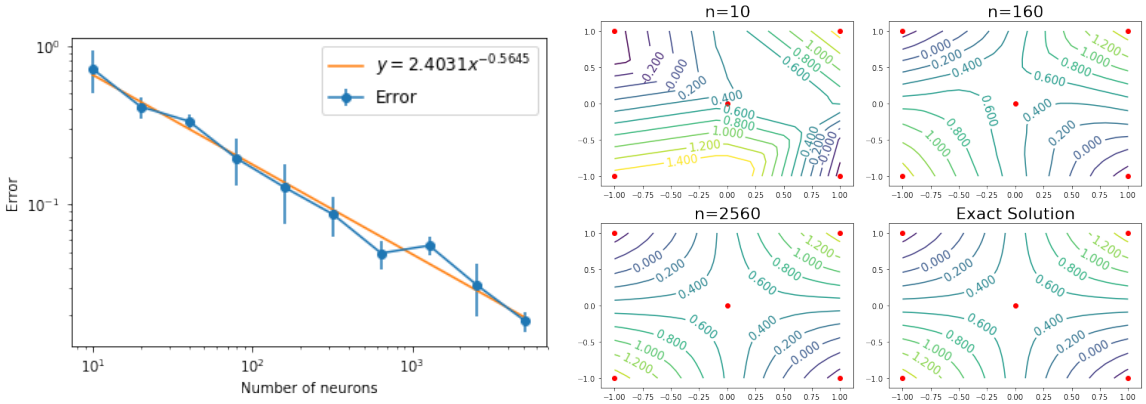

Figure 2: Illustration of Theorem 6. Left: Uniform error between the solution g^* to the variational problem and the functions $f(\cdot, \theta^*)$ obtained by gradient descent training of a neural network (in this case with initialization $\boldsymbol{\mathcal{W}} \sim \mathrm{Unif}(\mathbb{S}^1)$, $\mathcal{B} \sim \mathrm{Unif}(-2, 2)$), 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.

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:

$$f^{\mathrm{lin}}(\mathbf{x}, \omega) = f(\mathbf{x}, \theta_0) + \nabla_\theta f(\mathbf{x}, \theta_0)(\omega - \theta_0). \tag{12}$$

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

$$L^{\mathrm{lin}}(\omega) = \frac{1}{M} \sum_{j=1}^{M} \ell(f^{\mathrm{lin}}(\mathbf{x}_j, \omega), y_j). \tag{13}$$

The gradient descent iteration for the linearized model is given by

$$\omega_0 = \theta_0, \quad \omega_{t+1} = \omega_t - \eta \nabla_\theta f(\mathcal{X}, \theta_0)^T \nabla_{f^{\mathrm{lin}}(\mathcal{X}, \omega_t)} L^{\mathrm{lin}}. \tag{14}$$

Next, we consider wide neural networks. According to Lee et al. (2019, Theorem H.1) and (Lai et al., 2023, Proposition 3.2),

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

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 $f^{\mathrm{lin}}(\mathcal{X}, \omega_\infty) = f(\mathcal{X}, \theta_\infty) = \mathcal{Y}$, 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 (14) of the linearized network converge to the unique global minimum that is closest to initialization (Gunasekar et al., 2018a; Zhang et al., 2020). More precisely, ω_∞ is the solution to following constrained optimization problem (further details are provided in Appendix D):

$$\min_{\omega} \|\omega - \theta_0\|_2 \quad \text{s.t. } f^{\text{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \tag{15}$$

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 (3), given $1 \le i \le n$, we have that $\|\overline{\mathbf{W}}_i^{(1)}\|, \overline{b}_i^{(1)} = O_p(1)$ and $\overline{W}_i^{(2)}, \overline{b}^{(2)} = O_p(n^{-\frac{1}{2}})$.[4] Therefore, writing H for the Heaviside function, we have

$$\begin{aligned}
\nabla_{\mathbf{W}_i^{(1)}, b_i^{(1)}} f(\mathbf{x}, \theta_0) &= \left[\overline{W}_i^{(2)} H(\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}\rangle + \overline{b}_i^{(1)}) \cdot \mathbf{x} \ , \ \overline{W}_i^{(2)} H(\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}\rangle + \overline{b}_i^{(1)})\right] = O_p(n^{-\frac{1}{2}}), \\
\nabla_{W_i^{(2)}, b^{(2)}} f(\mathbf{x}, \theta_0) &= \left[[\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}\rangle + \overline{b}_i^{(1)}]_+ \ , \ 1\right] = O_p(1).
\end{aligned} \tag{16}$$

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, $W_i^{(2)}, i = 1, \ldots, n$, and fix all other parameters, which corresponds to training a smaller linear model. Let $\widetilde{\omega}_t = \text{vec}(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}, \widetilde{\mathbf{W}}_t^{(2)}, \overline{b}^{(2)})$ 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

$$\widetilde{\mathbf{W}}_0^{(2)} = \overline{\mathbf{W}}^{(2)}, \quad \widetilde{\mathbf{W}}_{t+1}^{(2)} = \widetilde{\mathbf{W}}_t^{(2)} - \eta \nabla_{\mathbf{W}^{(2)}} L^{\text{lin}}(\widetilde{\omega}_t). \tag{17}$$

Let $\widetilde{\omega}_\infty = \lim_{t\to\infty} \widetilde{\omega}_t$. By the above discussion, we expect that $f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_\infty)$ will be close to $f^{\text{lin}}(\mathbf{x}, \omega_\infty)$. 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 $\ell(\widehat{y}, y) = \frac{1}{2}\|\widehat{y} - y\|_2^2$ and $\inf_n \lambda_{\min}(\hat{\Theta}_n) > 0$. Let ω_t denote the parameters of the linearized model at time t when we train all parameters using* (14)*, and let $\widetilde{\omega}_t$ denote the parameters at time t when we only train weights of the output layer using* (17)*. If we use the same learning rate η in these two*

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

training processes and $\eta < \frac{2}{n\lambda_{\max}(\hat{\Theta}_n)}$*, then for any compact set* $D \subset \mathbb{R}^d$*, we have*

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

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 E. By combining Theorem 10 and the fact that training a linearized model approximates training a wide network (Lai et al., 2023, Proposition 3.2), 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* $D \subset \mathbb{R}^d$*,* $\sup_{\mathbf{x}\in D} \sup_t \|f^{\text{lin}}(\mathbf{x}, \tilde{\omega}_t) - f(\mathbf{x}, \theta_t)\|_2 = O_p(n^{-\frac{1}{2}})$.

The proof is given in Appendix 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.

5. Infinite Width Limit of Shallow Networks

According to (15), gradient descent training of the output weights (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

$$\min_{\mathbf{W}^{(2)}} \|\mathbf{W}^{(2)} - \overline{\mathbf{W}}^{(2)}\|_2^2 \quad \text{s.t.} \quad \sum_{i=1}^n (W_i^{(2)} - \overline{W}_i^{(2)})[\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}_j\rangle + \overline{b}_i^{(1)}]_+ = y_j - f(\mathbf{x}_j, \theta_0), \ j = 1, \dots, M. \tag{18}$$

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 B.2). The analysis still goes through without this simplification (see Appendix H).

We reformulate problem (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., $\mu_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A\left((\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)})\right)$, where $\mathbb{1}_A$ denotes the indicator function for measurable subsets A in $\mathbb{R}^{d+1}$. We further consider a function $\alpha_n \colon \mathbb{R}^{d+1} \to \mathbb{R}$ 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 (18) with ASI can be rewritten as

$$\min_{\alpha_n \in C(\mathbb{R}^{d+1})} \int_{\mathbb{R}^2} \alpha_n^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) \ \text{ s.t.} \int_{\mathbb{R}^{d+1}} \alpha_n(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j\rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) = y_j, \tag{19}$$

where j ranges from 1 to M. Here we minimize over functions α_n in $C(\mathbb{R}^{d+1})$, 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 (19) as follows:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^{d+1})} \quad & \int_{\mathbb{R}^{d+1}} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^{d+1}} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = y_j, \quad j = 1, \dots, M. \end{aligned} \tag{20}$$

Using that μ_n weakly converges to μ, the following theorem shows that in fact the solution of problem (19) converges to the solution of (20). The proof is given in Appendix 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 (19). Let $\overline{\alpha}(\mathbf{W}^{(1)}, b)$ be the solution of (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 $g_n(\mathbf{x}, \alpha_n) = \int_{\mathbb{R}^{d+1}} \alpha_n(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b)$ is the function represented by a network with n hidden neurons after training, and $g(\mathbf{x}, \alpha) = \int_{\mathbb{R}^{d+1}} \alpha(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)$ is the function represented by the infinite-width network.*

6. Implicit Bias for Univariate Regression

In this section we solve the optimization problem (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 $(\mathcal{W}, \mathcal{C})$.[5] Then, writing $\gamma(W^{(1)}, c) = \alpha(W^{(1)}, -cW^{(1)})$, the optimization problem (20) is equivalently given as

$$\min_{\gamma \in C(\mathbb{R}^2)} \int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c) \, \mathrm{d}\nu(W^{(1)}, c) \;\; \text{s.t.} \int_{\mathbb{R}^2} \gamma(W^{(1)}, c)[W^{(1)}(x_j - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c) = y_j, \tag{21}$$

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 $\mathrm{supp}(\nu_{\mathcal{C}})$ 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 H.1). This shows that $\gamma(W^{(1)}, c)$ is closely related to $g''(x, \gamma)$. In the following we seek to express (21) in terms of $g''(x, \gamma)$. Since $g''(x, \gamma)$ determines $g(x, \gamma)$ only up to

5. 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.

linear functions, we consider the following problem:

$$\min_{\gamma\in C(\mathbb{R}^2),u\in\mathbb{R},v\in\mathbb{R}} \quad \int_{\mathbb{R}^2} \gamma^2(W^{(1)},c)\ \mathrm{d}\nu(W^{(1)},c)$$
$$\text{subject to} \quad ux_j + v + \int_{\mathbb{R}^2} \gamma(W^{(1)},c)[W^{(1)}(x_j-c)]_+\ \mathrm{d}\nu(W^{(1)},c) = y_j, \quad j=1,\dots,M. \tag{22}$$

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 (22) are small, then the solution is close to the solution of (21). Ongie et al. (2020) also use this trick to simplify the characterization of neural networks in function space. Next we study the solution of (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 (22), and consider the corresponding output function*

$$g(x,(\overline{\gamma},\overline{u},\overline{v})) = \overline{u}x + \overline{v} + \int_{\mathbb{R}^2} \overline{\gamma}(W^{(1)},c)[W^{(1)}(x-c)]_+\ \mathrm{d}\nu(W^{(1)},c). \tag{23}$$

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, $x_i\in\operatorname{supp}(\zeta)$, $i=1,\dots,m$, 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:

$$\min_{h\in C^2(S)} \int_S \frac{(h''(x))^2}{\zeta(x)}\ \mathrm{d}x \quad s.t. \quad h(x_j)=y_j, \quad j=1,\dots,m. \tag{24}$$

The proof is provided in Appendix 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 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 $(\mathbb{E}(W^2|C=x)p_{\mathcal{C}}(x))^{-1}=(\mathbb{E}(W^2)p_{\mathcal{C}}(x))^{-1}\propto\varrho(x)$, $x\in\mathbb{R}$.*

7. Implicit Bias for Multivariate Regression

In this section we solve the optimization problem (20) in the multivariate case. Similar to Section 6, we can relax the optimization problem to

$$\begin{aligned}&\min_{\substack{\alpha\in C(\mathbb{R}^d\times\mathbb{R}),\\ \mathbf{u}\in\mathbb{R}^d, v\in\mathbb{R}}} \int_{\mathbb{R}^d\times\mathbb{R}} \alpha^2(\mathbf{W}^{(1)},b)\ \mathrm{d}\mu(\mathbf{W}^{(1)},b)\\ &\text{subject to } \int_{\mathbb{R}^d\times\mathbb{R}} \alpha(\mathbf{W}^{(1)},b)[\langle \mathbf{W}^{(1)},\mathbf{x}_j\rangle + b]_+\ \mathrm{d}\mu(\mathbf{W}^{(1)},b) + \langle \mathbf{u},\mathbf{x}_j\rangle + v = y_j,\ j=1,\dots,M.\end{aligned} \tag{25}$$

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 (25) is equivalently expressed as

$$\begin{aligned}&\min_{\substack{\alpha\in C(\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}),\\ \mathbf{u}\in\mathbb{R}^d, v\in\mathbb{R}}} \int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma^2(u,\boldsymbol{V},c)\ \mathrm{d}\nu(u,\boldsymbol{V},c)\\ &\text{subject to } \int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma(u,\boldsymbol{V},c)\cdot u\cdot[\langle \boldsymbol{V},\mathbf{x}_j\rangle - c]_+\ \mathrm{d}\nu(u,\boldsymbol{V},c) + \langle \mathbf{u},\mathbf{x}_j\rangle + v = y_j,\ j=1,\dots,M.\end{aligned} \tag{26}$$

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

$$\begin{aligned}\Delta g(\mathbf{x},(\gamma,\mathbf{u},v)) &= \int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma(u,\boldsymbol{V},c)\cdot u\cdot\ \delta(\langle \boldsymbol{V},\mathbf{x}\rangle - c)\ \mathrm{d}\nu(u,\boldsymbol{V},c)\\ &= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\int_{\mathbb{R}^+} \gamma(u,\boldsymbol{V},c)\cdot u\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u)\right)\delta(\langle \boldsymbol{V},\mathbf{x}\rangle - c)\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V},c),\end{aligned} \tag{27}$$

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

$$\kappa(\boldsymbol{V},c) = \int_{\mathbb{R}^+} \gamma(u,\boldsymbol{V},c)\cdot u\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u). \tag{28}$$

Then (27) becomes

$$\begin{aligned}\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) &= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \kappa(\boldsymbol{V},c)\ \delta(\langle \boldsymbol{V},\mathbf{x}\rangle - c)\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V},c)\\ &= \int_{\mathbb{S}^{d-1}} \left(\int_{\mathbb{R}} \kappa(\boldsymbol{V},c)\ \delta(\langle \boldsymbol{V},\mathbf{x}\rangle - c) p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\mathrm{d}c\right)\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\\ &= \int_{\mathbb{S}^{d-1}} \kappa(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\ p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(\langle \boldsymbol{V},\mathbf{x}\rangle)\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}),\end{aligned} \tag{29}$$

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 (29) becomes

$$\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) = \int_{\mathbb{S}^{d-1}} \kappa(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\; p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(\langle \boldsymbol{V},\mathbf{x}\rangle) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\; \mathrm{d}\sigma^{d-1}(\boldsymbol{V}). \tag{30}$$

Now, defining

$$\beta(\boldsymbol{V},c) = \kappa(\boldsymbol{V},c)\; p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\; p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}), \tag{31}$$

we observe that

$$\begin{aligned}\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) &= \int_{\mathbb{S}^{d-1}} \beta(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\; \mathrm{d}\boldsymbol{V} \\ &= \mathcal{R}^*\{\beta\}(\mathbf{x}).\end{aligned} \tag{32}$$

The right-hand side of (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 $\Delta g(\mathbf{x},(\alpha,\mathbf{u},v)) = \int_{\mathbb{S}^{d-1}} \beta^+(\boldsymbol{V},\langle \boldsymbol{V},\mathbf{x}\rangle)\; \mathrm{d}\boldsymbol{V}$. Ongie et al. (2020) observed that β^+ can be recovered from Δg by using the inversion formula of the dual Radon transform. According to Ongie et al. (2020, Lemma 3),

$$\beta^+ = -\frac{1}{2(2\pi)^{d-1}} \mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot,\alpha)\}, \tag{33}$$

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 (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 Kwaśnicki, 2017). Then according to (33) and the definition of β, we have

$$\begin{aligned}&\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot,\alpha)\}(\boldsymbol{V},c) - 2(2\pi)^{d-1}\beta^- \\ =& -2(2\pi)^{d-1}\kappa(\boldsymbol{V},c) p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \\ =& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \int_{\mathbb{R}^+} \gamma(u,\boldsymbol{V},c)\cdot u\; \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u).\end{aligned} \tag{34}$$

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 (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 (26), and assume that (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

$$g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v})) = \int_{\mathbb{R}^+ \times \mathbb{S}^{d-1} \times \mathbb{R}} \overline{\gamma}(u, \boldsymbol{V}, c) \cdot u \cdot [\langle \boldsymbol{V}, \mathbf{x} \rangle - c]_+\ \mathrm{d}\nu(u, \boldsymbol{V}, c) + \langle \overline{\mathbf{u}}, \mathbf{x} \rangle + \overline{v}. \tag{35}$$

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 $\zeta \colon \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$,

$$\zeta(\boldsymbol{V}, c) = p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c). \tag{36}$$

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

$$\begin{aligned} \min_{h \in \mathrm{Lip}(\mathbb{R}^d) \cap C(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \frac{\big(\mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c)\big)^2}{\zeta(\boldsymbol{V}, c)}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \mathrm{d}c \\ \text{subject to} \quad & h(\mathbf{x}_j) = y_j, \quad j = 1, \ldots, M, \\ & \mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c) = 0,\ \forall (\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta), \\ & (-\Delta)^{(d+1)/2} h \in L^p(\mathbb{R}^d),\ 1 \le p < d/(d-1), \end{aligned} \tag{37}$$

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 I.1. The optimization problem (37) characterizes the implicit bias of the gradient descent in function space for the multivariate setting. Zhang et al. (2020) 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 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 (37) under appropriate conditions.

To conclude this section, we discuss the function ζ in the variational problem (37). The proofs of the following statements are presented in Appendix 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 $\{(\boldsymbol{V}, c) : |c| \le a\}$ 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*

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

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., Wendland, 2004).

Zhang et al. (2020) 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 Zhang et al. (2020), 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 (Rahaman et al., 2019). Cao et al. (2021) 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.

Acknowledgments

This project has been supported by ERC Starting Grant 757983, NSF CAREER Grant 2145630, DFG SPP 2298 Grant 464109215.

Appendix

The appendix is organized as follows.

- In Appendix A we illustrate our theoretical results numerically, and provide details on the numerical implementation.
- In Appendix 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 C, D, E, F, G, we provide proofs and supporting results for the results presented in Sections 3, 4.1, 4.2, and 5.
- In Appendices H and I, we provide the proofs of the results in Sections 6 and 7 for univariate and multivariate regression respectively.
- In Appendix J, we prove a corresponding result for activation functions other than ReLU.
- In Appendix 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 L, we introduce the network with skip connections and obtain the same implicit bias result without adjusting the training data.
- In Appendix 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 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 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 P we comment on possible extensions and generalizations of the analysis.

Appendix 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 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 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 O. For a general non-constant curvature penalty function $1/\zeta$, we can obtain a numerical solution to problem (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 (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 (8). Hence we numerically solve (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 (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 3 and 4 we show analogous experiments to those in the left panel of Figure 1, but using two types of Gaussian initialization instead of the uniform initialization. As we already observed in the right panel of Figure 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 5 we show an experiment similar to that of Figure 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 1 and use uniform initialization. Starting from the same initial weights, we train the network in two

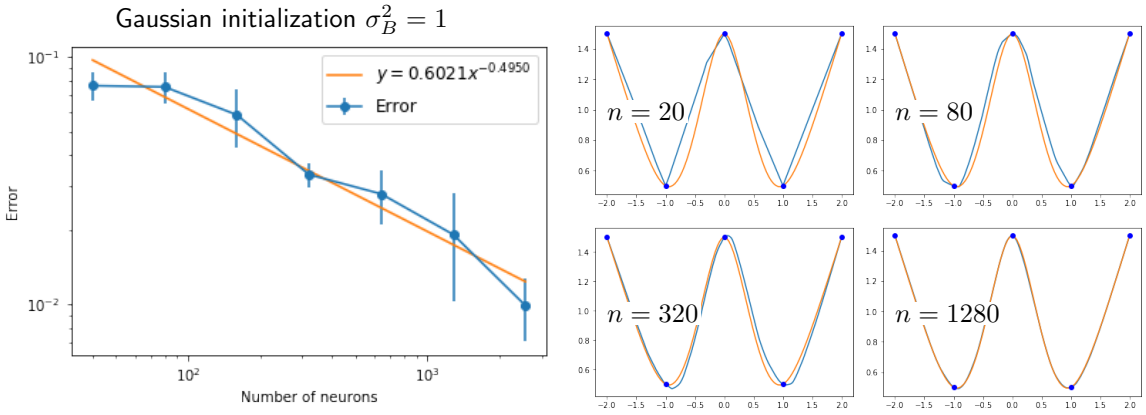

Figure 3: Illustration of Theorem 1. Shown is the error between the output function $f(\cdot, \theta^*)$ of the trained neural network and the solution g^* to the variational problem (24) against the number of neurons, n. Shown is the average over 5 repetitions, with error bars indicating the standard 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.

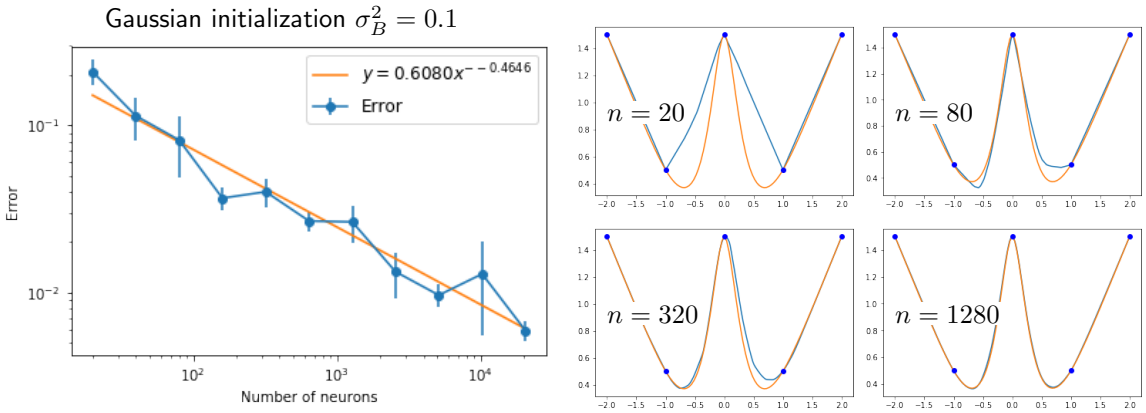

Figure 4: Illustration of Theorem 1. Similar to Figure 3, but with a different initialization $\mathcal{W} \sim \mathcal{N}(0, 1)$ and $\mathcal{B} \sim \mathcal{N}(0, 0.1)$, which gives rise to a curvature penalty function ζ that is more strongly peaked around $x = 0$ (see Figure 1). We observe in particular that the solutions are more curvy around $x = 0$.

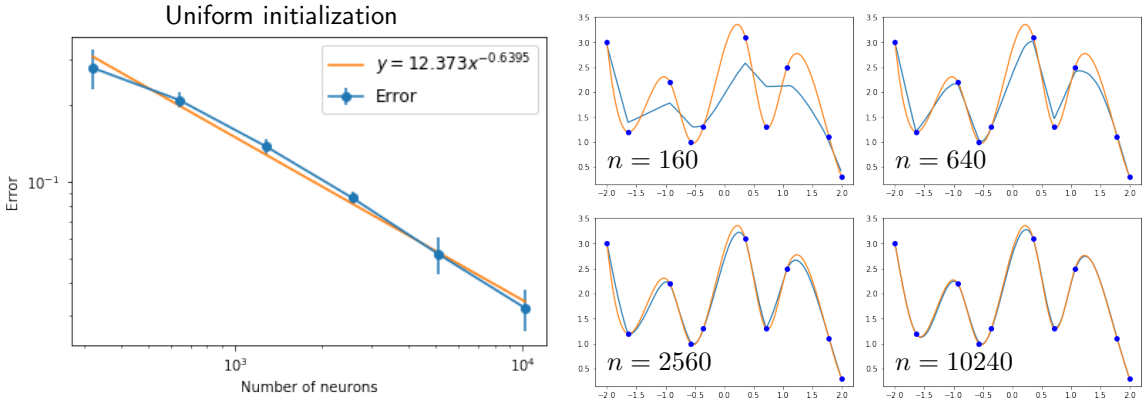

Figure 5: Illustration of Theorem 1. Similar to Figure 1, with uniform initialization, but with a larger data set and larger networks.

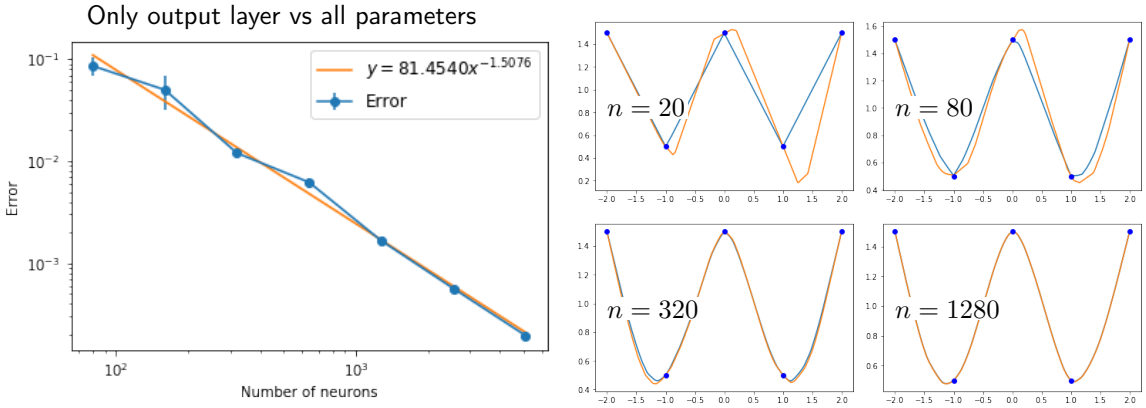

Figure 6: Illustration of Theorem 10. Training only output layer vs training all parameters of the network. We use uniform initialization and the same training set as in Figure 1. The left panel plots the error between two trained network functions against the number of neurons n. For one network, we 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.

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 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 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

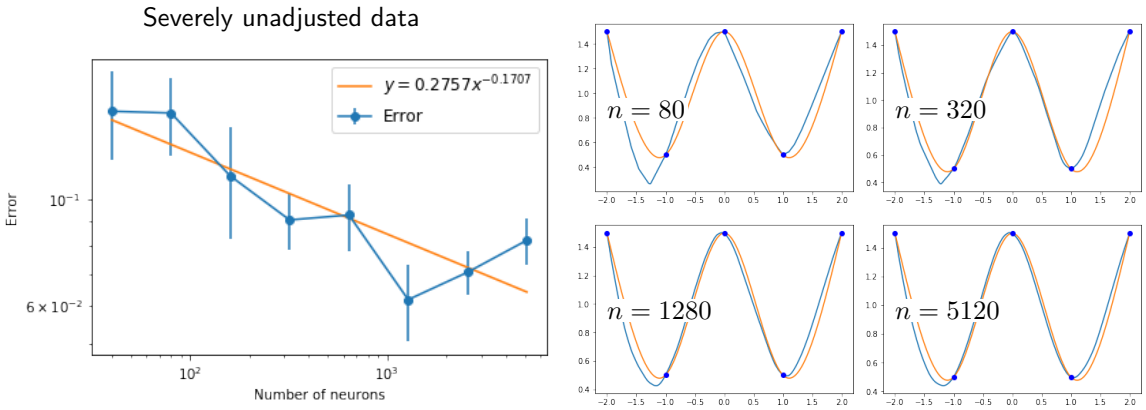

Figure 7: Effect of not adjusting the data. We use uniform initialization and add a linear function $10x + 10$ to the training data of Figure 1. In order to clearly show the difference between trained network function and the solution to the variational problem, we subtract $10x + 10$ from these two functions in the right panel. In the right panel we see that if we ignore u and v in the variational problem (22), the solution is slightly different from (24).

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 K. In order evaluate the effect of this linear function on the implicit bias, we conduct the following experiment. Similar to Figure 1, we use uniform initialization. We add a linear function $10x+10$ to the training data in Figure 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 7 we show analogous experiments to those in the left panel of Figure 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 7. From the right panel of Figure 7, we see that the difference between plotted two functions is relatively larger than that in Figure 1. From the left panel of Figure 7, we see that the error between these two functions stops to decrease when number of neurons n 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. We conduct experiments similar to Figure 1 and Figure 3 for the bivariate case. The initialization used in Figure 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 (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 8 we show an experiment similar to that of Figure 2, but for a larger data set.

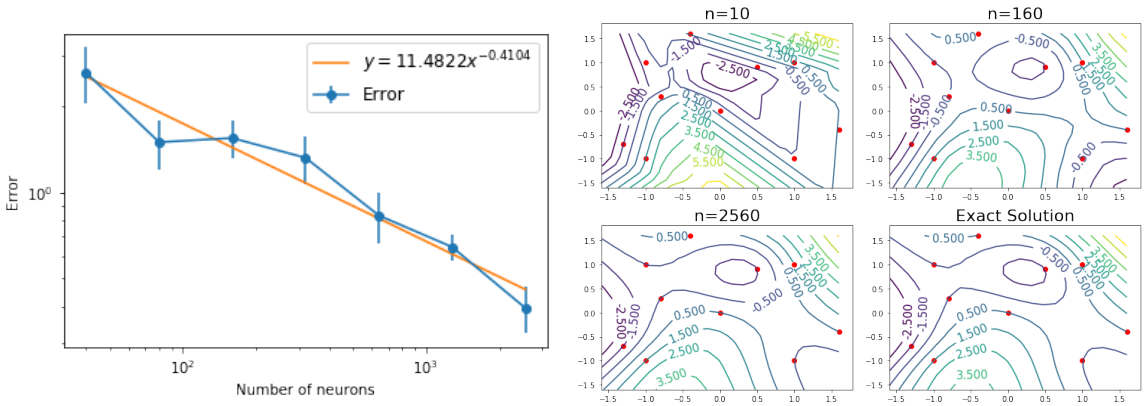

Figure 8: Illustration of Theorem 6. Similar to Figure 2, with the same initialization, but with a larger data set.

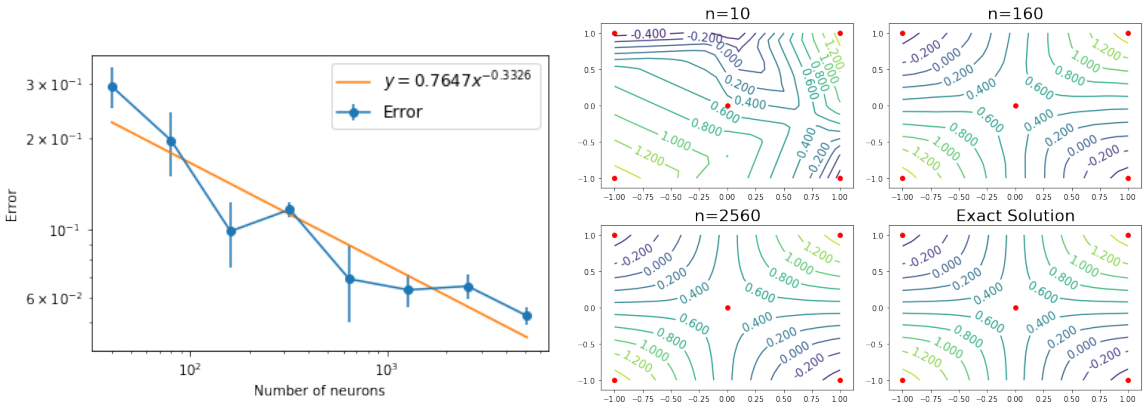

Figure 9: Illustration of Theorem 6. Similar to Figure 2, but with the Gaussian initialization $\boldsymbol{\mathcal{W}} \sim \mathcal{N}(0, I_d)$ and $\mathcal{B} \sim \mathcal{N}(0, 1)$.

To illustrate Theorem 6 across different initialization procedures, in Figures 9 and 10 we show analogous experiments to Figure 2, but using Gaussian initialization instead. The initialization used in Figure 9 is $\boldsymbol{\mathcal{W}} \sim \mathcal{N}(0, I_d)$ and $\mathcal{B} \sim \mathcal{N}(0, 1)$, and the initialization used in Figure 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 (8) numerically.

Appendix 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.

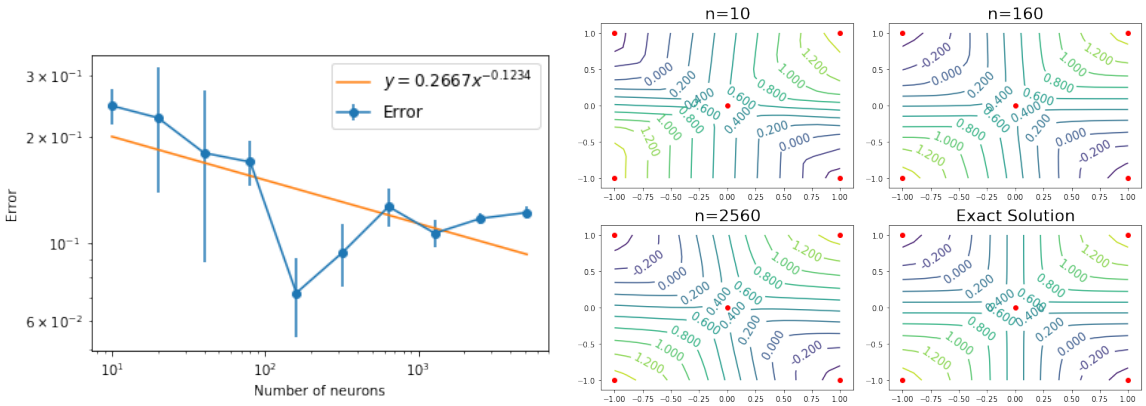

Figure 10: Illustration of Theorem 6. Similar to Figure 2, but with the Gaussian initialization $\boldsymbol{\mathcal{W}} \sim N(0, I_d)$ and $\mathcal{B} \sim N(0, 0.1)$. Because of the linear adjustment, the exact solution of the variational problem (8) is slightly different from the network output with a large number of hidden neurons.

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 Jacot et al. (2018) 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 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 $\partial_{\theta_i} f(\cdot, \theta)$, $i = 1, \dots, d$, becomes dense in function space with respect to $\|\cdot\|_p$ as the width of the network tends to infinity (Jacot et al., 2018). For a finite data set $\mathbf{x}_1, \dots, \mathbf{x}_M$, positive definiteness of the corresponding Gram matrix is equivalent to $\partial_{\theta_i} f(\mathbf{x}_j, \cdot)$ being linearly independent (Du et al., 2019, 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 (Neal, 1996). 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.

B.2 Anti-Symmetrical Initialization (ASI)

The AntiSymmetrical Initialization (ASI) trick as proposed by Zhang et al. (2020) 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 (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 Zhang et al. (2020). A related discussion also appears in Sahs et al. (2020a).

B.3 Standard vs NTK Parametrization

We have focused on the standard parametrization of the neural network. Jacot et al. (2018) 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 (20) to describe the result of gradient descent as $n \to \infty$.

B.4 Weight Norm Minimization

Savarese et al. (2019) 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 Savarese et al. (2019) 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 (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 Chizat et al. (2019) and we further illustrate it in Appendix 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 Savarese et al. (2019) 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 Savarese et al. (2019). 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 Savarese et al. (2019), 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 Parhi and Nowak (2019) follows the approach of Savarese et al. (2019) and generalizes the result of Savarese et al. (2019) 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 J.

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 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.

Appendix C. Proof of Theorem 1 and Theorem 6

The proof of Theorem 1 and Theorem 6 is the compilation of results from Sections 4, 5, 6 and 7. Next we give the proof of Theorem 6. Theorem 1 can be similarly proved.

Proof [Proof of Theorem 6| The convergence to zero training error for ReLU networks is by now a well known result (Du et al., 2019; Allen-Zhu et al., 2019). We proceed with the implicit bias result.

For simplicity, we give the proof under ASI (see Appendix B.2). In Section 7, we relax the optimization problem (20) to (25). Suppose $(\overline{\alpha}, \overline{\boldsymbol{u}}, \overline{v})$ is the solution of (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 (25). Then $\overline{\alpha}$ is the solution of (20) on the adjusted data. Furthermore, the solution of (20) in function space, $g(\mathbf{x}, \overline{\alpha})$, equals to the solution of (25) in function space, $g(\mathbf{x}, (\overline{\alpha}, 0, 0))$, i.e.,

$$g(\mathbf{x}, \overline{\alpha}) = g(\mathbf{x}, (\overline{\alpha}, \mathbf{0}, 0)). \tag{38}$$

It we change the variable α to γ as in Section 7, we get

$$g(\mathbf{x}, (\overline{\alpha}, \mathbf{0}, 0)) = g(\mathbf{x}, (\overline{\gamma}, \mathbf{0}, 0)), \tag{39}$$

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

$$\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, \overline{\alpha}_n) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2}), \tag{40}$$

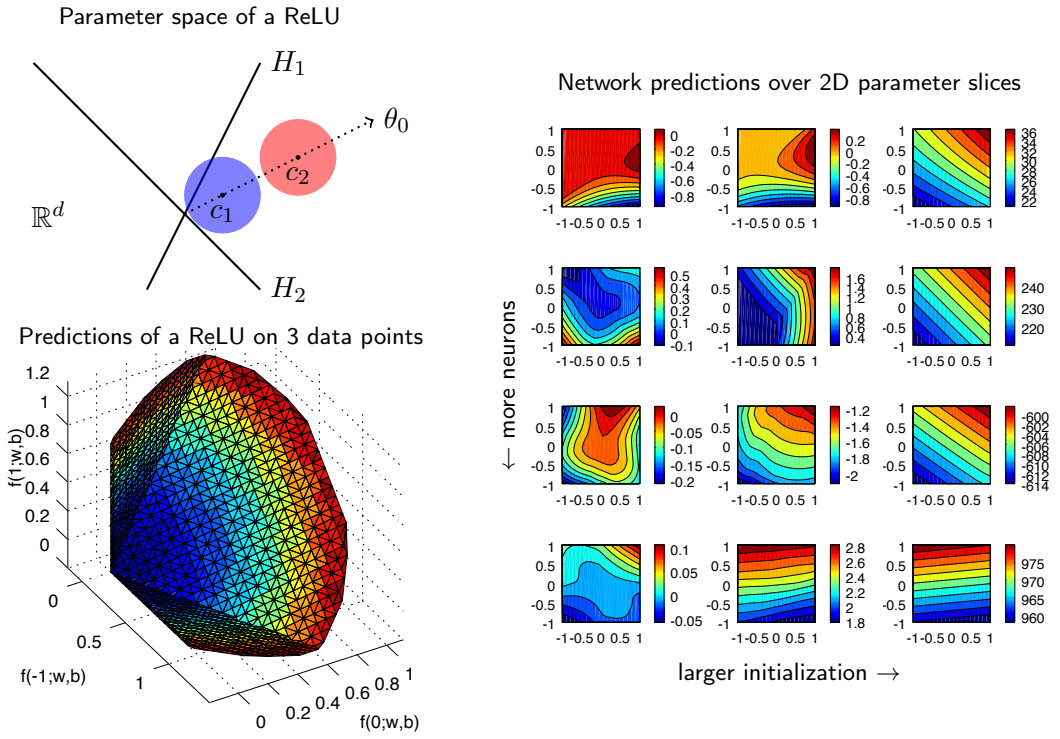

Figure 11: Left: For a single ReLU, the map $\theta \mapsto f(\mathcal{X}, \theta)$ from parameters to prediction vectors over a set $\mathcal{X} = \{x_1, \ldots, x_m\}$ of m input data points is piecewise linear, with pieces separated by m central hyperplanes. Right: Shown is the prediction $f(x, \theta)$ 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.

where $g_n(\mathbf{x}, \overline{\alpha}_n)$ is the solution of problem (19) in function space. Since problem (19) is equivalent to problem (18), $g_n(\mathbf{x}, \overline{\alpha}_n)$ is also the solution of (18) in function space. According to discussion in Section 5, $f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_\infty)$ is the solution of (18). Then we have

$$g_n(\mathbf{x}, \overline{\alpha}_n) = f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_\infty). \tag{41}$$

According to Corollary 11, we get

$$\sup_{\mathbf{x} \in D} |f^{\text{lin}}(\mathbf{x}, \tilde{\omega}_\infty) - f(\mathbf{x}, \theta^*)| = O_p(n^{-\frac{1}{2}}). \tag{42}$$

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 (8), which is $g^*(\mathbf{x})$. It means that

$$g(\mathbf{x}, (\overline{\gamma}, 0, 0)) = g^*(\mathbf{x}). \tag{43}$$

Combining (38), (39), (40), (41), (42), (43), we prove the theorem. ■

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

Zhang et al. (2020) 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 (14) with learning rate $\eta \leq \frac{M}{Kn\lambda_{\max}(\hat{\Theta}_n)}$ converges to the unique solution of following constrained optimization problem:*

$$\min_{\omega} \|\omega - \theta_0\|_2 \quad s.t.\ f^{\text{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \tag{44}$$

The derivative $\frac{\mathrm{d}}{\mathrm{d}y}$ is with respect to the first argument of ℓ and the gradient ∇_θ is with respect to the second argument of f (see notation in Section 2).

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 the proof of Theorem 20. We note that Zhang et al. (2020) prove a similar result for gradient flow. Our proof is for finite step size and different from theirs.

Proof [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 (44) is a strongly convex optimization problem. The first order optimality condition of the problem is

$$
\begin{cases}
\omega - \theta_0 + \nabla_\theta f^{\text{lin}}(\mathcal{X},\theta_0)^T \lambda = 0, \\
f^{\text{lin}}(\mathcal{X},\omega) = \mathcal{Y}.
\end{cases} \tag{45}
$$

Let $\lambda = \sum_{t=0}^{\infty} \nabla_{f^{\text{lin}}(\mathcal{X},\theta_t)} L$, we can easily check that ω_∞ satisfies condition (45). So ω_∞ is the solution of problem (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. Oymak and Soltanolkotabi (2019) 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:

$$\omega_0 = \theta_0, \quad \omega_{t+1} = \omega_t - \eta_t \frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\mathrm{lin}}(\mathbf{x}_{r(t)}, \omega_t), y_{r(t)})\nabla_\theta f(\mathbf{x}_{r(t)}, \theta_0), \tag{46}$$

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., $|\frac{\mathrm{d}}{\mathrm{d}y}\ell(y_1, \hat{y}) - \frac{\mathrm{d}}{\mathrm{d}y}\ell(y_2, \hat{y})| \le K|y_1 - y_2|$. If $\mathrm{rank}(\nabla_\theta f(\mathcal{X}, \theta_0)) = M$, the stochastic gradient descent iteration (46) with constant learning rate $\eta_t = \eta \le \frac{1}{K \max_j \|\nabla_\theta f(\mathbf{x}_j, \theta_0)\|_2^2}$ converges to the unique solution of following constrained optimization problem with probability 1:*

$$\min_\omega \|\omega - \theta_0\|_2 \quad s.t.\ f^{\mathrm{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \tag{47}$$

Proof [Proof of Theorem 24] Let ω^* be the solution to the optimization problem (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 $\hat{\eta} \le 1$, we have

$$\left\|\omega_t - \hat{\eta}\langle \omega_t - \omega^*, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}\rangle \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} - \omega^*\right\|_2^2 = \|\omega_t - \omega^*\|_2^2 - (1 - (1 - \hat{\eta})^2)\left|\langle \omega_t - \omega^*, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2^2}\rangle\right|^2 \tag{48}$$

$$\le \|\omega_t - \omega^*\|_2^2. \tag{49}$$

The length of the stochastic gradient in (46) can be bounded as follows:

$$\begin{aligned}
&\eta_t \frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\mathrm{lin}}(\mathbf{x}_{r(t)},\omega_t),y_{r(t)})\|\mathbf{z}_{r(t)}\|_2\\
&\le \eta_t K|f^{\mathrm{lin}}(\mathbf{x}_{r(t)},\omega_t)-y_{r(t)}|\|\mathbf{z}_{r(t)}\|_2\\
&\le K\frac{1}{K\max_j\|\nabla_\theta f(\mathbf{x}_j,\theta_0)\|_2^2}|f^{\mathrm{lin}}(\mathbf{x}_{r(t)},\omega_t)-y_{r(t)}|\|\mathbf{z}_{r(t)}\|_2\\
&= \frac{1}{\max_j\|\mathbf{z}_j\|_2^2}\langle\omega_t-\omega^*,\mathbf{z}_{r(t)}\rangle\|\mathbf{z}_{r(t)}\|_2\\
&\le \frac{1}{\max_j\|\mathbf{z}_j\|_2^2}\|\mathbf{z}_{r(t)}\|_2^2\langle\omega_t-\omega^*,\frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2}\rangle\\
&\le \langle\omega_t-\omega^*,\frac{\mathbf{z}_{r(t)}}{\|\mathbf{z}_{r(t)}\|_2}\rangle.
\end{aligned}$$

Then according to (49), we have

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

The above equation means that

$$\|\omega_{t+1}-\omega^*\|_2\le\|\omega_t-\omega^*\|_2. \tag{50}$$

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 $\{\omega_{t_k}\}_{k\ge1}$ of $\{\omega_t\}_{t\ge1}$, 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 (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 $t_k'\ge t_k$ such that $r(t_k')\in A$ and $r(t)\notin A$ for $t_k\le t<t_k'$.

When we prove (50), we only use the property that $f^{\mathrm{lin}}(\mathbf{x}_{r(t)},\omega^*)=y_{r(t)}$. When $t_k\le t<t_k'$, 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 (50), we have $\|\omega_{t+1}-\bar\omega\|_2\le\|\omega_t-\bar\omega\|_2$ when $t_k\le t<t_k'$. Then $\|\omega_{t_k'}-\bar\omega\|_2\le\|\omega_{t_k}-\bar\omega\|_2$. Then $\lim_{k\to\infty}\omega_{t_k'}=\lim_{k\to\infty}\omega_{t_k}=\bar\omega$. According to (48), we have

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

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

Since $\lim_{k\to\infty}\omega_{t_k'}=\bar{\omega}$, for sufficiently large k we have

$$\begin{aligned}\left|\langle\omega_{t_k'}-\omega^*,\frac{\mathbf{z}_{r(t_k')}}{\|\mathbf{z}_{r(t_k')}\|_2}\rangle\right|^2 &\geq \frac{1}{2}\min_{j\in A}\left|\langle\bar{\omega}-\omega^*,\frac{\mathbf{z}_{r(j)}}{\|\mathbf{z}_{r(j)}\|_2}\rangle\right|^2\\ &=\Omega(1),\end{aligned}\tag{51}$$

and

$$\begin{aligned}\tilde{\eta}_t &\geq \frac{1}{2}\frac{\eta\min_{j\in A}\frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\mathrm{lin}}(\mathbf{x}_j,\bar{\omega}),y_j)\min_{j\in A}\|\mathbf{z}_j\|_2}{\|\bar{\omega}-\omega^*\|_2}\\ &=\Omega(1)\min_{j\in A}\frac{\mathrm{d}}{\mathrm{d}y}\ell(f^{\mathrm{lin}}(\mathbf{x}_j,\bar{\omega}),y_j)\\ &=\Omega(1),\end{aligned}\tag{52}$$

where (52) 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 (51) and (52) 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.*

Appendix 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 B.1.

Proof [Proof of Theorem 10] Since set D is compact and $\mathbf{x}\in D$, we have $\|\mathbf{x}\|_2\leq C$ for a fixed constant C. According to (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 (12), we get

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

Then we have

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

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

$$\omega_t = -n^{-1}\nabla_\theta f(\mathcal{X},\theta_0)^T\hat{\Theta}_n^{-1}[I-(I-n\eta\hat{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y})+\theta_0. \tag{53}$$

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

$$\begin{aligned} f^{\text{lin}}(\mathbf{x},\omega_t) &= f(\mathbf{x},\theta_0)+\nabla_\theta f(\mathbf{x},\theta_0)(\omega_t-\theta_0) \\ &= f(\mathbf{x},\theta_0)-n^{-1}\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T\hat{\Theta}_n^{-1}[I-(I-n\eta\hat{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y}). \end{aligned} \tag{54}$$

For the training process (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

$$\widetilde{\mathbf{W}}_t^{(2)} = -n^{-1}\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\tilde{\Theta}_n^{-1}[I-(I-n\eta\tilde{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y})+\overline{\mathbf{W}}_0^{(2)} \tag{55}$$

and

$$f^{\text{lin}}(\mathbf{x},\widetilde{\omega}_t) = f(\mathbf{x},\theta_0)-n^{-1}\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\tilde{\Theta}_n^{-1}[I-(I-n\eta\tilde{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y}). \tag{56}$$

According to (54) and (56), we have

$$\begin{aligned} &|f^{\text{lin}}(\mathbf{x},\widetilde{\omega}_t)-f^{\text{lin}}(\mathbf{x},\omega_t)| \\ =&n^{-1}\Big|\nabla_\theta f(\mathbf{x},\theta_0)\nabla_\theta f(\mathcal{X},\theta_0)^T\hat{\Theta}_n^{-1}[I-(I-n\eta\hat{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y}) \\ &-\nabla_{\mathbf{W}^{(2)}}f(\mathbf{x},\theta_0)\nabla_{\mathbf{W}^{(2)}}f(\mathcal{X},\theta_0)^T\tilde{\Theta}_n^{-1}[I-(I-n\eta\tilde{\Theta}_n)^t](f(\mathcal{X},\theta_0)-\mathcal{Y})\Big|. \end{aligned} \tag{57}$$

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

$$\begin{aligned}(\Delta\Theta)_{ij} =& \frac{1}{n}\left[\sum_{k=1}^{n}\left(\left\langle \nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x}_i,\theta_0), \nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x}_j,\theta_0)\right\rangle + \nabla_{b_k^{(1)}} f(\mathbf{x}_i,\theta_0)\nabla_{b_k^{(1)}} f(\mathbf{x}_j,\theta_0)\right)\right. \\ &\left. + \nabla_{b^{(2)}} f(\mathbf{x}_i,\theta_0)\nabla_{b^{(2)}} f(\mathbf{x}_j,\theta_0)\right].\end{aligned} \tag{58}$$

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

$$\|\nabla_{\mathbf{W}_k^{(1)}} f(\mathbf{x},\theta_0)\| = \|W_k^{(2)} H(\langle \mathbf{W}_k^{(1)}, \mathbf{x}\rangle + b_k^{(1)})\cdot \mathbf{x}\| \le C|W_k^{(2)}| \tag{59}$$

$$|\nabla_{b_k^{(1)}} f(\mathbf{x},\theta_0)| = |W_k^{(2)} H(\langle \mathbf{W}_k^{(1)}, \mathbf{x}\rangle + b_k^{(1)})| \le |W_k^{(2)}| \tag{60}$$

$$|\nabla_{W_k^{(2)}} f(\mathbf{x},\theta_0)| = [\langle \mathbf{W}_k^{(1)}, \mathbf{x}\rangle + b_k^{(1)}]_+ \le C\|\mathbf{W}_k^{(1)}\|\| + b_k^{(1)} \tag{61}$$

$$|\nabla_{b^{(2)}} f(\mathbf{x},\theta_0)| = 1. \tag{62}$$

Therefore,

$$\begin{aligned}|(\Delta\Theta)_{ij}| &\le \frac{1}{n}\left[\sum_{k=1}^{n}\left(|W_k^{(2)}|^2\|\mathbf{x}_i\|\|\mathbf{x}_j\| + |W_k^{(2)}|^2\right) + 1\right] \\ &= \frac{C^2+1}{n}\sum_{k=1}^{n}|W_k^{(2)}|^2 + \frac{1}{n}.\end{aligned} \tag{63}$$

According to initialization (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 (57). According to assumption $\inf_n \lambda_{\min}(\hat{\Theta}_n) > 0$, we have

$$\lambda_{\min}(\hat{\Theta}_n^{-1}) \ge \frac{1}{\inf_n \lambda_{\min}(\hat{\Theta}_n)} = O_p(1) \tag{64}$$

$$\lambda_{\min}(\tilde{\Theta}_n^{-1}) \ge \frac{1}{\inf_n \lambda_{\min}(\hat{\Theta}_n) - O_p(n^{-1})} = O_p(1). \tag{65}$$

Therefore

$$\begin{aligned}\|\hat{\Theta}_n^{-1} - \tilde{\Theta}_n^{-1}\|_2 &= \|\hat{\Theta}_n^{-1}(\tilde{\Theta}_n - \hat{\Theta}_n)\tilde{\Theta}_n^{-1}\|_2 \\ &\le \|\hat{\Theta}_n^{-1}\|_2\|\Delta\Theta\|_2\|\tilde{\Theta}_n^{-1}\|_2 \\ &= O_p(n^{-1}).\end{aligned} \tag{66}$$

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

$$\|I - n\eta\hat{\Theta}_n\|_2 < 1, \tag{67}$$

and

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

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,

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

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

$$\begin{aligned}\|[I-(I-n\eta\hat{\Theta}_n)^t]-[I-(I-n\eta\tilde{\Theta}_n)^t]\|_2 &\leq O(\eta\|\hat{\Theta}_n-\tilde{\Theta}_n\|_2)\\ &=O_p(n^{-1}).\end{aligned}\tag{68}$$

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

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

Similar to (63), we have

$$(\Delta\Theta(\mathbf{x},\mathcal{X}))_i|=O_p(n^{-1}).\tag{69}$$

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

$$\|\Delta\Theta(\mathbf{x},\mathcal{X})\|_2=O_p(n^{-1}).\tag{70}$$

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

$$\begin{aligned}
|(\tilde{\Theta}_n(\mathbf{x}, \mathcal{X}))_i| &\leq \frac{1}{n} \sum_{k=1}^{n} |\nabla_{W_k^{(2)}} f(\mathbf{x}, \theta_0) \nabla_{W_k^{(2)}} f(\mathbf{x}_i, \theta_0)| \\
&\leq \frac{1}{n} \sum_{k=1}^{n} |(\|\mathbf{W}_k^{(1)}\| \|\mathbf{x}\|_2 + b_k^{(1)})(\|\mathbf{W}_k^{(1)}\| \|\mathbf{x}_i\|_2 + b_k^{(1)})|. \qquad (71) \\
&\leq \frac{1}{n} \sum_{k=1}^{n} |(C\|\mathbf{W}_k^{(1)}\| + b_k^{(1)})(C\|\mathbf{W}_k^{(1)}\| + b_k^{(1)})|.
\end{aligned}$$

According to initialization (3), $(\mathbf{W}_k^{(1)}, b_k^{(1)}) \overset{d}{=} (\boldsymbol{\mathcal{W}}, \mathcal{B})$. Then according to the law of large numbers,

$$|(\tilde{\Theta}_n(\mathbf{x}, \mathcal{X}))_i| = O_p(1). \qquad (72)$$

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

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

Neal (1996), Lee et al. (2018) 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

$$\|f(\mathcal{X}, \theta_0) - \mathcal{Y}\|_2 = O_p(1). \qquad (73)$$

Then following (57) and (73), 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 (66), (67), (68), (70) and (72), we have that

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

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 and $\mathbf{x}$. Then we get

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

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 (53) and (55),

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

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 (59), 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

$$\sup_t \|\overline{\mathbf{b}}^1 - \widehat{\mathbf{b}}_t^1\|_2 = O_p(n^{-1}), \text{ as } n \to \infty, \tag{74}$$

$$\sup_t \|\overline{b}^2 - \widehat{b}_t^2\| = O_p(n^{-1}), \text{ as } n \to \infty. \tag{75}$$

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

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

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 (61), 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

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

■

Appendix 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 (Lai et al., 2023, Proposition 3.2). Although Lai et al. (2023, Proposition 3.2) consider Gaussian initialization, the arguments extend to sub-Gaussian initialization if the initialization distribution has a continuous probability density.

Proof [Proof of Corollary 11| Using Theorem 10, we have that

$$\sup_t |f^{\text{lin}}(\mathbf{x}, \widetilde{\omega}_t) - f^{\text{lin}}(\mathbf{x}, \omega_t)| = O_p(n^{-1}), \text{ as } n \to \infty. \tag{76}$$

According to Lai et al. (2023, Proposition 3.2), 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. Combining the above equation with (76) concludes the proof. ■

Appendix G. Proof of Theorem 12

Proof [Proof of Theorem 12| The Lagrangian of problem (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 (19), we can let

$$\overline{\alpha}_n(\mathbf{W}^{(1)}, b) = -\frac{1}{2}\sum_{j=1}^{M} \lambda_j^{(n)} [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \quad \forall (\mathbf{W}^{(1)}, b) \in \mathbb{R}^{d+1} \tag{77}$$

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

$$-\frac{1}{2}\sum_{j=1}^{M} \lambda_j^{(n)} \int_{\mathbb{R}^2} [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ [\langle \mathbf{W}^{(1)}, \mathbf{x}_i \rangle + b]_+ \ \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) = y_i,\ i = 1, \ldots, M. \tag{78}$$

Similarly, the Lagrangian of problem (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

$$\overline{\alpha}(\mathbf{W}^{(1)}, b) = -\frac{1}{2}\sum_{j=1}^{M} \lambda_j [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \quad \forall (\mathbf{W}^{(1)}, b) \in \mathbb{R}^{d+1}. \tag{79}$$

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

$$-\frac{1}{2}\sum_{j=1}^{M} \lambda_j \int_{\mathbb{R}^2} [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ [\langle \mathbf{W}^{(1)}, \mathbf{x}_i \rangle + b]_+ \ \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = y_i, \quad i = 1, \ldots, M. \tag{80}$$

Compare (78) and (80). 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 (78) and (80) are systems of linear equations and coefficients of (78) converge to coefficients of (80) at the rate of $O_p(n^{-1/2})$, then we get

$$|\lambda_j^n - \lambda_j| = O_p(n^{-1/2}), \quad j = 1, \ldots, M. \tag{81}$$

Compare (77) and (79). Given $(\mathbf{W}^{(1)}, b)$, we have

$$|\overline{\alpha}_n(\mathbf{W}^{(1)}, b) - \overline{\alpha}(\mathbf{W}^{(1)}, b)| = O_p(n^{-1/2}). \tag{82}$$

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,

$$\begin{aligned} \mathbb{E}(g_n(\mathbf{x}, \overline{\alpha})) &= g(\mathbf{x}, \overline{\alpha}) \\ \mathrm{Var}(g_n(\mathbf{x}, \overline{\alpha})) &= \frac{1}{n} \int_{\mathbb{R}^2} [\overline{\alpha}(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+ - g(\mathbf{x}, \overline{\alpha})]^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b). \end{aligned} \tag{83}$$

Here the expectation and the variance are with respect to $(\mathbf{W}_i^{(1)}, b_i)_{i=1}^n$. According to (79) and the assumption that $\mathcal{W}$ and $\mathcal{B}$ have finite fourth moments, the integral in (83) is bounded on D. So $\sup_{\mathbf{x}\in D} \mathrm{Var}\, g_n(\mathbf{x}, \overline{\alpha}) = O(n^{-1})$. 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,

$$\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, \overline{\alpha}) - g(\mathbf{x}, \overline{\alpha})| = O_p(n^{-1/2}). \tag{84}$$

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 (81),

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

Combined with (84), 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. ■

Appendix H. Proofs of Results for Univariate Regression

H.1 Proof of Theorem 13

The second derivative g'' is given by

$$g''(x,\gamma)=p_{\mathcal{C}}(x)\int_{\mathbb{R}}\gamma(W^{(1)},x)\left|W^{(1)}\right|\,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}). \tag{85}$$

The detailed calculation of (85) is as follows:

$$
\begin{aligned}
g''(x,\gamma)&=\int_{\mathbb{R}^2}\gamma(W^{(1)},c)\left|W^{(1)}\right|\delta(x-c)\,\mathrm{d}\nu(W^{(1)},c)\\
&=\int_{\mathrm{supp}(\nu_{\mathcal{C}})}\left(\int_{\mathbb{R}}\gamma(W^{(1)},c)\left|W^{(1)}\right|\,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right)\delta(x-c)\,\mathrm{d}\nu_{\mathcal{C}}(c)\\
&=\int_{\mathrm{supp}(\nu_{\mathcal{C}})}\left(\int_{\mathbb{R}}\gamma(W^{(1)},c)\left|W^{(1)}\right|\,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right)\delta(x-c)p_{\mathcal{C}}(c)\mathrm{d}c\\
&=p_{\mathcal{C}}(x)\int_{\mathbb{R}}\gamma(W^{(1)},x)\left|W^{(1)}\right|\,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}).
\end{aligned}
\tag{86}
$$

Proof [Proof of Theorem 13| First, if $x \notin \text{supp}(\zeta)$, similar to (85), we have

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

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

$$\begin{aligned} \min_{h\in C^2(\text{supp}(\zeta))} \quad & \int_{\text{supp}(\zeta)} \frac{(h''(x))^2}{\zeta(x)}\,\mathrm{d}x \\ \text{subject to} \quad & h(x_j)=y_j, \quad j=1,\ldots,m. \end{aligned} \tag{87}$$

Let $L(f)=\int_{\text{supp}(\zeta)} \frac{(f''(x))^2}{p(x)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)}\mathrm{d}x$. 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 (87) is unique.

Suppose $h(x)$ is the minimizer of problem (87) and $h(x)$ is different from $g(x,(\overline{\gamma},\overline{u},\overline{v}))$ restricted on $\text{supp}(\zeta)$. Then by uniqueness of the solution,

$$L(h) < L(g(\cdot,(\overline{\gamma},\overline{u},\overline{v}))). \tag{88}$$

Now our goal is to find a different (γ,u,v) with smaller cost in problem (22). Then $(\overline{\gamma},\overline{u},\overline{v})$ is not the solution of (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 (85),

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

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 $\text{supp}(\zeta)$. So $g(x_j,(\gamma,u,v))=h(x_j)=y_j$. It means that (γ,u,v) satisfies the condition in problem (22). Next we compute the cost of

(γ, u, v):

$$\begin{aligned}
&\int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c)\ \mathrm{d}\nu(W^{(1)}, c) \\
=&\int_{\mathbb{R}^2} \left(\frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2\ \mathrm{d}\nu(W^{(1)}, c) \\
=&\int_{\mathrm{supp}(\zeta)} \left(\int_{\mathbb{R}} \left(\frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) \mathrm{d}\nu_{\mathcal{C}}(c) \\
=&\int_{\mathrm{supp}(\zeta)} \left(\frac{h''(c)}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2 \left(\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) p_{\mathcal{C}}(c)\mathrm{d}c \\
=&\int_{\mathrm{supp}(\zeta)} \left(\frac{h''(c)}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\right)^2 \left(\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) p_{\mathcal{C}}(c)\mathrm{d}c \\
=&\int_{\mathrm{supp}(\zeta)} \frac{(h''(c))^2}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\mathrm{d}x \\
=&L(h).
\end{aligned} \tag{89}$$

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

$$\begin{aligned}
&\int_{\mathbb{R}^2} \overline{\gamma}^2(W^{(1)}, c)\ \mathrm{d}\nu(W^{(1)}, c) \\
=&\int_{\mathrm{supp}(\zeta)} \left(\int_{\mathbb{R}} \overline{\gamma}^2(W^{(1)}, c)\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right) p_{\mathcal{C}}(c)\mathrm{d}c \\
\geq&\int_{\mathrm{supp}(\zeta)} \frac{\left(\int_{\mathbb{R}} \overline{\gamma}(W^{(1)}, c)|W^{(1)}|\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}\right)^2}{\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}} p_{\mathcal{C}}(c)\mathrm{d}c \quad \text{(Cauchy-Schwarz inequality)} \\
=&\int_{\mathrm{supp}(\zeta)} \frac{\left(g''(c, \overline{\gamma})/p_{\mathcal{C}}(c))\right)^2}{\int_{\mathbb{R}} |W^{(1)}|^2\ \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}} p_{\mathcal{C}}(c)\mathrm{d}c \quad \text{(according to (85))} \\
=&\int_{\mathrm{supp}(\zeta)} \frac{\left(g''(c, \overline{\gamma})\right)^2}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}\mathrm{d}c \\
=&L(g(\cdot, \overline{\gamma})) \\
=&L(g(\cdot, (\overline{\gamma}, \overline{u}, \overline{v}))) \quad (g(\cdot, (\overline{\gamma}, \overline{u}, \overline{v})) \text{ has the same second derivative as } g(\cdot, \overline{\gamma})).
\end{aligned} \tag{90}$$

From this we have

$$\begin{aligned}
\int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c)\ \mathrm{d}\nu(W^{(1)}, c) &= L(h) \quad \text{(according to (89))} \\
&< L(g(\cdot, (\overline{\gamma}, \overline{u}, \overline{v}))) \quad \text{(according to (88))} \\
&\leq \int_{\mathbb{R}^2} \overline{\gamma}^2(W^{(1)}, c)\ \mathrm{d}\nu(W^{(1)}, c) \quad \text{(according to (90))}.
\end{aligned}$$

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 (87), 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 (87). 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 (87). We only need to prove these statements for $h(x)$, which is the solution of (87).

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 (87), so $h(x)$ can be replaced by following $\tilde{h}(x)$ which also satisfies the constraints of problem (87):

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

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 (87) 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 (87), we get $h(x)|_S$ is the solution to problem (87). This concludes the proof. ■

In the case of not using ASI, problem (22) becomes

$$\begin{aligned}\min_{\gamma\in C(\mathbb{R}^2),u\in\mathbb{R},v\in\mathbb{R}}\quad & \int_{\mathbb{R}^2}\gamma^2(W^{(1)},c)\,\mathrm{d}\nu(W^{(1)},c)\\ \text{subject to}\quad & ux_j+v+\int_{\mathbb{R}^2}\gamma(W^{(1)},c)[W^{(1)}(x_j-c)]_+\,\mathrm{d}\nu(W^{(1)},c)=y_j-f(x_j,\theta_0),j=1,\dots,M.\end{aligned}\tag{91}$$

Then Theorem 13 without ASI is stated as follows.

Theorem 26 (Theorem 13 without ASI) *Suppose $(\overline{\gamma},\overline{u},\overline{v})$ is the solution of (91), and consider the corresponding output function*

$$g(x,(\overline{\gamma},\overline{u},\overline{v}))=\overline{u}x+\overline{v}+\int_{\mathbb{R}^2}\overline{\gamma}(W^{(1)},c)[W^{(1)}(x-c)]_+\,\mathrm{d}\nu(W^{(1)},c)+f(x,\theta_0).\tag{92}$$

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:

$$\begin{aligned}\min_{h\in C^2(S)}\quad & \int_S\frac{(h''(x)-f''(x,\theta_0))^2}{\zeta(x)}\,\mathrm{d}x\\ \text{subject to}\quad & h(x_j)=y_j,\quad j=1,\dots,M.\end{aligned}\tag{93}$$

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

Proof [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

$$\begin{aligned}\mathbb{E}(W^2|C=x)p_{\mathcal{C}}(x) &= \int_{\mathbb{R}} W^2 p_{\mathcal{W}|\mathcal{C}=x}(W)\ \mathrm{d}W \cdot p_{\mathcal{C}}(x)\\ &= \int_{\mathbb{R}} W^2 p_{\mathcal{W},\mathcal{C}}(W,x)\ \mathrm{d}W \qquad (94)\\ &= \int_{\mathbb{R}} |W|^3 p_{\mathcal{W},\mathcal{B}}(W,-Wx)\ \mathrm{d}W.\end{aligned}$$

■

Proof [Proof of Proposition 15] The construction is given in the statement of the proposition. ■

Remark 27 (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 Sahs et al. (2020a). We provide the explicit form of the smoothness penalty function for several types of initialization in Appendix H.3.

Remark 28 (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,* (94) *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 (Nasim, 1973) and one can write a formal solution, although it is not always clear whether it will be a density.

H.3 Proof of Theorem 2

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

Proof [Proof of Theorem 2 for Gaussian initialization] Using (94), 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 [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 [Proof of Theorem 2 for uniform initialization] According to Theorem 1 in Sahs et al. (2020a), the density function $p_{\mathcal{C}}(c)$ of $\nu_{\mathcal{C}}$ is

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

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

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

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

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 $c\in[-I,I]$. ■

Appendix I. Proofs of Results for Multivariate Regression

I.1 Proof of Theorem 16

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

Lemma 29 *Let $f\in\mathrm{Lip}(\mathbb{R}^d)$ be considered as a tempered distribution and $(-\Delta)^s f\equiv 0$, $s>0$. Then f is linear, i.e., $f(\mathbf{x})=\langle\mathbf{u},\mathbf{x}\rangle+v$.*

Proof [Proof of Lemma 29] 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

$$S(\mathbb{R}^d):=\left\{\psi\in C^\infty(\mathbb{R}^d)\mid\forall\alpha,\beta\in\mathbb{N}^d,\sup_{\mathbf{x}\in\mathbb{R}^d}|\mathbf{x}^\beta D^\alpha\psi(\mathbf{x})|<\infty\right\}. \tag{95}$$

The details of defining norms and the topology on $S(\mathbb{R}^d)$ is shown in (Melrose and Uhlmann, 2008, Chapter 1).

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

$$T_f:S(\mathbb{R}^d)\mapsto\mathbb{R},\ T_f(\psi)=\int_{\mathbb{R}^d}f\psi\mathrm{d}\mathbf{x}. \tag{96}$$

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

$$\mathcal{F}:S'(\mathbb{R}^d)\mapsto S'(\mathbb{R}^d),\ (\mathcal{F}T_f)(\psi)=\int_{\mathbb{R}^d}f\cdot\mathcal{G}\psi\mathrm{d}\mathbf{x}, \tag{97}$$

where $\mathcal{G}$ is the adjoint of $\mathcal{F}$. Details of Fourier transform on tempered distributions can be found in (Melrose and Uhlmann, 2008, Chapter 1.7).

Similarly the fractional Laplacian of a tempered distribution is defined by

$$(-\Delta)^s : S'(\mathbb{R}^d) \mapsto S'(\mathbb{R}^d),\ ((-\Delta)^s T_f)(\psi) = \int_{\mathbb{R}^d} f \cdot (-\Delta)^s \psi \mathrm{d}\mathbf{x}, \tag{98}$$

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 $\{0\}$. According to Folland (1999, Chapter 9), $\mathcal{F}f$ is a linear combination of δ (Dirac's Delta) and derivatives of δ.[6] Then f is a polynomial. Since f is Lipschitz continuous, we conclude that f is linear. ■

Lemma 30 *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 [Proof of Lemma 30] 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 (Ongie et al., 2020, Proposition 8).

Lemma 31 *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 [Proof of Lemma 31] 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

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

6. The k-th derivative of δ can be defined as a tempered distribution on the Schwartz test function ϕ by: $\delta^{(k)}(\phi) = (-1)^k \phi^{(k)}(0)$.

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 32 *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 $\psi := -\frac{1}{2(2\pi)^{d-1}}\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}$. 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 (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) $(-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\le p<d/(d-1)$. 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 [Proof of Lemma 32] Since $\|\boldsymbol{\mathcal{W}}\|_2$ and $\mathcal{B}$ both have finite second moments, we have

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

and

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

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

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

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 $\int_{\text{supp}(\zeta)} |\psi(\boldsymbol{V}, c)| \, \|\boldsymbol{V}\|_2 \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c$, which is finite. So $\tilde{h}(\mathbf{x})$ is Lipschitz continuous. Then we have

$$\begin{aligned}(-\Delta)^{(d+1)/2}\tilde{h} &= -(-\Delta)^{(d-1)/2} \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \psi(\boldsymbol{V}, c)\delta(\langle \boldsymbol{V}, \mathbf{x}\rangle - c) \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\ &= -(-\Delta)^{(d-1)/2} \int_{\mathbb{S}^{d-1}} \psi(\boldsymbol{V}, \langle \boldsymbol{V}, \mathbf{x}\rangle) \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V}) \\ &= -(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\psi\}.\end{aligned} \tag{99}$$

Since $(-\Delta)^{(d+1)/2}h \in L^p(\mathbb{R}^d),\ 1 \le p < d/(d-1)$, we can apply the inversion formula of the Radon transform (Solmon, 1987):

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

According to Lemma 29, we have that $h - \tilde{h}$ is linear, which gives the claim. ■

Lemma 32 immediately gives the following corollary:

Corollary 33 *If $\mathcal{R}\{(-\Delta)^{(d+1)/2}g\} \equiv \mathcal{R}\{(-\Delta)^{(d+1)/2}h\}$, and $(-\Delta)^{(d+1)/2}g, (-\Delta)^{(d+1)/2}h \in L^p(\mathbb{R}^d),\ 1 \le p < d/(d-1)$, then $g - h$ is linear.*

The next lemma shows that the minimizer $h(\mathbf{x})$ of problem (37) satisfies that $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}$ is compactly supported.

Lemma 34 *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 (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 [Proof of Lemma 34] Define $\psi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ by $\psi := -\frac{1}{2(2\pi)^{d-1}}\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}$. Then we construct the function $\overline{\psi} : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ as follows:

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

Since the Radon transform is even, we have that ψ and $\overline{\psi}$ are both even. Since h is the solution of (37), ψ satisfies all assumptions of Lemma 32. Then according to Lemma 32, $h(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \psi(\boldsymbol{V}, c)[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+ \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c + \langle \mathbf{u}, \mathbf{x}\rangle + v$. 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 $|c| \le R$, $\psi - \overline{\psi} = 0$. When $|c| > R$, $[\langle \boldsymbol{V}, \mathbf{x}\rangle - c]_+$ is linear with respect to x on $\{\mathbf{x} : \|\mathbf{x}\|_2 \le R\}$. It means that $\overline{h}(\mathbf{x}) - h(\mathbf{x})$ is linear on $\{\mathbf{x} : \|\mathbf{x}\|_2 \le R\}$. 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 $\{\mathbf{x} : \|\mathbf{x}\|_2 \le R\}$. Let $\widetilde{h}(\mathbf{x}) = \overline{h}(\mathbf{x}) + \langle \overline{\mathbf{u}}, \mathbf{x}\rangle + \overline{v}$. Since all training inputs satisfy $\|\mathbf{x}_i\| \le R$, we have that $\widetilde{h}(\mathbf{x})$ fits all training data. Similar to (99), we have that $\Delta\widetilde{h} = \mathcal{R}^*\{\overline{\psi}\}$. Since $\overline{\psi}$ has compact support, the inversion formula of

the Radon transform (Solmon, 1987) 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 $\mathcal{R}\{(-\Delta)^{(d+1)/2}\widetilde{h}\}(\boldsymbol{V}, c) = 0,\ \forall(\boldsymbol{V}, c) \notin \text{supp}(\zeta)$. 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 (Solmon, 1987, Lemma 4.1). The above argument shows that $\widetilde{h}$ satisfies all constrains of the problem (37). Since h is the solution of (37), we have $\int_{\text{supp}(\zeta)} \frac{(\psi(\boldsymbol{V},c))^2}{\zeta(\boldsymbol{V},c)}\, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \leq \int_{\text{supp}(\zeta)} \frac{\left(\overline{\psi}(\boldsymbol{V},c)\right)^2}{\zeta(\boldsymbol{V},c)}\, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c$. It means that $\psi(\boldsymbol{V}, c) = 0$ when $|c| > R$, which gives the claim. ■

The proof of Lemma 34 also applies to the optimization problem without the constraint $\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c) = 0,\ \forall(\boldsymbol{V}, c) \notin \text{supp}(\zeta)$. Then we have the following corollary.

Corollary 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 following optimization problem:*

$$\begin{aligned} \min_{h\in \text{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\text{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)}\, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\ \text{subject to} \quad & h(\mathbf{x}_j) = y_j, \quad j = 1, \ldots, M, \\ & (-\Delta)^{(d+1)/2}h \in L^p(\mathbb{R}^d),\ 1 \leq p < d/(d-1). \end{aligned} \tag{100}$$

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 $\mathbb{S}^{d-1} \times [0, R] \subset \text{supp}(\zeta)$, $h(\mathbf{x})$ is also the solution of (37).

Now we are ready to prove Theorem 16. We use the proof technique of Theorem 13 and (34).

Proof [Proof of Theorem 16] First, according to (28) and (34), if $(\boldsymbol{V}, c) \notin \text{supp}(\zeta)$, we have

$$\begin{aligned} &|\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot, (\overline{\gamma}, \mathbf{u}, v))\}(\boldsymbol{V}, c)| \\ =&|2(2\pi)^{d-1} \int_{\mathbb{R}} \gamma(u, \boldsymbol{V}, c) \cdot u\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u) \cdot p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})| \\ \leq&|2(2\pi)^{d-1} \int_{\mathbb{R}} \gamma^2(u, \boldsymbol{V}, c)\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u) \cdot \mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c) p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})| \\ =&0. \end{aligned} \tag{101}$$

By Lemma 31, 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 (37). Next, we prove that $g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))$ is the solution of (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 (37) is unique.

Suppose f_1, f_2 are two different functions in the feasible set of (37). Then $\mathcal{R}\{(-\Delta)^{(d+1)/2}f_1\}$ and $\mathcal{R}\{(-\Delta)^{(d+1)/2}f_2\}$ should be different. Otherwise, according to Corollary 33, $f_1 - f_2$ 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 (37) and $h(\mathbf{x})$ is different from $g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))$. Then by uniqueness of the solution,

$$L(h) < L(g(\mathbf{x}, (\overline{\gamma}, \overline{\mathbf{u}}, \overline{v}))). \tag{102}$$

Now our goal is to find a different $(\gamma, \mathbf{u}, v)$ with smaller cost in problem (26). Then $(\overline{\gamma}, \overline{\mathbf{u}}, \overline{v})$ is not the solution of (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 (32), we have $\Delta g(\cdot, (\gamma, \mathbf{0}, 0)) = \mathcal{R}^*\{\beta\}$ where β is defined in (28) and (31). Using Lemma 34, 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 (Solmon, 1987, Lemma 4.1), $(-\Delta)^{(d-1)/2}\mathcal{R}^*\{\beta\} \in L^p(\mathbb{R}^d),\ 1 \le p < d/(d-1)$, which means that $g(\cdot, (\gamma, \mathbf{0}, 0))$ satisfies the third constraint of the optimization problem (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 (28) and β (31), we have that κ and β are even. Then the odd part β^- of β is 0. According to (34),

$$\begin{aligned}
&\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, (\gamma, \mathbf{0}, 0)))\}(\boldsymbol{V}, c) \\
=& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \int_{\mathbb{R}^+} \gamma(u, \boldsymbol{V}, c) \cdot u \, \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}(u) \\
=& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \int_{\mathbb{R}^+} \frac{\mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c) \cdot u^2}{-2(2\pi)^{d-1}\zeta(\boldsymbol{V}, c)} \, \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}, \mathcal{C}=c}(u) \\
=& -2(2\pi)^{d-1} p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c) p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V}) \frac{\mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c) \cdot \mathbb{E}(\mathcal{U}^2 | \boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c)}{-2(2\pi)^{d-1}\zeta(\boldsymbol{V}, c)} \\
=& \mathcal{R}\{(-\Delta)^{(d+1)/2} h\}(\boldsymbol{V}, c), \quad (\boldsymbol{V}, c) \in \operatorname{supp}(\zeta).
\end{aligned} \tag{103}$$

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 (101). Then, according to (103), $\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, (\gamma, \mathbf{0}, 0)))\} \equiv \mathcal{R}\{(-\Delta)^{(d+1)/2} h\}$. According to Corollary 33, 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

(26). Next we compute the cost of $(\gamma, \mathbf{u}, v)$:

$$
\begin{aligned}
&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma^2(u,\boldsymbol{V},c)\ \mathrm{d}\nu(u,\boldsymbol{V},c)\\
=&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\cdot u}{-2(2\pi)^{d-1}\zeta(\boldsymbol{V},c)}\right)^2\ \mathrm{d}\nu(u,\boldsymbol{V},c)\\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)}{\zeta(\boldsymbol{V},c)}\right)^2 \left(\frac{\int_{\mathbb{R}^+} u^2\ \mathrm{d}\nu_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u)}{4(2\pi)^{2(d-1)}}\right)\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V},c)\\
=&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)}{\zeta(\boldsymbol{V},c)}\right)^2 \frac{\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c)p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})}{4(2\pi)^{2(d-1)}}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c\\
=&\frac{1}{4(2\pi)^{2(d-1)}}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\right)^2}{\zeta(\boldsymbol{V},c)}\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c\\
=&\frac{1}{4(2\pi)^{2(d-1)}}L(h).
\end{aligned}
\tag{104}
$$

According to (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{1}{\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c)}\ \mathrm{d}\nu_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V},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{105}
$$

where the first inequality is by the Cauchy-Schwarz inequality and the second inequality is by the Lemma 30 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

$$\begin{aligned}
&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \gamma^2(u,\boldsymbol{V},c)\ \mathrm{d}\nu(u,\boldsymbol{V},c)\\
=&\frac{1}{4(2\pi)^{2(d-1)}}L(h) \quad \text{(according to (104))}\\
<&\frac{1}{4(2\pi)^{2(d-1)}}L(g(\cdot,(\overline{\gamma},\mathbf{u},v))) \quad \text{(according to (102))}\\
\leq&\int_{\mathbb{R}^+\times\mathbb{S}^{d-1}\times\mathbb{R}} \frac{1}{4(2\pi)^{2(d-1)}}\overline{\gamma}^2(u,\boldsymbol{V},c)\ \mathrm{d}\nu(u,\boldsymbol{V},c) \quad \text{(according to (105))}.
\end{aligned}$$

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 (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 (26). This concludes the proof. ■

I.2 Proof of Theorem 7

Proof [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 35, we have that the variational problem (8) is equivalent to the following variational problem:

$$\begin{aligned}
\min_{h\in\mathrm{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \Big(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\Big)^2\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c\\
\text{subject to} \quad & h(\mathbf{x}_j)=y_j, \quad j=1,\ldots,M,\\
& (-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\leq p<d/(d-1).
\end{aligned} \tag{106}$$

The solution $h(\mathbf{x})$ of (106) 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:

$$\begin{aligned}
\min_{h\in\mathrm{Lip}(\mathbb{R}^d)\cap C(\mathbb{R}^d)} \quad & \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \Big(\mathcal{R}\{(-\Delta)^{(d+1)/2}h\}(\boldsymbol{V},c)\Big)^2\ \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c\\
\text{subject to} \quad & h(\mathbf{x}_j)=y_j, \quad j=1,\ldots,M.\\
& (-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\leq p<d/(d-1).
\end{aligned} \tag{107}$$

So it is sufficient to prove that if $h\in\mathrm{Lip}(\mathbb{R}^d)$ and $(-\Delta)^{(d+1)/2}h\in L^p(\mathbb{R}^d),\ 1\leq p<d/(d-1)$, 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 Solmon (1987, 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

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

■

I.3 Proof of Theorem 8

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

Lemma 36 *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 [Proof of Lemma 36| In order to prove the lemma, we need the following simple fact that

$$(-\Delta)\|\mathbf{x}\|^p = \tilde{C}_p \|\mathbf{x}\|^{p-2}, \tag{108}$$

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 (108) 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 constants.

If d is even, we apply (108) 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 fractional Laplacian can be written as a singular integral (Kwaśnicki, 2017), 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. Since the fractional Laplacian of a radially symmetric function is also radially symmetric, we have that $g(\mathbf{x})$ is radially symmetric, which means $g(\mathbf{x})$ only depends on $\|\mathbf{x}\|$. 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 $g(\mathbf{x}) = \|\mathbf{x}\|^{2-d} g(\frac{\mathbf{x}}{\|\mathbf{x}\|}) = C\|\mathbf{x}\|^{2-d}$ 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

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

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

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

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

The problem (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 37 *Suppose $h(\mathbf{x})$ is the solution of the variational problem* (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:*

$$\begin{aligned}
&\min_{\substack{f\in C(\mathbb{R}^d),\\ \mathbf{u}\in\mathbb{R}^d, v\in\mathbb{R}}} \quad \int_{\mathrm{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d-1)/2} f\}(\boldsymbol{V}, c)\right)^2}{\zeta(\boldsymbol{V}, c)} \, \mathrm{d}\sigma^{d-1}(\boldsymbol{V})\mathrm{d}c \\
&\text{subject to} \quad \int_{\mathbb{R}^d} f(\mathbf{s}) \left[\Gamma(\mathbf{x}_j - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}_j, \nabla\Gamma(-\mathbf{s})\rangle\right] \mathrm{d}\mathbf{s} + \langle \mathbf{u}, \mathbf{x}_j\rangle + v = y_j, \quad j = 1, \dots, M, \\
&\qquad \mathcal{R}\{(-\Delta)^{(d-1)/2} f\}(\boldsymbol{V}, c) = 0, \ \forall (\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta), \\
&\qquad (-\Delta)^{(d-1)/2} f \in L^p(\mathbb{R}^d), \ 1 \le p < d/(d-1), \\
&\qquad \sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |f(\mathbf{x})| < \infty,
\end{aligned} \tag{109}$$

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 [Proof of Lemma 37] First we prove that $\sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |-\Delta h| < \infty$. According to Lemma 34 and Lemma 32, we have $-\Delta h = \mathcal{R}^*\{\psi\}$, where ψ is tightly supported. Then (Solmon, 1987, 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

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

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

$$\Gamma(\mathbf{x} - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}, \nabla\Gamma(-\mathbf{s})\rangle = \mathbf{x}^T H_\Gamma(c\mathbf{x} - \mathbf{s})\mathbf{x} \text{ for some } c \in [0, 1], \tag{110}$$

where H_Γ is the Hessian matrix of Γ. Since

$$\frac{\partial\Gamma}{\partial s_i \partial s_j}(\mathbf{s}) = -\frac{\delta_{ij}\|\mathbf{s}\|^2 - d s_i s_j}{d V_d \|\mathbf{s}\|^{d+2}} = O(\|\mathbf{s}\|^{-d}), \tag{111}$$

where $\delta_{ij} = 1$ when $i = j$, and $\delta_{ij} = 0$ otherwise. According to (110) 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 $\|\mathbf{x}\| \ge 2$. The partial derivative of $\overline{g}$ is given by

$$\frac{\partial\overline{g}}{\partial x_i}(\mathbf{x}) = \int_{\mathbb{R}^d} -\frac{1}{d V_d} [-\Delta\overline{h}(\mathbf{s})] \left[\frac{x_i - s_i}{\|\mathbf{x} - \mathbf{s}\|^d} + \frac{s_i}{\|\mathbf{s}\|^d}\right] \mathrm{d}\mathbf{s}. \tag{112}$$

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 (112) is $O(\|\mathbf{s}\|^{d+1})$. So $|\frac{\partial\overline{g}}{\partial x_i}(\mathbf{x})|<\infty$. Next we estimate the integral (112) on $\mathbb{R}^d\backslash B(0,\|\mathbf{x}\|/2)$:

$$\begin{aligned}
&\left|\int_{\|\mathbf{s}\|>\|\mathbf{x}\|/2}[-\Delta\overline{h}(\mathbf{s})]\left[\frac{x_i-s_i}{\|\mathbf{x}-\mathbf{s}\|^d}+\frac{s_i}{\|\mathbf{s}\|^d}\right]\mathrm{d}\mathbf{s}\right|\\
\leq&\left|\int_{\|\mathbf{s}\|>\|\mathbf{x}\|/2}\frac{1}{\|\mathbf{s}\|}\left[\frac{x_i-s_i}{\|\mathbf{x}-\mathbf{s}\|^d}+\frac{s_i}{\|\mathbf{s}\|^d}\right]\mathrm{d}\mathbf{s}\right|\\
=&\left|\int_{\|\mathbf{s}\|>1/2}\frac{1}{\|\mathbf{s}\|}\left[\frac{x_i/\|\mathbf{x}\|-s_i}{\|\mathbf{x}/\|\mathbf{x}\|-\mathbf{s}\|^d}+\frac{s_i}{\|\mathbf{s}\|^d}\right]\mathrm{d}\mathbf{s}\right|\\
\leq&\max_{\|\widehat{\mathbf{x}}\|=1}\left|\int_{\|\mathbf{s}\|>1/2}\frac{1}{\|\mathbf{s}\|}\left[\frac{\widehat{x}_i-s_i}{\|\widehat{\mathbf{x}}-\mathbf{s}\|^d}+\frac{s_i}{\|\mathbf{s}\|^d}\right]\mathrm{d}\mathbf{s}\right|.
\end{aligned}\tag{113}$$

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 (112) on $B(0,\|\mathbf{x}\|/2)$:

$$\begin{aligned}
&\left|\int_{\|\mathbf{s}\|\leq\|\mathbf{x}\|/2}[-\Delta\overline{h}(\mathbf{s})]\left[\frac{x_i-s_i}{\|\mathbf{x}-\mathbf{s}\|^{d-1}}+\frac{s_i}{\|\mathbf{s}\|^d}\right]\mathrm{d}\mathbf{s}\right|\\
\leq&\left|\int_{\|\mathbf{s}\|\leq1}C\left[\frac{1}{\|\mathbf{x}-\mathbf{s}\|^d}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right|+\left|\int_{1<\|\mathbf{s}\|\leq\|\mathbf{x}\|/2}\frac{C}{\|\mathbf{s}\|}\left[\frac{1}{\|\mathbf{x}-\mathbf{s}\|^{d-1}}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right|\\
\leq&\left|\int_{\|\mathbf{s}\|\leq1}C\left[\frac{2^d}{\|\mathbf{x}\|^d}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right|+\left|\int_{1<\|\mathbf{s}\|\leq\|\mathbf{x}\|/2}\frac{C}{\|\mathbf{s}\|}\left[\frac{2^d}{\|\mathbf{x}\|^{d-1}}+\frac{1}{\|\mathbf{s}\|^{d-1}}\right]\mathrm{d}\mathbf{s}\right|\\
\leq&C_1+\frac{2^dC}{\|\mathbf{x}\|^{d-1}}\left|\int_{1<\|\mathbf{s}\|\leq\|\mathbf{x}\|/2}\frac{1}{\|\mathbf{s}\|}\mathrm{d}\mathbf{s}\right|+C\left|\int_{1<\|\mathbf{s}\|\leq\|\mathbf{x}\|/2}\frac{1}{\|\mathbf{s}\|^d}\mathrm{d}\mathbf{s}\right|\\
\leq&C_1+\frac{2^dC_2}{\|\mathbf{x}\|^{d-1}}\|\mathbf{x}\|^{d-1}+C_3\log\|\mathbf{x}\|\\
\leq&C_4+C_3\log\|\mathbf{x}\|,
\end{aligned}\tag{114}$$

where C_1, C_2, C_3 and C_4 are some constants. Combining (113) and (114) 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 $-\Delta(\overline{g}-\overline{h})\equiv0$. 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 29, 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 [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 (9). Then Lemma 37 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:

$$\begin{aligned}
\min_{\substack{f\in C(\mathbb{R}^d),\\ \mathbf{u}\in\mathbb{R}^d, v\in\mathbb{R}}} \quad & \int_{\mathbb{R}^d} \left((-\Delta)^{(d-1)/4} f(\mathbf{x})\right)^2 \,\mathrm{d}\mathbf{x} \\
\text{subject to} \quad & \int_{\mathbb{R}^d} f(\mathbf{s}) \left[\Gamma(\mathbf{x}_j - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}_j, \nabla\Gamma(-\mathbf{s})\rangle\right] \mathrm{d}\mathbf{s} + \langle \mathbf{u}, \mathbf{x}_j\rangle + v = y_j, \quad j = 1, \ldots, M \\
& (-\Delta)^{(d-1)/2} f \in L^p(\mathbb{R}^d),\ 1 \le p < d/(d-1) \\
& \sup_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\| \cdot |f(\mathbf{x})| < \infty,
\end{aligned} \tag{115}$$

Suppose that $f(\mathbf{x})$ is the solution of (115). 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 $G_j(f, \mathbf{u}, v) = \int_{\mathbb{R}^d} f(\mathbf{s}) \left[\Gamma(\mathbf{x}_j - \mathbf{s}) - \Gamma(-\mathbf{s}) - \langle \mathbf{x}_j, \nabla\Gamma(-\mathbf{s})\rangle\right] \mathrm{d}\mathbf{s} + \langle \mathbf{u}, \mathbf{x}_j\rangle + v$. For any function φ in Schwartz space $\mathcal{S}(\mathbb{R}^d)$,[7] $\tilde{\mathbf{u}} \in \mathbb{R}^d$ 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

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

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

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

which can be simplified to

$$\begin{cases}
(-\Delta)^{(d+1)/2} h(\mathbf{x}) = \sum_{j=1}^M \bar{\lambda}_j \left[\Gamma(\mathbf{x} - \mathbf{x}_j) - \Gamma(\mathbf{x})\right] \\
\sum_{j=1}^M \bar{\lambda}_j = 0 \\
\sum_{j=1}^M \bar{\lambda}_j \mathbf{x}_j = \mathbf{0}
\end{cases}. \tag{116}$$

7. 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.

According to Lemma 36 and Lemma 29, 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 (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 (116), the coefficients $\bar{\lambda}_j$, $\mathbf{u}$ and v satisfy (11). Now $h(\mathbf{x})$ satisfies the first-order optimality condition and fits all training data. Since the variational problem (107) 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 (107). Using (110), we have

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

According to (111), 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

$$\frac{\partial \|\cdot\|^3}{\partial x_i \partial x_j}(\mathbf{x}) = 3\delta_{ij}\|\mathbf{x}\| - 3\frac{x_i x_j}{\|\mathbf{x}\|} = O(\|\mathbf{x}\|), \tag{117}$$

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

I.4 Explicit Form of the Curvature Penalty Function

Proof [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 [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

$$\begin{aligned}
&p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c)\\
=&p_{\mathcal{C}|\boldsymbol{\mathcal{V}}=\boldsymbol{V}}(c)\ p_{\boldsymbol{\mathcal{V}}}(\boldsymbol{V})\cdot\int_{\mathbb{R}^+}u^2 p_{\mathcal{U}|\boldsymbol{\mathcal{V}}=\boldsymbol{V},\mathcal{C}=c}(u)\ \mathrm{d}u\\
=&\int_{\mathbb{R}^+}u^2 p_{\mathcal{U},\boldsymbol{\mathcal{V}},\mathcal{C}}(u,\boldsymbol{V},c)\ \mathrm{d}u\\
=&\int_{\mathbb{R}^+}u^{d+2} p_{\boldsymbol{\mathcal{W}},\mathcal{B}}(u\boldsymbol{V},-uc)\ \mathrm{d}u.
\end{aligned}\tag{118}$$

■

Proof [Proof of Theorem 19| Using (118), 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

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

■

Appendix 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 Parhi and Nowak (2019) 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 Parhi and Nowak (2019) show that minimizing the weight "norm"[8] 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 Parhi and Nowak (2019) 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 [Proof of Corollary 4] Use the same notation as in Section 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 (19) becomes:

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha_n^2(W^{(1)}, b)\ \mathrm{d}\mu_n(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha_n(W^{(1)}, b)\sigma(W^{(1)}x_j + b)\ \mathrm{d}\mu_n(W^{(1)}, b) = y_j, \quad j = 1, \dots, M. \end{aligned} \tag{119}$$

The limit of the problem (119) as width $n \to \infty$ is

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b)\ \mathrm{d}\mu(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)\sigma(W^{(1)}x_j + b)\ \mathrm{d}\mu(W^{(1)}, b) = y_j, \quad j = 1, \dots, M. \end{aligned} \tag{120}$$

As in Section 6, we can change the variables and relax the optimization problem (120) to

$$\begin{aligned} \min_{\substack{\gamma \in C(\mathbb{R}^2), \\ p \in C(\mathbb{R})}} \quad & \int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c)\ \mathrm{d}\nu(W^{(1)}, c) \\ \text{subject to} \quad & p(x_j) + \int_{\mathbb{R}^2} \gamma(W^{(1)}, c)\sigma\left(W^{(1)}(x_j - c)\right)\ \mathrm{d}\nu(W^{(1)}, c) = y_j, \quad j = 1, \dots, M \\ & \mathrm{L}\ p \equiv 0. \end{aligned} \tag{121}$$

If the activation function σ is ReLU, p is a linear function. Then (121) becomes the optimization problem (22). Define the output function g of the neural network by

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

8. 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.

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 (86), we have

$$\begin{aligned}
(\mathrm{L}g)(x,(\gamma,p)) &= \mathrm{L}\left(\int_{\mathbb{R}^2}\gamma(W^{(1)},c)\left|W^{(1)}\right|^k\sigma\left(\mathrm{sign}(W^{(1)})\cdot(x-c)\right)\,\mathrm{d}\nu(W^{(1)},c)\right)\\
&= \int_{\mathbb{R}^2}\gamma(W^{(1)},c)\left|W^{(1)}\right|^k\delta(x-c)\,\mathrm{d}\nu(W^{(1)},c)\\
&= \int_{\mathrm{supp}(\nu_{\mathcal{C}})}\left(\int_{\mathbb{R}}\gamma(W^{(1)},c)\left|W^{(1)}\right|^k\,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right)\delta(x-c)\,\mathrm{d}\nu_{\mathcal{C}}(c) \qquad (122)\\
&= \int_{\mathrm{supp}(\nu_{\mathcal{C}})}\left(\int_{\mathbb{R}}\gamma(W^{(1)},c)\left|W^{(1)}\right|^k\,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\right)\delta(x-c)p_{\mathcal{C}}(c)\mathrm{d}c\\
&= p_{\mathcal{C}}(x)\int_{\mathbb{R}}\gamma(W^{(1)},x)\left|W^{(1)}\right|^k\,\mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}).
\end{aligned}$$

Then similar to Theorem 13, we show that the solution of (121) in function space actually solves the following optimization problem:

$$\min_{h\in C^2(S)}\int_S\frac{((\mathrm{L}h)(x))^2}{\zeta(x)}\,\mathrm{d}x\quad\text{s.t.}\quad h(x_j)=y_j,\quad j=1,\dots,m, \qquad (123)$$

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 (123) and the technique used in proof of Theorem 1. ■

Appendix 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 (25) is close to the solution of training with the original training data (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 (20):

$$\begin{aligned}
&\min_{\alpha\in C(\mathbb{R}^d\times\mathbb{R})}\quad\int_{\mathbb{R}^d\times\mathbb{R}}\alpha^2(\mathbf{W}^{(1)},b)\,\mathrm{d}\mu(\mathbf{W}^{(1)},b)\\
&\text{subject to}\quad\int_{\mathbb{R}^d\times\mathbb{R}}\alpha(\mathbf{W}^{(1)},b)[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+\,\mathrm{d}\mu(\mathbf{W}^{(1)},b)=y_j,\quad j=1,\dots,M,
\end{aligned} \qquad (124)$$

and the linearly adjusted variational problem:

$$\begin{aligned}
&\min_{\substack{\alpha\in C(\mathbb{R}^d\times\mathbb{R}),\\ \mathbf{u}\in\mathbb{R}^d,v\in\mathbb{R}}}\int_{\mathbb{R}^d\times\mathbb{R}}\alpha^2(\mathbf{W}^{(1)},b)\,\mathrm{d}\mu(\mathbf{W}^{(1)},b)\\
&\text{subject to}\int_{\mathbb{R}^d\times\mathbb{R}}\alpha(\mathbf{W}^{(1)},b)[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+\,\mathrm{d}\mu(\mathbf{W}^{(1)},b)+\langle\mathbf{u},\mathbf{x}_j\rangle+v=y_j,\ j=1,\dots,M.
\end{aligned} \qquad (125)$$

In this paper, our main focus is on the variational problem (125), 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 (124) and (125), and thus show that to what extent the variational problem (5) and (8) in Theorem 1 and Theorem 6 describes the implicit bias of gradient descent on original training data.

Suppose the solution of problem (124) 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 (125) 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.,

$$\int_{\mathbb{R}^2} \alpha_s(\mathbf{W}^{(1)}, b)[\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = \langle \bar{\mathbf{u}}, \mathbf{x}\rangle + \bar{v}. \tag{126}$$

Then $\bar{\alpha}_2 + \alpha_s$ is a feasible solution of the problem (124). 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 38 *Suppose that the solution of the optimization problem* (124) *is* $\bar{\alpha}_1$ *and the solution of the optimization problem* (125) *is* $(\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v})$. *Suppose that* α_s *satisfies* (126). *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 [Proof of Theorem 38| Since $(\bar{\alpha}_2, \bar{u}, \bar{v})$ is the minimizer of (125), we have that $(\bar{\alpha}_1, 0, 0)$ is a feasible solution of (124) but not optimal, which means

$$\int_{\mathbb{R}^2} \bar{\alpha}_1^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \geq \int_{\mathbb{R}^2} \bar{\alpha}_2^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b). \tag{127}$$

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 (124), we have that there exist $\lambda_j \in \mathbb{R}$ such that

$$\alpha_1(\mathbf{W}^{(1)}, b) = \sum_{j=1}^{M} \lambda_j [\langle \mathbf{W}^{(1)}, \mathbf{x}\rangle + b]_+. \tag{128}$$

Since both $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$ are the feasible solutions of the problem (120),

$$\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \cdot [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) = 0, \quad j = 1, \ldots, M. \tag{129}$$

Using (128) and (129), we have

$$\begin{aligned}
&\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s)\bar{\alpha}_1 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \sum_{j=1}^{M} \lambda_j [\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\sum_{j=1}^{M} \lambda_j \int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \cdot [\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&0.
\end{aligned} \tag{130}$$

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

$$\begin{aligned}
&\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s)^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - (2\bar{\alpha}_2 + 2\alpha_s - \bar{\alpha}_1)\bar{\alpha}_1 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - \bar{\alpha}_1^2 + (2\bar{\alpha}_2 + 2\alpha_s - 2\bar{\alpha}_1)\bar{\alpha}_1 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \quad \text{(use (130))} \\
=&\int_{\mathbb{R}^2} (\bar{\alpha}_2^2 + 2\bar{\alpha}_2\alpha_s + \alpha_s^2) - \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
\leq&\int_{\mathbb{R}^2} (\bar{\alpha}_1^2 + 2\bar{\alpha}_2\alpha_s + \alpha_s^2) - \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \quad \text{(use (127))} \\
\leq&\int_{\mathbb{R}^2} 2\bar{\alpha}_2\alpha_s + \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
\leq&2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_2^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\
\leq&2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \quad \text{(use (127))}.
\end{aligned}$$

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

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

	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 1: Experimental settings.

■

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 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 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 $D_1 \coloneqq \frac{\int_{\mathbb{R}^2} \alpha_s^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b)}$. 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 (124) 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 (125) and get $(\bar{\alpha}_2, \bar{\mathbf{u}}, \bar{v})$. Then we need to find out α_s which satisfies (126). We can give 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 (126).

Next, we conduct some experiments to verify the above argument. We try three different settings and they are summarized in Table 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 12 we plot 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 (124) and (125) is small in practice.

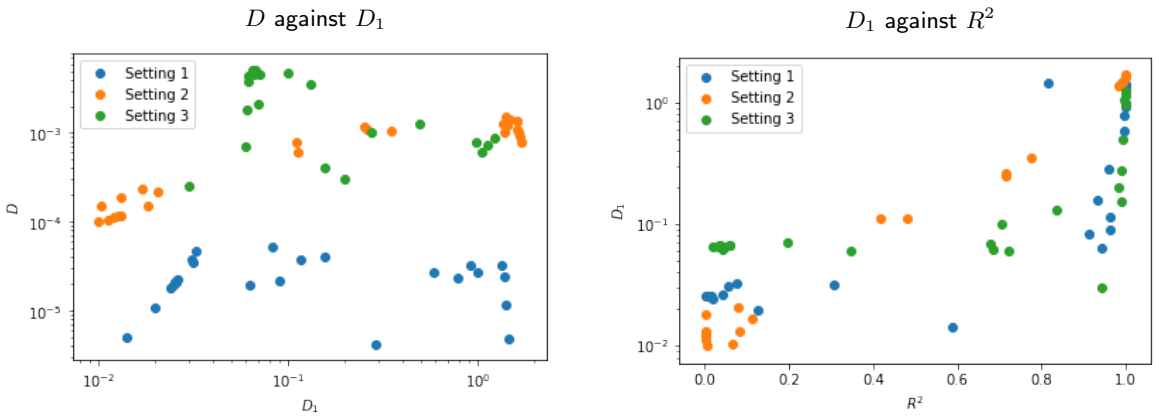

Figure 12: Scatter plots of D_1, D and R^2. The left panel is the scatter plot of D against D_1, which shows that D_1 is a very loose upper bound of D. Even when D_1 is around 1, D is still around 10^{-3}. The right panel is the scatter plot of D_1 against R^2, which shows that D_1 is small when training data are not highly linearly correlated and D_1 is large when training data are highly linearly correlated.

Appendix L. Neural Networks with Skip Connections

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

$$f(\mathbf{x}, \theta) = \sum_{i=1}^{n} W_i^{(2)} \phi(\langle \mathbf{W}_i^{(1)}, \mathbf{x} \rangle + b_i^{(1)}) + \langle \mathbf{u}, \mathbf{x} \rangle + v. \tag{131}$$

The skip connection corresponds to the term $\langle \mathbf{u}, \mathbf{x} \rangle$. The initializations of $\mathbf{W}_i^{(1)}, b_i^{(1)}, W_i^{(2)}$ are the same as (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

$$\begin{aligned}
&\mathbf{W}_0^{(1)} = \overline{\mathbf{W}}^{(1)}, && \mathbf{W}_{t+1}^{(1)} = \mathbf{W}_t^{(1)} - \eta_r \nabla_{\mathbf{W}^{(1)}} L^{\mathrm{lin}}(\theta_t) \\
&\mathbf{b}_0^{(1)} = \overline{\mathbf{b}}^{(1)}, && \mathbf{b}_{t+1}^{(1)} = \mathbf{b}_t^{(1)} - \eta_r \nabla_{\mathbf{b}^{(1)}} L^{\mathrm{lin}}(\theta_t) \\
&\mathbf{W}_0^{(2)} = \overline{\mathbf{W}}^{(2)}, && \mathbf{W}_{t+1}^{(2)} = \mathbf{W}_t^{(2)} - \eta_r \nabla_{\mathbf{W}^{(2)}} L^{\mathrm{lin}}(\theta_t) \\
&\mathbf{u}_0 = \mathbf{0}, && \mathbf{u}_{t+1} = \mathbf{u}_t - \eta_s \nabla_{\mathbf{u}} L^{\mathrm{lin}}(\theta_t) \\
&v_0 = 0, && v_{t+1} = v_t - \eta_s \nabla_v L^{\mathrm{lin}}(\theta_t)
\end{aligned} \tag{132}$$

Let $\widetilde{\omega}_t = \mathrm{vec}(\overline{\mathbf{W}}^{(1)}, \overline{\mathbf{b}}^{(1)}, \widetilde{\mathbf{W}}_t^{(2)}, \widetilde{\mathbf{u}}, \widetilde{v})$ 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

$$\begin{aligned} &\widetilde{\mathbf{W}}_0^{(2)} = \overline{\mathbf{W}}^{(2)}, \quad \widetilde{\mathbf{W}}_{t+1}^{(2)} = \widetilde{\mathbf{W}}_t^{(2)} - \eta_r \nabla_{\mathbf{W}^{(2)}} L^{\mathrm{lin}}(\widetilde{\omega}_t) \\ &\widetilde{\mathbf{u}}_0 = \mathbf{0}, \qquad\qquad\quad \widetilde{\mathbf{u}}_{t+1} = \widetilde{\mathbf{u}}_t - \eta_s \nabla_{\mathbf{u}} L^{\mathrm{lin}}(\widetilde{\omega}_t) \\ &\widetilde{v}_0 = 0, \qquad\qquad\quad \widetilde{v}_{t+1} = \widetilde{v}_t - \eta_s \nabla_{v} L^{\mathrm{lin}}(\widetilde{\omega}_t) \end{aligned} \tag{133}$$

Let $\Psi = \sum_{j=1}^{M} (\mathbf{x}_j, 1)^T (\mathbf{x}_j, 1)$. Using the similar argument in Section 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 44 with some modifications and show that gradient descent training of the output weights (133) 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:

$$\begin{aligned} \min_{\mathbf{W}^{(2)}} \quad & \frac{1}{\eta_r} \|\mathbf{W}^{(2)} - \overline{\mathbf{W}}^{(2)}\|_2^2 + \frac{1}{\eta_s} \left(\|\mathbf{u}\|_2^2 + v^2 \right) \\ \text{s.t.} \quad & \sum_{i=1}^{n} (W_i^{(2)} - \overline{W}_i^{(2)}) [\langle \overline{\mathbf{W}}_i^{(1)}, \mathbf{x}_j \rangle + \overline{b}_i^{(1)}]_+ + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j - f(\mathbf{x}_j, \theta_0), \ j = 1, \dots, M. \end{aligned} \tag{134}$$

Similar to Section 5, we let $f^{\mathrm{lin}}(\mathbf{x}, \theta_0) \equiv 0$ 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., $\mu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A \left((\overline{\mathbf{W}}_i^{(1)}, \overline{b}_i^{(1)}) \right)$, 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 (134) with ASI can be rewritten as

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha_n^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) + \frac{n\eta_r}{\eta_s} \left(\|\mathbf{u}\|_2^2 + v^2 \right) \\ \text{s.t.} \quad & \int_{\mathbb{R}^2} \alpha_n(\mathbf{W}^{(1)}, b) [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu_n(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j, \ j = 1, \dots, M. \end{aligned} \tag{135}$$

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 (135) as follows:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(\mathbf{W}^{(1)}, b) \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) \\ \text{s.t.} \quad & \int_{\mathbb{R}^2} \alpha(\mathbf{W}^{(1)}, b) [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ \, \mathrm{d}\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j, \ j = 1, \dots, M. \end{aligned} \tag{136}$$

Using that μ_n weakly converges to μ, we show that in fact the solution of problem (135) converges to the solution of (136) in Theorem 39.

Theorem 39 (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* (135)*. Let* $(\overline{\alpha}, \overline{\mathbf{u}}, \overline{v})$ *be the solution of* (136)*. 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 39 is provided at the end of the section. In Section 6 and Section 7, we show that the optimization problem (136) is equivalent to (24) in the univariate case and equivalent to (37) in the multivariate case. From this we immediately obtain our main theorems for networks with skip connections without adjusting the training data, namely the following Theorem 40 and Theorem 41.

Theorem 40 (Implicit bias of networks with skip connections, univariate) *Consider a two-layer feedforward network with skip connections* (131)*. Assume parameter initialization* (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* (132) *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:*

$$\begin{aligned} \min_{g\in C^2(S)} \quad & \int_S \frac{1}{\zeta(x)} (g''(x) - f''(x, \theta_0))^2 \, \mathrm{d}x \\ \text{subject to} \quad & g(x_j) = y_j, \quad j = 1, \ldots, M. \end{aligned} \tag{137}$$

Theorem 41 (Implicit bias of networks with skip connections, multivariate) *Consider the same network settings as in Theorem 40 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* (132) *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:

$$\begin{aligned}\min_{g\in \mathrm{Lip}(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \frac{\left(\mathcal{R}\{(-\Delta)^{(d+1)/2}(g-f(\cdot,\theta_0))\}(\boldsymbol{V},c)\right)^2}{\zeta(\boldsymbol{V},c)}\,\mathrm{d}\boldsymbol{V}\mathrm{d}c \\ \text{subject to} \quad & g(\mathbf{x}_j)=y_j, \quad j=1,\dots,M \\ & \mathcal{R}\{(-\Delta)^{(d+1)/2}(g-f(\cdot,\theta_0))\}(\boldsymbol{V},c)=0, \quad (\boldsymbol{V},c)\notin \mathrm{supp}(\zeta) \\ & (-\Delta)^{(d+1)/2}(g-f(\cdot,\theta_0))\in L^p(\mathbb{R}^d),\ 1\le p<d/(d-1).\end{aligned} \tag{138}$$

Proof [Proof of Theorem 39] The Lagrangian of problem (135) is

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

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

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

$$\frac{2n\eta_r}{\eta_s}\mathbf{u}_n+\sum_{j=1}^M\lambda_j^{(n)}\mathbf{x}_j=0$$

$$\frac{2n\eta_r}{\eta_s}v_n+\sum_{j=1}^M\lambda_j^{(n)}=0.$$

Since only function values on $(\mathbf{W}_i^{(1)},b_i)_{i=1}^M$ are taken into account in problem (135), we can let

$$\overline{\alpha}_n(\mathbf{W}^{(1)},b)=-\frac{1}{2}\sum_{j=1}^M\lambda_j^{(n)}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+ \quad \forall(\mathbf{W}^{(1)},b)\in\mathbb{R}^{d+1} \tag{139}$$

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:

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

$$\sum_{j=1}^M\lambda_j^{(n)}\mathbf{x}_j+\frac{2n\eta_r}{\eta_s}\mathbf{u}_n=0$$

$$\sum_{j=1}^M\lambda_j^{(n)}+\frac{2n\eta_r}{\eta_s}v_n=0. \tag{140}$$

Similarly, the Lagrangian of problem (136) 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

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

Then we get

$$\overline{\alpha}(\mathbf{W}^{(1)},b)=-\frac{1}{2}\sum_{j=1}^{M}\lambda_j[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+\quad\forall(\mathbf{W}^{(1)},b)\in\mathbb{R}^2.\tag{141}$$

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

$$\begin{aligned}-\frac{1}{2}\sum_{j=1}^{M}\lambda_j\int_{\mathbb{R}^2}[\langle\mathbf{W}^{(1)},\mathbf{x}_j\rangle+b]_+[\langle\mathbf{W}^{(1)},\mathbf{x}_i\rangle+b]_+\ \mathrm{d}\mu(\mathbf{W}^{(1)},b)+\langle\mathbf{u},\mathbf{x}_i\rangle+v=y_i,\ i=1,\dots,M\\ \sum_{j=1}^{M}\lambda_j\mathbf{x}_j+0\cdot\mathbf{u}=0\\ \sum_{j=1}^{M}\lambda_j+0\cdot v=0\\ .\end{aligned}\tag{142}$$

Compare (140) and (142). 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 $\forall i=1,\dots,M,\ \forall j=1,\dots,M$,

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

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 (140) converge to coefficients of (142) at the rate of $O_p(n^{-1/2})$, then we get

$$|\lambda_j^n - \lambda_j| = O_p(n^{-1/2}), \quad j = 1, \ldots, M. \tag{143}$$

Compare (139) and (141). Given $(\mathbf{W}^{(1)}, b)$, we have

$$|\overline{\alpha}_n(\mathbf{W}^{(1)}, b) - \overline{\alpha}(\mathbf{W}^{(1)}, b)| = O_p(n^{-1/2}). \tag{144}$$

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 (84) in the proof of Theorem 12 in Appendix 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

$$\sup_{\mathbf{x}\in D} |g_n(\mathbf{x}, (\overline{\alpha}, \overline{\mathbf{u}}, \overline{v})) - g(\mathbf{x}, (\overline{\alpha}, \overline{\mathbf{u}}, \overline{v}))| = O_p(n^{-1/2}). \tag{145}$$

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 (143),

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

Combined with (145), 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. ■

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

In this section we demonstrate that NTK norm minimization (Zhang et al., 2020), which characterizes the implicit bias of training a linearized model by gradient descent, is equivalent to our characterization in Section 5 and Section 6. For simplicity, we only consider univariate regression in this section. Following Jacot et al. (2018), Zhang et al. (2020) 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

$$\tilde{\Theta}(x_1, x_2) = \int_{\mathbb{R}^2} [W^{(1)} x_1 + b^{(1)}]_+ [W^{(1)} x_2 + b^{(1)}]_+ \, \mathrm{d}\mu(W^{(1)}, b). \tag{146}$$

Equivalently, using the notation in Section 6, we have

$$\tilde{\Theta}(x_1, x_2) = \int_{\mathbb{R}^2} [W^{(1)}(x_1 - c)]_+ [W^{(1)}(x_2 - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c). \tag{147}$$

Next, Zhang et al. (2020) 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:

$$\text{(i)} \quad \forall x \in S, \tilde{\Theta}(\cdot, x) \in \mathcal{H}_{\tilde{\Theta}}(S); \tag{148}$$

$$\text{(ii)} \quad \forall x \in S, \forall f \in \mathcal{H}_{\tilde{\Theta}}, \langle f(\cdot), \tilde{\Theta}(\cdot, x)\rangle_{\tilde{\Theta}} = f(x); \tag{149}$$

$$\text{(iii)} \quad \forall x, y \in S, \langle \tilde{\Theta}(\cdot, x), \tilde{\Theta}(\cdot, y)\rangle_{\tilde{\Theta}} = \tilde{\Theta}(x, y). \tag{150}$$

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, Zhang et al. (2020) prove that $f^{\mathrm{lin}}(x, \widetilde{\omega}_\infty)$ (defined in Section 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

$$\min_{g\in\mathcal{H}_{\tilde{\Theta}}(S)} \|g\|_{\tilde{\Theta}} \quad \text{s.t. } g(x_j)=y_j,\ j=1,\dots,M. \tag{151}$$

In Section 5, we show that $f^{\text{lin}}(x,\widetilde{\omega}_\infty)$ is the solution of the optimization problem (19) in function space. As width n tends to infinity, the optimization problem (19) becomes (20), which we repeat below:

$$\begin{aligned} \min_{\alpha\in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(W^{(1)},b)\ \mathrm{d}\mu(W^{(1)},b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha(W^{(1)},b)[W^{(1)}x_j+b]_+\ \mathrm{d}\mu(W^{(1)},b)=y_j, \quad j=1,\dots,M. \end{aligned} \tag{152}$$

Since optimization problems (151) and (152) 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 42 (Equivalence of our variational problem and NTK norm minimization) *Assume that optimization problems* (151) *and* (152) *are both feasible. Suppose* $\overline{\alpha}$ *is the solution of* (152)*, and consider the corresponding output function:*

$$\overline{g}(x)=\int_{\mathbb{R}^2} \overline{\alpha}(W^{(1)},b)[W^{(1)}x+b]_+\ \mathrm{d}\mu(W^{(1)},b). \tag{153}$$

Then $\overline{g}(x)$ *restricted on* S *is the solution of the optimization problem* (151)*.*

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 [Proof of Theorem 42| Since $\overline{\alpha}(W^{(1)},b)$ is the solution of (152), according to (79) 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 $\lambda_j, j=1,\dots,M$. Then we write $\overline{\alpha}(W^{(1)},b)$ in the following form:

$$\overline{\alpha}(W^{(1)},b)=\int_S h(x)[W^{(1)}x+b]_+\mathrm{d}x, \tag{154}$$

where $h(x)$ can be a combination of Dirac delta functions. Then substitute (154) into the expression of $\overline{g}(x)$ (153) to obtain

$$\begin{aligned} \overline{g}(x) &= \int_{\mathbb{R}^2\times S} h(\tilde{x})[W^{(1)}\tilde{x}+b]_+[W^{(1)}x+b]_+\ \mathrm{d}\mu(W^{(1)},b)\mathrm{d}\tilde{x} \\ &= \int_S h(\tilde{x})\tilde{\Theta}(x,\tilde{x})\mathrm{d}\tilde{x}, \end{aligned} \tag{155}$$

where we use the expression of the NTK in equation (146). Then we get

$$\begin{aligned}\langle g(x), g(x)\rangle_{\tilde{\Theta}} &= \langle g(x), \int_S h(\tilde{x})\tilde{\Theta}(x,\tilde{x})\mathrm{d}\tilde{x}\rangle_{\tilde{\Theta}}\mathrm{d}\tilde{x} \\ &= \int_S h(\tilde{x})\langle g(x), \tilde{\Theta}(x,\tilde{x})\rangle_{\tilde{\Theta}}\mathrm{d}\tilde{x} \\ &= \int_S h(\tilde{x})g(\tilde{x})\mathrm{d}\tilde{x} \quad \text{(here we use the property of RKHS norm (149))} \\ &= \int_{S\times S} h(\tilde{x})h(\bar{x})\tilde{\Theta}(\tilde{x},\bar{x})\mathrm{d}\tilde{x}\mathrm{d}\bar{x} \quad \text{(use (155))}.\end{aligned} \tag{156}$$

On the other hand, using (154), the objective of (152) becomes

$$\begin{aligned}&\int_{S^2} \overline{\alpha}^2(W^{(1)},b)\ \mathrm{d}\mu(W^{(1)},b) \\ =&\int_{S\times S\times\mathbb{R}^2} h(\tilde{x})[W^{(1)}\tilde{x}+b]_+h(\bar{x})[W^{(1)}\bar{x}+b]_+\ \mathrm{d}\tilde{x}\mathrm{d}\bar{x}\mathrm{d}\mu(W^{(1)},b) \\ =&\int_{S\times S} h(\tilde{x})h(\bar{x})\int_{\mathbb{R}^2}[W^{(1)}\tilde{x}+b]_+[W^{(1)}\bar{x}+b]_+\mathrm{d}\mu(W^{(1)},b)\ \mathrm{d}\tilde{x}\mathrm{d}\bar{x} \\ =&\int_{S\times S} h(\tilde{x})h(\bar{x})\tilde{\Theta}(\bar{x},\tilde{x})\ \mathrm{d}\tilde{x}\mathrm{d}\bar{x} \quad \text{(use (146))}.\end{aligned} \tag{157}$$

Comparing (156) and (157), we have that optimization problems (151) and (152) are equivalent if $\alpha(W^{(1)},b)$ has the form (154) and $g(x)$ has the form (155). 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 (154) such that $g(x)$ is expressed in the form of (155). In this sense we show that optimization problems (151) and (152) are equivalent. ■

In Section 6, we relax the optimization problem (21) to (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 (151) to the following problem:

$$\min_{g\in\mathcal{H}_{\tilde{\Theta}}(S),u\in\mathbb{R},v\in\mathbb{R}} \|g-ux-v\|_{\tilde{\Theta}} \quad \text{s.t. } g(x_j)=y_j,\ j=1,\dots,M. \tag{158}$$

Then the optimization problems (22) and (158) are equivalent. Theorem 13 shows that (22) and (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 (158) and (24) are equivalent, which means that

$$\min_{u\in\mathbb{R},v\in\mathbb{R}} \|g-ux-v\|_{\tilde{\Theta}} = \int_S \frac{(g''(x))^2}{\zeta(x)}\ \mathrm{d}x, \quad \forall g\in\mathcal{H}_{\tilde{\Theta}}(S). \tag{159}$$

Next, we directly prove the above equation (159). 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 (159) which we want to prove in the following theorem:

Theorem 43 (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:*

$$\|h\|^2_{\tilde{\Theta}} = \int_S \frac{(h''(x))^2}{\zeta(x)}\, \mathrm{d}x, \quad \forall h \in G. \tag{160}$$

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 [Proof of Theorem 43| 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

$$\begin{aligned}
\langle h, \tilde{\Theta}_{x,G}\rangle_{\tilde{\Theta}} &= \langle h, \tilde{\Theta}_x - ux - v\rangle_{\tilde{\Theta}} \\
&= \langle h, \tilde{\Theta}_x\rangle_{\tilde{\Theta}} - u\langle h, x\rangle_{\tilde{\Theta}} - v\langle h, 1\rangle_{\tilde{\Theta}} \\
&= \langle h, \tilde{\Theta}_x\rangle_{\tilde{\Theta}} \\
&= h(x) \quad \text{(use the reproducing property of the kernel (149)).}
\end{aligned} \tag{161}$$

Next, using the notation from Section 6 we have

$$\begin{aligned}
\tilde{\Theta}''_{x,G}(y) &= (\tilde{\Theta}_x(y) - uy - v)'' = \tilde{\Theta}_x(y)'' = \frac{\partial^2}{\partial y^2}\tilde{\Theta}(x, y) \\
&= \frac{\partial^2}{\partial y^2}\int_{\mathbb{R}^2} [W^{(1)}(x - c)]_+ [W^{(1)}(y - c)]_+ \, \mathrm{d}\nu(W^{(1)}, c) \quad \text{(use (147))} \\
&= \frac{\partial^2}{\partial y^2}\int_{\mathbb{R}^2} (W^{(1)})^2 [\mathrm{sign}(W^{(1)})(x - c)]_+ [\mathrm{sign}(W^{(1)})(y - c)]_+ \, \mathrm{d}\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})\mathrm{d}\nu_{\mathcal{C}}(c) \\
&= \frac{\partial^2}{\partial y^2}\int_{\mathbb{R}} \big(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} \geq 0)|\mathcal{C} = c)[x - c]_+[y - c]_+ \\
&\qquad + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} < 0)|\mathcal{C} = c)[c - x]_+[c - y]_+\big)\, p_{\mathcal{C}}(c)\, \mathrm{d}c \\
&= \int_{\mathbb{R}} \Bigg(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} \geq 0)|\mathcal{C} = c)[x - c]_+ \frac{\partial^2}{\partial y^2}[y - c]_+ \\
&\qquad + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} < 0)|\mathcal{C} = c)[c - x]_+ \frac{\partial^2}{\partial y^2}[c - y]_+\Bigg)\, p_{\mathcal{C}}(c)\, \mathrm{d}c \\
&= \int_{\mathbb{R}} \big(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} \geq 0)|\mathcal{C} = c)[x - c]_+\delta(y - c) \\
&\qquad + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} < 0)|\mathcal{C} = c)[c - x]_+\delta(y - c)\big)\, p_{\mathcal{C}}(c)\, \mathrm{d}c \\
&= \big(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} \geq 0)|\mathcal{C} = y)[x - y]_+ + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W} < 0)|\mathcal{C} = y)[y - x]_+\big)\, p_{\mathcal{C}}(y).
\end{aligned}$$

Then we have

$$\begin{aligned}
&\int_S \frac{h''(y)\tilde{\Theta}''_{x,G}(y)}{\zeta(y)}\,\mathrm{d}y \\
=&\int_S \frac{h''(y)\left(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)[x-y]_+ + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)[y-x]_+\right)p_{\mathcal{C}}(y)}{\zeta(y)}\,\mathrm{d}y \\
=&\int_S \frac{h''(y)\left(\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)[x-y]_+ + \mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)[y-x]_+\right)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}\,\mathrm{d}y \\
=&\int_S \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}\geq 0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)[x-y]_+ + \frac{\mathbb{E}(\mathcal{W}^2\mathbb{1}(\mathcal{W}<0)|\mathcal{C}=y)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=y)}h''(y)[y-x]_+\,\mathrm{d}y.
\end{aligned}$$

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. ■

Appendix 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.

N.1 Regularized Regression and Early Stopping

Bishop (1995) 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 $X = [\mathbf{x}_1, \dots, \mathbf{x}_M]^T$ is the matrix of training inputs, $\mathbf{y} = [y_1, \dots, y_M]^T$ 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 $\tilde{\mathbf{w}} = \sum_j \tilde{w}_j \mathbf{u}_j$. Bishop (1995) shows that

$$\tilde{w}_j = \frac{\lambda_j}{\lambda_j + c} w_j^*. \tag{162}$$

For early stopping, consider the gradient descent on $E(\mathbf{w})$ with zero initial weight vector:

$$\begin{aligned} \mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)} - \eta \nabla E \\ &= \mathbf{w}^{(\tau-1)} - \eta H (\mathbf{w}^{(\tau-1)} - \mathbf{w}^*), \\ \mathbf{w}^{(0)} &= \mathbf{0}. \end{aligned}$$

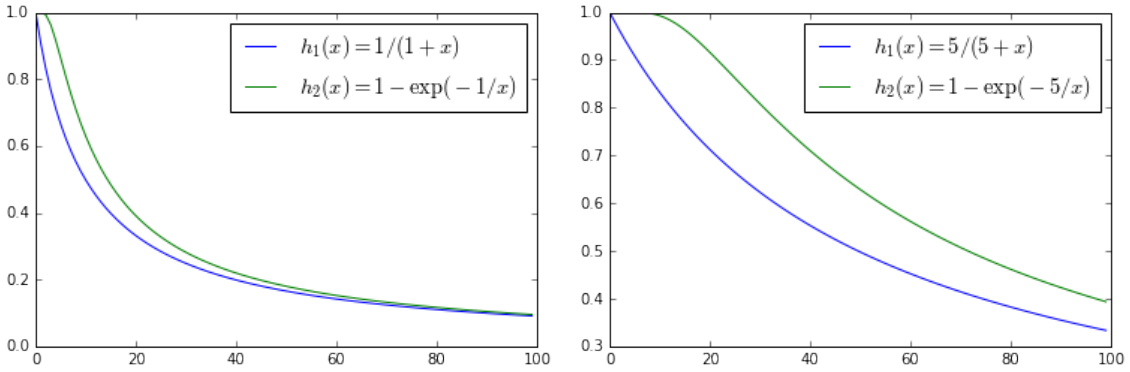

Figure 13: Plot of functions $h_1(x)$ and $h_2(x)$. The left panel plots the two function when $\lambda_j = 1$. The right panel plots the two function when $\lambda_j = 5$.

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

$$w_j^{(\tau)} = (1 - e^{-\eta\tau\lambda_j}) w_j^*. \tag{163}$$

From (162) and (163), Bishop (1995) 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 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 (162) and (163), 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 (17) here:

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

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:

$$\min_{W^{(2)}} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} (W_i^{(2)} - \overline{W}_i^{(2)})[W_i^{(1)} x_j + b_i]_+ - (y_j - f(x_j, \theta_0)) \right)^2 + c\|W^{(2)} - \overline{W}\|_2^2, \quad (164)$$

where $c = \frac{1}{n_n t}$. Using the same approach and notation as in Section 5, the optimization problem (164) is equivalent to

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \quad & \sum_{j=1}^{M} \left(\int_{\mathbb{R}^2} \alpha_n(W^{(1)}, b)[W^{(1)} x_j + b]_+ \, \mathrm{d}\mu_n(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{n\eta_n t} \int_{\mathbb{R}^2} \alpha_n^2(W^{(1)}, b) \, \mathrm{d}\mu_n(W^{(1)}, b), \end{aligned} \quad (165)$$

where we use the ASI trick (see Appendix B.2). Here (165) has an extra factor $\frac{1}{n}$ compared to (164). 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 (165) as the width n tends to infinity is:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \sum_{j=1}^{M} \left(\int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)} x_j + b]_+ \, \mathrm{d}\mu(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{\bar{\eta} t} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) \, \mathrm{d}\mu(W^{(1)}, b). \end{aligned} \quad (166)$$

Following the same reasoning of Section 6, we relax the optimization problem (166) to the following one:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2), u \in \mathbb{R}, v \in \mathbb{R}} \quad & \sum_{j=1}^{M} \left(u x_j + v + \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)} x_j + b]_+ \, \mathrm{d}\mu(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{\bar{\eta} t} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) \, \mathrm{d}\mu(W^{(1)}, b). \end{aligned} \quad (167)$$

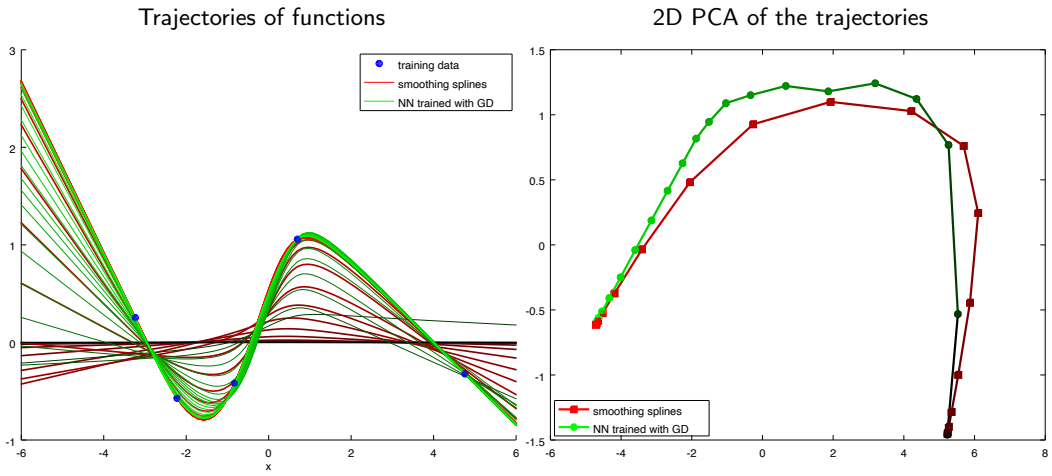

Figure 14: Trajectories of functions obtained by gradient descent training a neural network and by smoothing splines of the training data with decreasing regularization strength (from dark to bright). The left panel plots 20 functions along each trajectory. The right panel shows the same functions in 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.

Using the same technique and notation as in Theorem 13, we can prove that the solution of (167) actually solves the following optimization problem:

$$\min_{h \in C^2(S)} \sum_{j=1}^{M} [h(x_j) - y_j]^2 + \frac{1}{\bar{\eta} t} \int_S \frac{(h''(x))^2}{\zeta(x)} \, \mathrm{d}x. \tag{168}$$

Then in order to study the trajectory of gradient descent, we can study the optimization problem (168) with varying t. Figure 14 illustrates smoothing spline and gradient descent trajectories. The solution of (168) 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 the solution of (168) 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 44 (Spectral bias) *We have thus that the gradient descent optimization trajectory 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 Rahaman et al. (2019). The spectral bias is that lower frequencies are learned first.

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 (168) becomes the following optimization problem:

$$\min_{h \in C^2(S)} \sum_{j=1}^{M} [h(x_j) - y_j]^2 + \lambda \int_S (h''(x))^2 \, \mathrm{d}x. \tag{169}$$

German (2001) gives the explicit form of the minimizer $\hat{h}$ of (169), 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

$$(\hat{h}(x_1), \ldots, \hat{h}(x_M))^\top = (I + \lambda A)^{-1} Y. \tag{170}$$

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 $j = i$. German (2001) gives 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.

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 (168) can be written as

$$\min_{h \in W_2(S)} \frac{1}{M} \sum_{i=1}^{M} [h(x_j) - y_j]^2 + \int_S \rho(x) (h''(x))^2 \, \mathrm{d}x, \tag{171}$$

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. Abramovich and Steinberg (1996); Pintore et al. (2006) give the solution of (171) explicitly, which is called a spatially adaptive smoothing spline.

According to Pintore et al. (2006), 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. Pintore et al. (2006) define $B_2(S)$ in this way just for simplicity. Then the kernel of the space $W_0^2(S)$ is given by

$$K_\rho(x_1, x_2) = \int_S \rho(u)^{-1}[x_1 - u]_+[x_2 - u]_+ \mathrm{d}u. \tag{172}$$

Then the minimizer $\hat{h}$ of (171) is given by

$$\hat{h}(x) = \sum_{j=1}^{M} c_j K_\rho(x_j, x) + a + bx. \tag{173}$$

Now define the $M \times M$ matrix

$$\Sigma_\rho = \{K_\rho(x_i, x_j)\}_{i,j=1,\ldots,M}, \tag{174}$$

and the $M \times 2$ matrix

$$T = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_M \end{bmatrix}. \tag{175}$$

Denote the vector of coefficients $\mathbf{c} = (c_1, \ldots, c_M)^T$ and the vector of output values $\mathbf{y} = (y_1, \ldots, y_M)^T$. Then the coefficients in (173) satisfy the following conditions:

$$\Sigma_\rho \left[(\Sigma_\rho + MI)\mathbf{c} + T \begin{pmatrix} a \\ b \end{pmatrix} \right] = \Sigma_\rho \mathbf{y} \quad \text{and} \quad T^\top \left[\Sigma_\rho \mathbf{c} + T \begin{pmatrix} a \\ b \end{pmatrix} \right] = T^\top \mathbf{y}. \tag{176}$$

After solving for (176), we get the values of $\mathbf{c}$, a and b. Plug them into (173), then we get the exact form of the minimizer of (171).

Appendix O. Solution to the Variational Problems for Univariate Regression after Training

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 (5) with ASI is solved by the cubic spline interpolation of the data with natural boundary conditions (Ahlberg et al., 1967).

Theorem 45 (Ahlberg et al. 1967) *For training samples $\{(x_i, y_i)\}_{i=1}^M$, suppose $x_j \in S$, $j = 1, \ldots, M$. Then cubic spline interpolation of data $\{(x_i, y_i)\}_{i=1}^M$ with natural boundary condition is the solution of*

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

As already mentioned in Appendix 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 $S_j(x) = a_j + b_j x + c_j x^2 + d_j x^3$, $j = 1, \dots, M-1$. These $(M-1)$ pieces satisfy equations $S_i(x_i) = y_i$, $S_i(x_{i+1}) = y_{i+1}$, $i = 1, \dots, M-1$ and $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$, $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1})$, $i = 1, \dots, M-2$, 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.

O.2 Spatially Adaptive Interpolating Splines

In the case that ζ is not constant, we can still give the form of the solution to the variational problem (5) with ASI by using the result in Appendix N. We multiply by a coefficient λ the regularization term in the optimization problem (171) and choose $\rho(x) = \frac{1}{\zeta(x)}$. Then we get

$$\min_{h \in W_2(S)} \frac{1}{M} \sum_{i=1}^{M} [h(x_j) - y_j]^2 + \lambda \int_S \frac{1}{\zeta(x)} (h''(x))^2 \, \mathrm{d}x. \tag{177}$$

As $\lambda \to 0$, the minimizer of (177) converges to the solution of the following optimization problem:

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

which is the variational problem (5) with ASI. According to Appendix N, the solution of (177) is given by:

$$\hat{h}^{(\lambda)}(x) = \sum_{j=1}^{M} c_j^{(\lambda)} K_{\frac{\lambda}{\zeta}}(x_j, x) + a^{(\lambda)} + b^{(\lambda)} x. \tag{178}$$

And the vector $\mathbf{c}^{(\lambda)} = (c_1^{(\lambda)}, \dots, c_M^{(\lambda)})^T$, $a^{(\lambda)}$ and $b^{(\lambda)}$ satisfy the following conditions:

$$\Sigma_{\frac{\lambda}{\zeta}} \left[(\Sigma_{\frac{\lambda}{\zeta}} + MI)\mathbf{c}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = \Sigma_{\frac{\lambda}{\zeta}} \mathbf{y} \quad \text{and} \quad T^\top \left[\Sigma_{\frac{\lambda}{\zeta}} \mathbf{c}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = T^\top \mathbf{y}, \tag{179}$$

where $K_{\frac{\lambda}{\zeta}}$, $\Sigma_{\frac{\lambda}{\zeta}}$ and T are defined in (172), (174) and (175). Next we show that $K_{\frac{\lambda}{\zeta}}$ is inversely proportional to λ:

$$\begin{aligned} K_{\frac{\lambda}{\zeta}}(x_1, x_2) &= \int_S \left(\frac{\lambda}{\zeta} \right)^{-1} [x_1 - u]_+ [x_2 - u]_+ \mathrm{d}u \\ &= \lambda^{-1} \int_S \left(\frac{1}{\zeta} \right)^{-1} [x_1 - u]_+ [x_2 - u]_+ \mathrm{d}u \\ &= \lambda^{-1} K_{\frac{1}{\zeta}}(x_1, x_2). \end{aligned} \tag{180}$$

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 (178) and (179) as

$$\hat{h}^{(\lambda)}(x) = \sum_{j=1}^{M} \bar{c}_j^{(\lambda)} K_{\frac{1}{\zeta}}(x_j, x) + a^{(\lambda)} + b^{(\lambda)} x, \tag{181}$$

where $\bar{\mathbf{c}}^{(\lambda)}$, $a^{(\lambda)}$ and $b^{(\lambda)}$ satisfy the following conditions:

$$\Sigma_{\frac{1}{\zeta}}\left[(\Sigma_{\frac{1}{\zeta}}+\lambda MI)\bar{\mathbf{c}}^{(\lambda)}+T\begin{pmatrix}a^{(\lambda)}\\ b^{(\lambda)}\end{pmatrix}\right]=\Sigma_{\frac{1}{\zeta}}\mathbf{y}\quad\text{and}\quad T^{\top}\left[\Sigma_{\frac{1}{\zeta}}\bar{\mathbf{c}}^{(\lambda)}+T\begin{pmatrix}a^{(\lambda)}\\ b^{(\lambda)}\end{pmatrix}\right]=T^{\top}\mathbf{y},\ (182)$$

Now, as $\lambda\to 0$, (181) and (182) become:

$$\hat{h}^{(0^+)}(x)=\sum_{j=1}^{M}\bar{c}_j^{(0^+)}K_{\frac{1}{\zeta}}(x_j,x)+a^{(0^+)}+b^{(0^+)}x,\tag{183}$$

where $\bar{\mathbf{c}}^{(0^+)}$, $a^{(0^+)}$, and $b^{(0^+)}$ satisfy the following conditions:

$$\Sigma_{\frac{1}{\zeta}}\left[\Sigma_{\frac{1}{\zeta}}\bar{\mathbf{c}}^{(0^+)}+T\begin{pmatrix}a^{(0^+)}\\ b^{(0^+)}\end{pmatrix}\right]=\Sigma_{\frac{1}{\zeta}}\mathbf{y}\quad\text{and}\quad T^{\top}\left[\Sigma_{\frac{1}{\zeta}}\bar{\mathbf{c}}^{(\lambda)}+T\begin{pmatrix}a^{(0^+)}\\ b^{(0^+)}\end{pmatrix}\right]=T^{\top}\mathbf{y}.\tag{184}$$

The expressions (183) and (184) give the solution of (177) as $\lambda\to 0$, which is also the solution to the variational problem (24).

Appendix P. Possible Generalizations

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 Lee et al. (2019); Lai et al. (2023) show that the model still is approximated by a linearized model. The result on kernel norm minimization (Zhang et al., 2020) 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.

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., Soudry et al. (2018). 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 Lee et al. (2018), 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 Lee et al. (2018) to other loss functions, then we can generalize our main result in Theorem 1 to other loss functions as well.

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 Gunasekar et al. (2018a) 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.

References

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. URL http://www.sciencedirect.com/science/article/pii/0378375895000216.

J. H. Ahlberg, Edwin N. Nilson, and J. L. Walsh. *The Theory of Splines and Their Applications*. ISSN. Elsevier Science, 1967. URL https://books.google.com/books?id=S7d1pjJHsRgC.

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*, pages 242–252, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/allen-zhu19a.html.

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*, pages 2269–2277. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/baratin21a.html.

Christopher Bishop. Regularization and complexity control in feed-forward networks. In *Proceedings International Conference on Artificial Neural Networks ICANN'95*, volume 1, pages 141–148. EC2 et Cie, January 1995. URL https://www.microsoft.com/en-us/research/publication/regularization-and-complexity-control-in-feed-forward-networks/.

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.

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*, pages 2205–2211. International Joint Conferences on Artificial Intelligence Organization, 8 2021. URL https://doi.org/10.24963/ijcai.2021/304. Main Track.

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*, pages 1305–1338. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/chizat20a.html.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

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. URL https://proceedings.neurips.cc/paper/2017/file/489d0396e6826eb0c1e611d82ca8b215-Paper.pdf.

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*, pages 1019–1028, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/dinh17b.html.

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. URL https://openreview.net/forum?id=S1eK3i09YQ.

Wilna Du Toit. *Radial basis function interpolation*. PhD thesis, Stellenbosch: Stellenbosch University, 2008.

PPB Eggermont and VN LaRiccia. Uniform error bounds for smoothing splines. *Lecture Notes-Monograph Series*, pages 220–237, 2006.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

G German. Smoothing and non-parametric regression. *International Journal of Systems Science*, 2001.

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, pages 1832–1841, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018a. PMLR. URL http://proceedings.mlr.press/v80/gunasekar18a.html.

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*, pages 9461–9471. Curran Associates, Inc., 2018b. URL http://papers.nips.cc/paper/8156-implicit-bias-of-gradient-descent-on-linear-convolutional-networks.pdf.

Charles A Hall and W Weston Meyer. Optimal error bounds for cubic spline interpolation. *Journal of Approximation Theory*, 16(2):105–122, 1976.

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.

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*, pages 8571–8580. Curran Associates, Inc., 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable 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*, pages 1772–1798, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/ji19a.html.

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. URL https://openreview.net/pdf?id=H1oyRlYgg.

Mateusz Kwaśnicki. Ten equivalent definitions of the fractional laplace operator. *Fractional Calculus and Applied Analysis*, 20(1):7–51, 2017.

Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization ability of wide neural networks on $\mathbb{R}$. *arXiv preprint arXiv:2302.05933*, 2023.

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. URL https://openreview.net/forum?id=B1EA-M-0Z.

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*, pages 8572–8583. Curran Associates, Inc., 2019.

Ziyue Liu and Wensheng Guo. Data driven adaptive spline smoothing. *Statistica Sinica*, pages 1143–1163, 2010.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018.

Richard B Melrose and Gunther Uhlmann. *An introduction to microlocal analysis*. Department of Mathematics, Massachusetts Institute of Technology, 2008.

C. Nasim. The solution of an integral equation. *Proceedings of the American Mathematical Society*, 40(1):95–101, 1973. URL `http://www.jstor.org/stable/2038642`.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

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. URL `http://arxiv.org/abs/1412.6614`.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

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. URL `https://openreview.net/forum?id=H1lNPxHKDH`.

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*, pages 4951–4960, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/oymak19a.html`.

Rahul Parhi and Robert D. Nowak. Minimum "norm" neural networks are splines. *arXiv preprint arXiv:1910.02333*, 2019.

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.

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*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Alexandre Pintore, Paul Speckman, and Chris C. Holmes. Spatially adaptive smoothing splines. *Biometrika*, 93(1):113–125, 03 2006. doi: 10.1093/biomet/93.1.113. URL https://doi.org/10.1093/biomet/93.1.113.

Evelyn Dianne Hatton Potter. *Multivariate polyharmonic spline interpolation*. Iowa State University, 1981.

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.

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*, pages 5301–5310, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/rahaman19a.html.

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, 2020a. URL https://openreview.net/forum?id=BJl9PRVKDS.

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, 2020b.

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*, pages 2667–2690, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/savarese19a.html.

Johannes Schmidt-Hieber. Rejoinder: "nonparametric regression using deep neural networks with ReLU activation function". *The Annals of Statistics*, 48(4):1916–1921, 2020.

Karel Segeth. Multivariate smooth interpolation that employs polyharmonic functions. *Programs and Algorithms of Numerical Mathematics*, pages 140–148, 2019.

Donald C Solmon. Asymptotic formulas for the dual radon transform and applications. *Mathematische Zeitschrift*, 195(3):321–343, 1987.

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.

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.

Xiao Wang, Pang Du, and Jinglai Shen. Smoothing splines with varying smoothing parameter. *Biometrika*, 100(4):955–970, 2013.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

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*, pages 8378–8387, 2019.

Lei Wu, Zhanxing Zhu, and E Weinan. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

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. URL `https://arxiv.org/abs/1611.03530`.

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*, pages 144–164, Princeton University, Princeton, NJ, USA, 20–24 Jul 2020. PMLR. URL `http://proceedings.mlr.press/v107/zhang20a.html`.