

SUPPLEMENTARY MATERIAL

IMPLICIT BIAS OF GRADIENT DESCENT FOR MEAN SQUARED ERROR REGRESSION WITH WIDE NEURAL NETWORKS

Anonymous authors

Paper under double-blind review

Appendices

The appendices are organized as follows. In Appendix A we illustrate our theoretical results numerically, and in Appendix B we provide details on the numerical implementation.

In Appendix C we briefly comment on 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 D E, F, G, H, I, J, K we provide the proofs of the formal results from the main part.

In Appendix L 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 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. We provide an interpretable description of the kernel norm.

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 upon gradient descent training in the infinite width limit.

In Appendix P we comment on some of the possible extensions and generalizations of the analysis. In particular, we give a generalization of Theorem 1 to the case of multi-dimensional inputs, and a formulation for neural networks with activation function different from ReLU.

A NUMERICAL ILLUSTRATION OF THE THEORETICAL RESULTS

Gradient descent training and variational problem To illustrate Theorem 1 across different initialization procedures, in Figures A1 and A2 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 A3 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 4, we conduct the following experiment. We use the same training set as in Figure 1 and use uniform initialization.

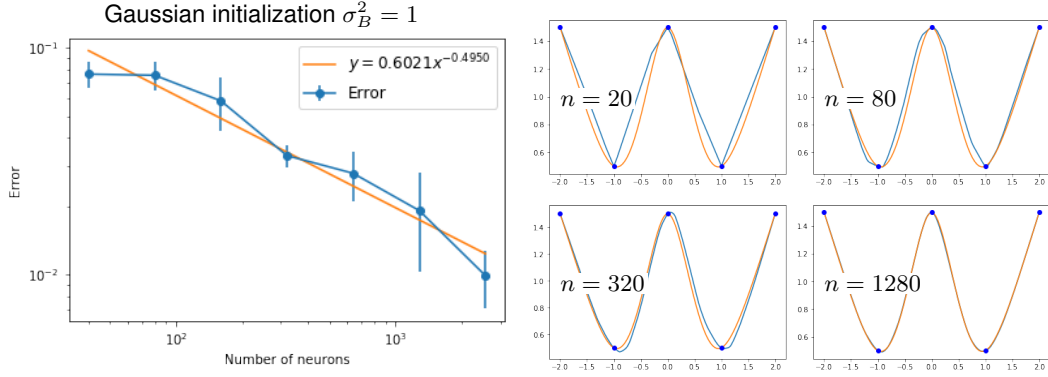


Figure A1: Illustration of Theorem 1. Shown is the error between the output function $f(\cdot, \theta^*)$ of the trained neural network and the solution g^* to the variational problem (19) against the number of neurons, n . Shown is the average over 5 repetitions, with error bars indicating the standard deviation. Here the training data is fixed, and the parameters were initialized with $W \sim N(0, 1)$ and $B \sim N(0, 1)$. 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.

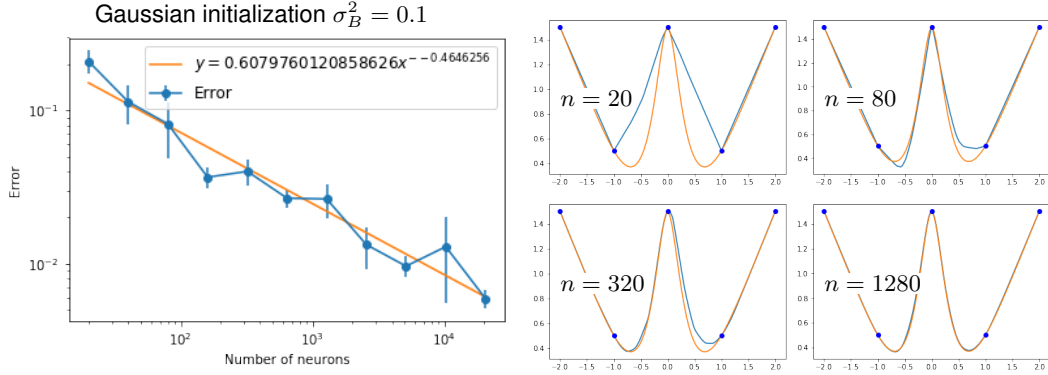


Figure A2: Illustration of Theorem 1. Similar to Figure A1, but with a different initialization $W \sim N(0, 1)$ and $N(0, 0.1)$, which gives rise to a curvature penalty function ζ that is more strongly peaked around $x = 0$ (see Figure 1). We observe in particular that the solutions are more curvy around $x = 0$.

Starting from the same initial weights, we train the network in two ways. One way is only training the output layer and another way is training all layers of the network. The result is shown in Figure A4. 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 4. 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 our main result Theorem 1, since the variational problem defines functions only up to addition of linear functions, we need to adjust training data by subtracting a specific linear function $ux + v$. However, in our previous experiments, we don't adjust the training data and the statement of Theorem 1 still approximately holds. The reason might be that the coefficients u and v of the linear function which we need to subtract are relatively small. In order to see the effect of linear function on 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 A5 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 A5. From the right panel

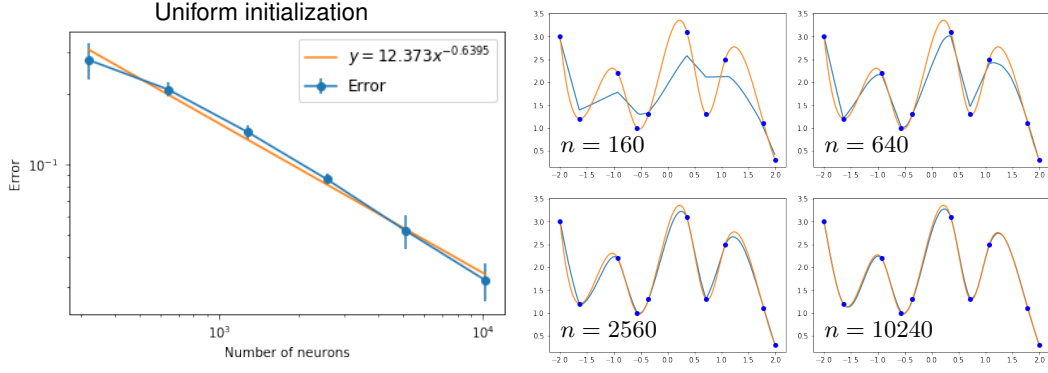


Figure A3: Illustration of Theorem 1. Similar to Figure 1, with uniform initialization, but with a larger dataset and larger networks.

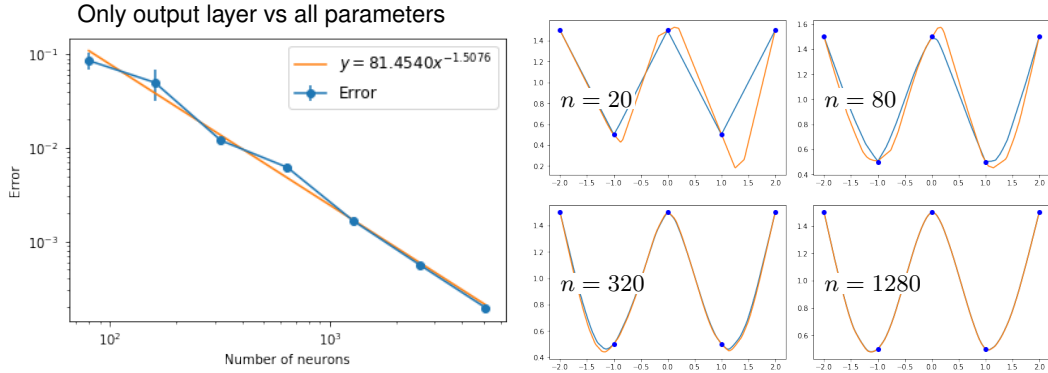


Figure A4: Illustration of Theorem 4. Training only output layer vs training all parameters of the network. We use uniform initialization and the same training set as in Figure 1. The left panel plots the error between two trained network functions against the number of neurons n . For one network, we 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.

of Figure A5, we see that the difference between plotted two functions is relatively larger than that in Figure 1. From the left panel of Figure A5, 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 $n \rightarrow \infty$ 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.

B DETAILS ON THE NUMERICAL IMPLEMENTATION

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 A1 with respect to the upper bound of the learning rate in order to converge.

We note that the implicit bias in parameter space Theorem A1 is independent of the specific step size that is used in the optimization, so long as it is small enough. See Appendix E. 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: $|L(\theta_t) - L(\theta_{t-1})| \leq 10^{-8}$.

For the comparison of the functions $f(\cdot, \theta^*)$ and g^* , the 2-norm of error $\|f(\cdot, \theta^*) - g^*\|_2$ is computed by numerical integration with step 0.01 over $[\min_i(x_i), \max_i(x_i)]$.

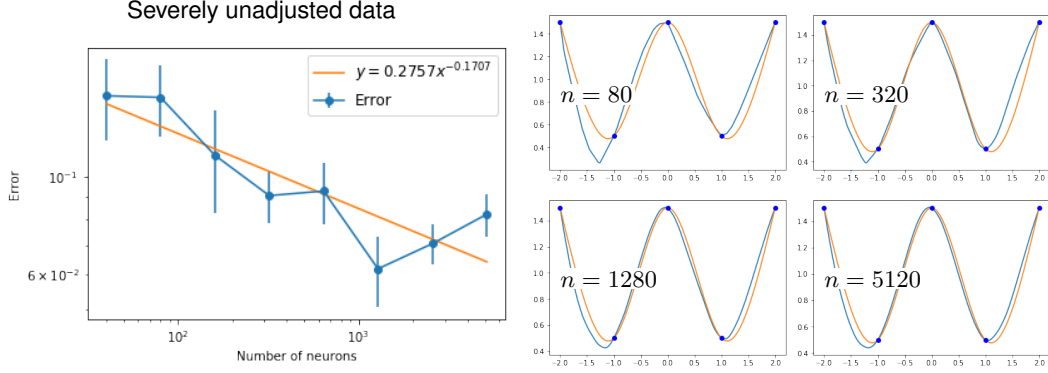


Figure A5: Effect of not adjusting the data. We use uniform initialization and add a linear function $10x + 10$ to the training data of Figure 1. In order to clearly show the difference between trained network function and the solution to the variational problem, we subtract $10x + 10$ from these two functions in the right panel. In the right panel we see that if we ignore u and v in the variational problem (17), the solution is slightly different from (19).

We use ASI (see Appendix C.2) at initialization. Then the initial output function of the network is $f(\cdot, \theta_0) \equiv 0$. Then according to Theorem 1, the weighted 2-norm of the second derivative of the trained network function is minimized. So in the figures the output function is actually equal to the difference from initialization.

Numerical solution of the variational problem The variational problem for cubic splines can be solved explicitly as described in Appendix O. For a general non-constant curvature penalty $1/\zeta$, we can obtain a numerical solution to problem (19) as follows. First we discretize the interval $[-L, L]$ evenly with points $x_j = -L + 2jL/n$, $j = 0, \dots, n$. For simplicity we suppose that the input training data points are among these grid points, and we denote them by x_{j_1}, \dots, 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 central differences 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 (19) 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 quadratic problem in $f(x_j)$, $j \in \{1, \dots, n\} \setminus \{j_1, \dots, j_m\}$. 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. We use a discretization of the interval $[-2, 2]$ into 200 pieces and use conjugate gradient descent for solving the linear system.

C ADDITIONAL COMMENTS

C.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 $\|f\|_p^2 = \mathbb{E}_{x \sim p}[f(x)^T f(x)]$, 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 x_1, \dots, x_M , positive definiteness of the corresponding Gram matrix is equivalent to $\partial_{\theta_i} f(x_j, \cdot)$ being linearly independent (Du et al., 2018, 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 $\sum_i W_i^{(2)} \sigma(W_i^{(1)} x + b_i)$ is a sum of independent random variables with finite variance.

C.2 ANTI-SYMMETRICAL INITIALIZATION (ASI)

The AntiSymmetrical Initialization (ASI) trick as proposed by Zhang et al. (2019) creates duplicate hidden units with opposite output weights, ensuring that $f(\cdot, \theta_0) \equiv 0$. More precisely, ASI defines $f_{\text{ASI}}(x, \vartheta) = \frac{\sqrt{2}}{2} f(x, \vartheta') - \frac{\sqrt{2}}{2} f(x, \vartheta'')$. Here $\vartheta = (\vartheta', \vartheta'')$ is initialized with $\vartheta'_0 = \vartheta''_0$, so that

$$f_{\text{ASI}}(x, \vartheta_0) = \sum_{i=1}^n \frac{\sqrt{2}}{2} \bar{V}_i^{(2)} [\bar{V}_i^{(1)} x + \bar{a}_i^{(1)}]_+ + \sum_{i=1}^n -\frac{\sqrt{2}}{2} \bar{V}_i^{(2)} [\bar{V}_i^{(1)} x + \bar{a}_i^{(1)}]_+ \equiv 0. \quad (\text{A1})$$

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

The basic statistics on the size of the parameters remains like (2), 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. (2019). A related discussion also appears in Sahs et al. (2020).

C.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 parameterization is described as

$$\begin{cases} h^{(l+1)} = \sqrt{\frac{1}{n_l}} w^{(l+1)} x^l + b^{(l+1)} \\ x^{(l+1)} = \phi(h^{(l+1)}) \end{cases} \quad \text{and} \quad \begin{cases} w_{ij}^{(l)} \sim \mathcal{N}(0, 1) \\ b_j^{(l)} \sim \mathcal{N}(0, 1) \end{cases}. \quad (\text{A2})$$

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 $\nabla_{w_i^{(1)}} f(x, \theta_0) = O(n^{-\frac{1}{2}})$ and $\nabla_{w_i^{(2)}} f(x, \theta_0) = O(n^{-\frac{1}{2}})$. In turn, during training the changes of $w_i^{(1)}$ and $w_i^{(2)}$ are comparable in magnitude. This implies that we can not ignore the changes of $w_i^{(1)}$ 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 parameterization. In particular, we can not use problem (A63) to describe the result of gradient descent as $n \rightarrow \infty$.

C.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 $\lim_{\epsilon \rightarrow 0} (\inf_{\theta} C(\theta) \text{ s.t. } \|f(\cdot, \theta) - g\|_{\infty} \leq \epsilon) = \max\{\int_{-\infty}^{\infty} |g''(x)| dx, |g'(-\infty) + g'(\infty)|\}$, 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 (2) with sub-Gaussian distribution, the neural network can be approximated by the linearized model. Then by Theorem A1, $\|\omega - \theta_0\|_2$ 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 C.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) can 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 Lf 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 P.

C.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 $\theta \mapsto f(\mathcal{X}, \theta) = [f(x_1, \theta), \dots, f(x_m, \theta)]$ 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 $f(x; \theta) = W^{(2)}[W^{(1)}x + b^{(1)}]_+ + b^{(2)}$, where $\theta = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$. The Jacobian is non-smooth whenever $\theta \in H_{xj} = \{W_{j1}^{(1)}x + b_j^{(1)} = 0\}$ for some $j = 1, \dots, n$. Hence for m input data points $x_i, i = 1, \dots, m$, the locus of non-smoothness is given by m central hyperplanes $H_{ij}, i = 1, \dots, m$ in the parameter space of each hidden unit $j = 1, \dots, n$. 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 $H_i, i = 1, \dots, m$, behaves roughly as $p = O(m\epsilon^{-1})$. Hence we can expect that the prediction function will be better approximated by its linearization $f^{\text{lin}}(x, \theta) = f(x, c\theta_0) + \nabla_{\theta} f(x, c\theta_0)(\theta - c\theta_0)$ at a point $c\theta_0$ if c is larger. This is well reflected numerically in Figure A6. 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.

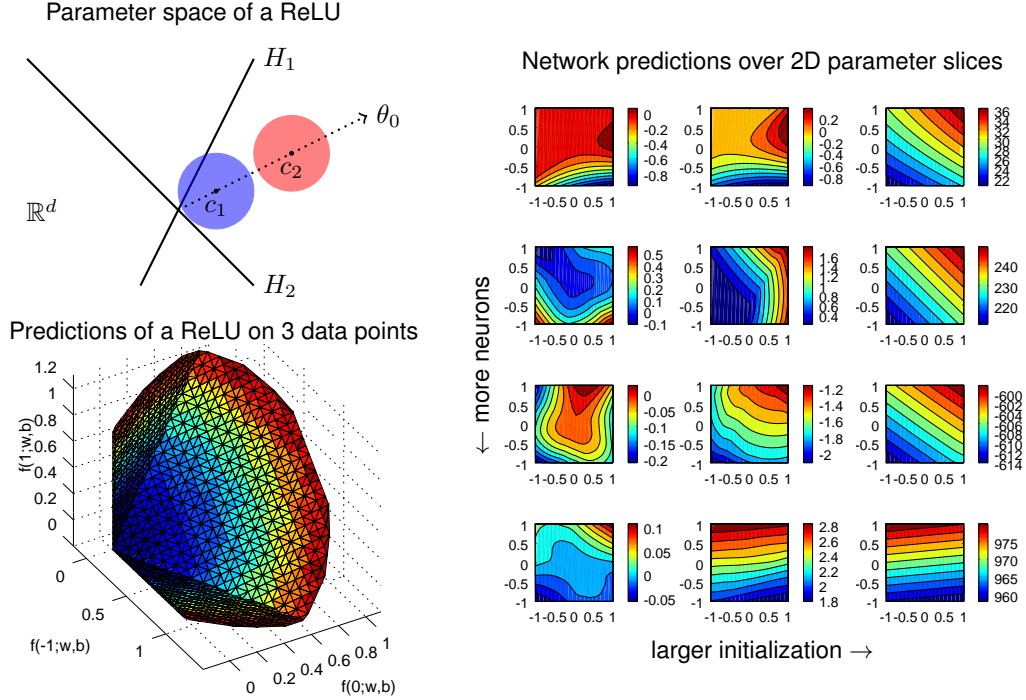


Figure A6: Left: For a single ReLU, the map $\theta \mapsto f(\mathcal{X}, \theta)$ from parameters to prediction vectors over a set $\mathcal{X} = \{x_1, \dots, x_m\}$ of m input data points is piecewise linear, with pieces separated by m central hyperplanes. Right: Shown is the prediction $f(x, \theta)$ 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.

D PROOF OF THEOREM 1

Proof of Theorem 1. The convergence to zero training error for ReLU networks is by now a well known result (Du et al., 2018; Allen-Zhu et al., 2019). We proceed with the implicit bias result.

For simplicity, we give out the proof under ASI (see Appendix C.2). In Section 5.2, we relax the optimization problem (16) to (17). Suppose $(\bar{\gamma}, \bar{u}, \bar{v})$ is the solution of (17). Then we can adjust the training samples $\{(x_i, y_i)\}_{i=1}^M$ to $\{(x_i, y_i - \bar{u}x_i - \bar{v})\}_{i=1}^M$. It's easy to see that on the adjusted training samples, $(0, 0, \bar{\gamma})$ is the solution of (17). Then $\bar{\gamma}$ is the solution of (16). Furthermore, the solution of (16) in function space, $g(x, \bar{\gamma})$, equals to the solution of (17) in function space, $g(x, (\bar{\gamma}, 0, 0))$, i.e.

$$g(x, \bar{\gamma}) = g(x, (\bar{\gamma}, 0, 0)). \quad (\text{A3})$$

If we change the variable γ to α as in Section 5.2, we get

$$g(x, \bar{\alpha}) = g(x, \bar{\gamma}), \quad (\text{A4})$$

where $g(x, \bar{\alpha})$ is the solution of the continuous version of problem (15) with μ in place of μ_n . On the set $S = \text{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$, according to Theorem 5,

$$\sup_{x \in S} |g_n(x, \bar{\alpha}_n) - g(x, \bar{\alpha})| = O(n^{-1/2}), \quad (\text{A5})$$

where $g_n(x, \bar{\alpha}_n)$ is the solution of problem (15) in function space. Since problem (15) is equivalent to problem (14), $g_n(x, \bar{\alpha}_n)$ is also the solution of (14) in function space. According to discussion in Section 5, $f^{\text{lin}}(x_j, \tilde{\omega}_\infty)$ is the solution of (14). Then

$$g_n(x, \bar{\alpha}_n) = f^{\text{lin}}(x_j, \tilde{\omega}_\infty). \quad (\text{A6})$$

According to Corollary A5,

$$\|f^{\text{lin}}(x, \tilde{\omega}_\infty) - f(x, \theta^*)\|_2 = O(n^{-\frac{1}{2}}). \quad (\text{A7})$$

Finally, according to Theorem 6 and Proposition 7, $g(x, (\bar{\gamma}, 0, 0))$ restricted on S is the solution of (4), which is $g^*(x)$. Then on the set S ,

$$g(x, (\bar{\gamma}, 0, 0)) = g^*(x) \quad (\text{A8})$$

Combining (A3), (A4), (A5), (A6), (A7), (A8), and using the fact that on domain S , $\|f\|_2 \leq \text{vol}(S)\|f\|_\infty$, we prove the theorem. \square

E IMPLICIT BIAS IN PARAMETER SPACE FOR A LINEARIZED MODEL

Zhang et al. (2019) 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 A1 (Bias of the linearized model in parameter space). *Consider a convex loss function ℓ with a unique finite minimum and which is K -Lipschitz continuous, i.e. $|\ell(\hat{y}_1, y) - \ell(\hat{y}_2, y)| \leq K|\hat{y}_1 - \hat{y}_2|$. If $\text{rank}(\nabla_\theta f(\mathcal{X}, \theta_0)) = M$, then the gradient descent iteration (9) with learning rate $\eta \leq \frac{1}{Kn\sqrt{M}\lambda_{\max}(\hat{\Theta}_n)}$ converges to the unique solution of following constrained optimization problem:*

$$\min_{\omega} \|\omega - \theta_0\|_2 \quad \text{s.t.} \quad f^{\text{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \quad (\text{A9})$$

Remark A2 (Remark on Theorem A1, 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 A3 (Remark on Theorem A1, rank assumption). The assumption $\nabla_\theta f(\mathcal{X}, \theta_0) = M$ is satisfied in most cases when $n \geq M$ (here n refers to the number of parameters in θ since we use the linearized model). This is because $\nabla_\theta f(\mathcal{X}, \theta_0)$ is a $M \times n$ matrix. The M rows corresponds to M training samples and they are almost always linearly independent.

Here we give out the proof of Theorem A1. We note that Zhang et al. (2019) prove a similar result for gradient flow. Our proof is for finite step size and different from theirs.

Proof of Theorem A1. We use gradient descent to minimize $L^{\text{lin}}(\omega) = \sum_{i=1}^M \ell(f^{\text{lin}}(x_i, \omega), y_i)$. First we prove that $\nabla_\omega L^{\text{lin}}(\omega)$ is Lipschitz continuous. Since

$$\begin{aligned} & \|\nabla_\omega L^{\text{lin}}(\omega_1) - \nabla_\omega L^{\text{lin}}(\omega_2)\|_2 \\ &= \|\nabla_\theta f(\mathcal{X}, \theta_0)^\top \nabla_{f^{\text{lin}}(\mathcal{X}, \omega_1)} L - \nabla_\theta f(\mathcal{X}, \theta_0)^\top \nabla_{f^{\text{lin}}(\mathcal{X}, \omega_2)} L\|_2 \\ &\leq \|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|\nabla_{f^{\text{lin}}(\mathcal{X}, \omega_1)} L - \nabla_{f^{\text{lin}}(\mathcal{X}, \omega_2)} L\|_2 \\ &\leq K \|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|f^{\text{lin}}(\mathcal{X}, \omega_1) - f^{\text{lin}}(\mathcal{X}, \omega_2)\|_1 \quad (\text{K-Lipschitz continuity of } \ell) \\ &\leq K\sqrt{M} \|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|f^{\text{lin}}(\mathcal{X}, \omega_1) - f^{\text{lin}}(\mathcal{X}, \omega_2)\|_2 \\ &= K\sqrt{M} \|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|\nabla_\theta f(\mathcal{X}, \theta_0)(\omega_1 - \omega_2)\|_2 \\ &\leq K\sqrt{M} \|\nabla_\theta f(\mathcal{X}, \theta_0)^\top\|_2 \|\nabla_\theta f(\mathcal{X}, \theta_0)\|_2 \|\omega_1 - \omega_2\|_2 \\ &\leq Kn\sqrt{M}\lambda_{\max}(\hat{\Theta}_n) \|\omega_1 - \omega_2\|_2. \end{aligned} \quad (\text{A10})$$

So $L^{\text{lin}}(\omega)$ is Lipschitz continuous with Lipschitz constant $Kn\sqrt{M}\lambda_{\max}(\hat{\Theta}_n)$. Since L^{lin} is convex over ω , gradient descent with learning rate $\eta = \frac{1}{Kn\sqrt{M}\lambda_{\max}(\hat{\Theta}_n)}$ converges to a global minimum 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 minimum of $L^{\text{lin}}(\omega)$ is zero and gradient descent converges to zero loss.

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

$$\begin{aligned}\omega_\infty &= \theta_0 - \sum_{t=0}^{\infty} \eta \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \nabla_{f^{\text{lin}}(\mathcal{X}, \omega_t)} L^{\text{lin}} \\ &= \theta_0 - \eta \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \sum_{t=0}^{\infty} \nabla_{f^{\text{lin}}(\mathcal{X}, \omega_t)} L^{\text{lin}}\end{aligned}\tag{A11}$$

Since f^{lin} is linear over weights ω and $\|\omega - \theta_0\|_2$ is strongly convex, the constrained optimization problem (A9) is a strongly convex optimization problem. The first order optimality condition of the problem is

$$\begin{cases} \omega - \theta_0 + \nabla_{\theta} f^{\text{lin}}(\mathcal{X}, \theta_0)^T \lambda = 0 \\ f^{\text{lin}}(\mathcal{X}, \omega) = \mathcal{Y}. \end{cases}\tag{A12}$$

Let $\lambda = \sum_{t=0}^{\infty} \nabla_{f^{\text{lin}}(\mathcal{X}, \theta_t)} L$, we can easily check out that ω_∞ satisfies condition (A12). So ω_∞ is the solution of problem (A9). \square

Remark A4 (Remark to Theorem A1). Making an analogous statement to Theorem A1 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.

F PROOF OF THEOREM 4

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

Proof of Theorem 4. According to (9),

$$\omega_{t+1} = \omega_t - \eta \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \nabla_{f^{\text{lin}}(\mathcal{X}, \omega_t)} L^{\text{lin}}.\tag{A13}$$

Since we use the MSE loss,

$$\omega_{t+1} = \omega_t - \eta \nabla_{\theta} f(\mathcal{X}, \theta_0)^T (f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y}).\tag{A14}$$

Using (8), we get

$$\begin{aligned}f^{\text{lin}}(\mathcal{X}, \omega_{t+1}) &= f^{\text{lin}}(\mathcal{X}, \omega_t) - \eta \nabla_{\theta} f(\mathcal{X}, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^T (f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y}) \\ &= f^{\text{lin}}(\mathcal{X}, \omega_t) - n\eta \hat{\Theta}_n (f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y}).\end{aligned}\tag{A15}$$

Then we have

$$f^{\text{lin}}(\mathcal{X}, \omega_{t+1}) - \mathcal{Y} = (I - n\eta \hat{\Theta}_n) (f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y}),\tag{A16}$$

and

$$\begin{aligned}f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y} &= (I - n\eta \hat{\Theta}_n)^t (f^{\text{lin}}(\mathcal{X}, \theta_0) - \mathcal{Y}) \\ &= (I - n\eta \hat{\Theta}_n)^t (f(\mathcal{X}, \theta_0) - \mathcal{Y}).\end{aligned}\tag{A17}$$

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

$$\begin{aligned}f^{\text{lin}}(\mathcal{X}, \omega_t) - \mathcal{Y} &= f^{\text{lin}}(\mathcal{X}, \omega_t) - f(\mathcal{X}, \theta_0) + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= \nabla_{\theta} f(\mathcal{X}, \theta_0) (\omega_t - \theta_0) + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= \nabla_{\theta} f(\mathcal{X}, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \xi + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= n\hat{\Theta}_n \xi + f(\mathcal{X}, \theta_0) - \mathcal{Y} \\ &= (I - n\eta \hat{\Theta}_n)^t (f(\mathcal{X}, \theta_0) - \mathcal{Y}).\end{aligned}\tag{A18}$$

From above equation we can solve for ξ :

$$\xi = -n^{-1}\hat{\Theta}_n^{-1}[I - (I - n\eta\hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}). \quad (\text{A19})$$

Therefore

$$\omega_t = -n^{-1}\nabla_{\theta}f(\mathcal{X}, \theta_0)^T\hat{\Theta}_n^{-1}[I - (I - n\eta\hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) + \theta_0. \quad (\text{A20})$$

For any $x \in \mathbb{R}$,

$$\begin{aligned} f^{\text{lin}}(x, \omega_t) &= f(x, \theta_0) + \nabla_{\theta}f(x, \theta_0)(\omega_t - \theta_0) \\ &= f(x, \theta_0) - n^{-1}\nabla_{\theta}f(x, \theta_0)\nabla_{\theta}f(\mathcal{X}, \theta_0)^T\hat{\Theta}_n^{-1}[I - (I - n\eta\hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}). \end{aligned} \quad (\text{A21})$$

For the training process (12), we can define the corresponding empirical neural tangent kernel in the following way:

$$\tilde{\Theta}_n = \frac{1}{n}\nabla_{W^{(2)}}f(\mathcal{X}, \theta_0)\nabla_{W^{(2)}}f(\mathcal{X}, \theta_0)^T. \quad (\text{A22})$$

Using the same argument, we have

$$\widetilde{W}_t^{(2)} = -n^{-1}\nabla_{W^{(2)}}f(\mathcal{X}, \theta_0)^T\tilde{\Theta}_n^{-1}[I - (I - n\eta\tilde{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) + \overline{W}_0^{(2)} \quad (\text{A23})$$

and

$$f^{\text{lin}}(x, \widetilde{\omega}_t) = f(x, \theta_0) - n^{-1}\nabla_{W^{(2)}}f(x, \theta_0)\nabla_{W^{(2)}}f(\mathcal{X}, \theta_0)^T\tilde{\Theta}_n^{-1}[I - (I - n\eta\tilde{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}). \quad (\text{A24})$$

Then

$$\begin{aligned} &|f^{\text{lin}}(x, \widetilde{\omega}_t) - f^{\text{lin}}(x, \omega_t)| \\ &= n^{-1} \left| \nabla_{\theta}f(x, \theta_0)\nabla_{\theta}f(\mathcal{X}, \theta_0)^T\hat{\Theta}_n^{-1}[I - (I - n\eta\hat{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) \right. \\ &\quad \left. - \nabla_{W^{(2)}}f(x, \theta_0)\nabla_{W^{(2)}}f(\mathcal{X}, \theta_0)^T\tilde{\Theta}_n^{-1}[I - (I - n\eta\tilde{\Theta}_n)^t](f(\mathcal{X}, \theta_0) - \mathcal{Y}) \right|. \end{aligned} \quad (\text{A25})$$

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

$$\begin{aligned} (\Delta\Theta)_{ij} &= \frac{1}{n} \left[\sum_{k=1}^n \left(\nabla_{W_k^{(1)}}f(x_i, \theta_0)\nabla_{W_k^{(1)}}f(x_j, \theta_0) + \nabla_{b_k^{(1)}}f(x_i, \theta_0)\nabla_{b_k^{(1)}}f(x_j, \theta_0) \right) \right. \\ &\quad \left. + \nabla_{b^{(2)}}f(x_i, \theta_0)\nabla_{b^{(2)}}f(x_j, \theta_0) \right]. \end{aligned} \quad (\text{A26})$$

According to initialization (2), we can find a $C > 0$ such that $|W_i^{(1)}|, |b_i^{(1)}| \leq C$ and $|W_i^{(2)}|, |b^{(2)}| \leq Cn^{-\frac{1}{2}}$ with probability at least $(1 - \delta/4)$. Then given $x \in \mathbb{R}$,

$$|\nabla_{W_i^{(1)}}f(x, \theta_0)| = |W_i^{(2)}H(W_i^{(1)}x + b) \cdot x| \leq Cn^{-\frac{1}{2}}x = O(n^{-\frac{1}{2}}), \quad (\text{A27})$$

$$|\nabla_{b_i^{(1)}}f(x, \theta_0)| = |W_i^{(2)}H(W_i^{(1)}x + b_i^{(1)})| \leq Cn^{-\frac{1}{2}} = O(n^{-\frac{1}{2}}) \quad (\text{A28})$$

$$|\nabla_{W_i^{(2)}}f(x, \theta_0)| = [W_i^{(1)}x + b_i^{(1)}]_+ \leq C|x| + C = O(1), \quad (\text{A29})$$

$$|\nabla_{b^{(2)}}f(x, \theta_0)| = 1 = O(1). \quad (\text{A30})$$

So

$$\begin{aligned} |(\Delta\Theta)_{ij}| &\leq \frac{1}{n} \left[\sum_{k=1}^n \left(O(n^{-\frac{1}{2}})O(n^{-\frac{1}{2}}) + O(n^{-\frac{1}{2}})O(n^{-\frac{1}{2}}) \right) + O(1)O(1) \right] \\ &= O(n^{-1}). \end{aligned} \quad (\text{A31})$$

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

Now we measure the difference of each part in (A25). According to assumption (2), $\inf_n \lambda_{\min}(\hat{\Theta}_n) > 0$, then

$$\lambda_{\min}(\hat{\Theta}_n) \geq \frac{1}{\inf_n \lambda_{\min}(\hat{\Theta}_n)} = O(1) \quad (\text{A32})$$

$$\lambda_{\min}(\tilde{\Theta}_n) \geq \frac{1}{\inf_n \lambda_{\min}(\hat{\Theta}_n) - O(n^{-1})} = O(1). \quad (\text{A33})$$

So

$$\begin{aligned} \|\hat{\Theta}_n^{-1} - \tilde{\Theta}_n^{-1}\|_2 &= \|\hat{\Theta}_n^{-1}(\tilde{\Theta}_n - \hat{\Theta}_n)\tilde{\Theta}_n^{-1}\|_2 \\ &\leq \|\hat{\Theta}_n^{-1}\|_2 \|\Delta\Theta\|_2 \|\tilde{\Theta}_n^{-1}\|_2 \\ &= O(n^{-1}). \end{aligned} \quad (\text{A34})$$

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

$$\|I - n\eta\hat{\Theta}_n\|_2 < 1, \quad (\text{A35})$$

And

$$\begin{aligned} \|I - n\eta\tilde{\Theta}_n\|_2 &\leq \|I - n\eta\hat{\Theta}_n\|_2 + n\eta\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2 \\ &\leq \max\left\{n\eta\frac{\lambda_{\max}(\Theta)}{2}, 1 - n\eta\lambda_{\min}(\hat{\Theta}_n)\right\} + O(n^{-1}). \end{aligned} \quad (\text{A36})$$

As n is large enough, we also have $\|I - n\eta\tilde{\Theta}_n\|_2 < 1$. Then as n is large enough,

$$\begin{aligned} &\|[I - (I - n\eta\hat{\Theta}_n)^t] - [I - (I - n\eta\tilde{\Theta}_n)^t]\|_2 \\ &= \|(I - n\eta\hat{\Theta}_n)^t - (I - n\eta\tilde{\Theta}_n)^t\|_2 \\ &\leq \|(I - n\eta\hat{\Theta}_n) - (I - n\eta\tilde{\Theta}_n)\|(I - n\eta\hat{\Theta}_n)^{t-1}\|_2 \\ &\quad + \|(I - n\eta\tilde{\Theta}_n)[(I - n\eta\hat{\Theta}_n) - (I - n\eta\tilde{\Theta}_n)](I - n\eta\hat{\Theta}_n)^{t-2}\|_2 \\ &\quad + \dots \\ &\quad + \|(I - n\eta\tilde{\Theta}_n)^{t-1}[(I - n\eta\hat{\Theta}_n) - (I - n\eta\tilde{\Theta}_n)]\|_2 \\ &\leq \eta\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2\|I - n\eta\hat{\Theta}_n\|_2^{t-1} \\ &\quad + \eta\|I - n\eta\tilde{\Theta}_n\|_2\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2\|I - n\eta\hat{\Theta}_n\|_2^{t-2} \\ &\quad + \dots \\ &\quad + \eta\|I - n\eta\tilde{\Theta}_n\|_2^{t-1}\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2 \\ &\leq \eta\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2 \cdot t \cdot (\max\{\|I - n\eta\hat{\Theta}_n\|_2, \|I - n\eta\tilde{\Theta}_n\|_2\})^{t-1}. \end{aligned} \quad (\text{A37})$$

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. So

$$\begin{aligned} \|[I - (I - n\eta\hat{\Theta}_n)^t] - [I - (I - n\eta\tilde{\Theta}_n)^t]\|_2 &\leq O(\eta\|\hat{\Theta}_n - \tilde{\Theta}_n\|_2) \\ &\leq O(n^{-1}). \end{aligned} \quad (\text{A38})$$

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

$$\begin{aligned} (\Delta\Theta(x, \mathcal{X}))_i &= \frac{1}{n} \left[\sum_{k=1}^n \left(\nabla_{W_k^{(1)}}f(x, \theta_0)\nabla_{W_k^{(1)}}f(x_i, \theta_0) + \nabla_{b_k^{(1)}}f(x, \theta_0)\nabla_{b_k^{(1)}}f(x_i, \theta_0) \right) \right. \\ &\quad \left. + \nabla_{b^{(2)}}f(x, \theta_0)\nabla_{b^{(2)}}f(x_i, \theta_0) \right]. \end{aligned} \quad (\text{A39})$$

According to (A27), (A28), (A29) and (A30), we have

$$\begin{aligned} |(\Delta\Theta(x, \mathcal{X}))_i| &\leq \frac{1}{n} \left[\sum_{k=1}^n \left(O(n^{-\frac{1}{2}})O(n^{-\frac{1}{2}}) + O(n^{-\frac{1}{2}})O(n^{-\frac{1}{2}}) \right) + O(1)O(1) \right] \\ &= O(n^{-1}). \end{aligned} \quad (\text{A40})$$

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

$$\|\Delta\Theta(x, \mathcal{X})\|_2 = O(n^{-1}). \quad (\text{A41})$$

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

$$\begin{aligned} |(\tilde{\Theta}_n(x, \mathcal{X}))_i| &\leq \frac{1}{n} \sum_{k=1}^n |\nabla_{W_k^{(2)}} f(x, \theta_0) \nabla_{W_k^{(2)}} f(x_i, \theta_0)| \\ &\leq \frac{1}{n} \sum_{k=1}^n |O(1)O(1)| \\ &= O(1). \end{aligned} \quad (\text{A42})$$

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

$$\|\Theta(x, \mathcal{X})\|_2 = O(n^{-1}). \quad (\text{A43})$$

Neal (1996), Lee et al. (2018) show that as n goes to infinity, the output function at initialization $f(\cdot, \theta_0)$ tends to a Gaussian process, 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. So as n is large enough, we can find a $C_1 > 0$ such that $f(x_i, \theta_0) \leq C_1, i = 1, \dots, M$ with probability at least $(1 - \delta/4)$. Then

$$\|f(\mathcal{X}, \theta_0) - \mathcal{Y}\|_2 = O(1). \quad (\text{A44})$$

Combine all these together, we have that with probability at least $(1 - \delta)$, (A32), (A33), (A34), (A35), (A38), (A41), (A42) and (A44) hold true. Then follow the equation (A25), we get

$$\begin{aligned} &|f^{\text{lin}}(x, \tilde{\omega}_t) - f(x, \theta_t)| \\ &= n^{-1} |\nabla_{\theta} f(x, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t] (f(\mathcal{X}, \theta_0) - \mathcal{Y}) \\ &\quad - \nabla_{W^{(2)}} f(x, \theta_0) \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t] (f(\mathcal{X}, \theta_0) - \mathcal{Y})| \\ &= n^{-1} \|\nabla_{\theta} f(x, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t] \\ &\quad - \nabla_{W^{(2)}} f(x, \theta_0) \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t]\|_2 \|f(\mathcal{X}, \theta_0) - \mathcal{Y}\|_2 \\ &= n^{-1} \|\nabla_{\theta} f(x, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t] \\ &\quad - \nabla_{W^{(2)}} f(x, \theta_0) \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t]\|_2 \cdot O(1). \end{aligned} \quad (\text{A45})$$

And

$$\begin{aligned} &n^{-1} \|\nabla_{\theta} f(x, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t] \\ &\quad - \nabla_{W^{(2)}} f(x, \theta_0) \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t]\|_2 \\ &\leq n^{-1} \|\nabla_{\theta} f(x, \theta_0) \nabla_{\theta} f(\mathcal{X}, \theta_0)^T - \nabla_{W^{(2)}} f(x, \theta_0) \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 \|\hat{\Theta}_n^{-1}\|_2 \|I - (I - n\eta \hat{\Theta}_n)^t\|_2 \\ &\quad + n^{-1} \|\nabla_{W^{(2)}} f(x, \theta_0) \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 \|\hat{\Theta}_n^{-1} - \tilde{\Theta}_n^{-1}\|_2 \|I - (I - n\eta \hat{\Theta}_n)^t\|_2 \\ &\quad + n^{-1} \|\nabla_{W^{(2)}} f(x, \theta_0) \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 \|\tilde{\Theta}_n^{-1}\|_2 \|I - (I - n\eta \hat{\Theta}_n)^t\|_2 - \|I - (I - n\eta \tilde{\Theta}_n)^t\|_2 \\ &\leq O(n^{-1})O(1)O(1) + O(1)O(n^{-1})O(1) + O(1)O(1)O(n^{-1}) \\ &= O(n^{-1}). \end{aligned} \quad (\text{A46})$$

So we have $|f^{\text{lin}}(x, \tilde{\omega}_t) - f(x, \theta_t)| = O(n^{-1})$, and $O(n^{-1})$ does not contain any constant factor which is related to t . Then

$$\sup_t |f^{\text{lin}}(x, \tilde{\omega}_t) - f^{\text{lin}}(x, \omega_t)| = O(n^{-1}), \text{ as } n \rightarrow \infty. \quad (\text{A47})$$

For the difference of parameters, we have

$$\tilde{\omega}_t - \omega_t = \text{vec}(\bar{W}_t^{(1)} - \widehat{W}_t^{(1)}, \bar{b}_t^{(1)} - \hat{b}_t^{(1)}, \bar{W}_t^{(2)} - \widehat{W}_t^{(2)}, \bar{b}_t^{(2)} - \hat{b}_t^{(2)}). \quad (\text{A48})$$

According to (A20) and (A23),

$$\begin{aligned} \|\bar{W}_t^{(1)} - \widehat{W}_t^{(1)}\|_2 &= \|n^{-1} \nabla_{W^{(1)}} f(\mathcal{X}, \theta_0)^T \hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t] (f(\mathcal{X}, \theta_0) - \mathcal{Y})\|_2 \\ &\leq \|n^{-1} \nabla_{W^{(1)}} f(\mathcal{X}, \theta_0)^T\|_2 \|\hat{\Theta}_n^{-1}\|_2 \|I - (I - n\eta \hat{\Theta}_n)^t\|_2 \|f(\mathcal{X}, \theta_0) - \mathcal{Y}\|_2 \\ &\leq n^{-1} \|\nabla_{W^{(1)}} f(\mathcal{X}, \theta_0)^T\|_2 \cdot O(1). \end{aligned} \quad (\text{A49})$$

Here $\nabla_{W^{(1)}} f(\mathcal{X}, \theta_0)^T$ is a $n \times M$ matrix, the ij -th entry of the matrix is $\nabla_{W_i^{(1)}} f(x_j, \theta_0)$. According to (A27), we have $\nabla_{W_i^{(1)}} f(x_j, \theta_0) = O(n^{-1/2})$. Then $\|\nabla_{W^{(1)}} f(\mathcal{X}, \theta_0)^T\|_2 = O(1)$. So we have $\|\bar{W}_t^{(1)} - \widehat{W}_t^{(1)}\|_2 = O(n^{-1})$, and $O(n^{-1})$ does not contain any constant factor which is related to t . Then

$$\sup_t \|\bar{W}_t^{(1)} - \widehat{W}_t^{(1)}\|_2 = O(n^{-1}), \text{ as } n \rightarrow \infty. \quad (\text{A50})$$

Similarly we can prove

$$\sup_t \|\bar{b}_t^{(1)} - \widehat{b}_t^{(1)}\|_2 = O(n^{-1}), \text{ as } n \rightarrow \infty. \quad (\text{A51})$$

$$\sup_t \|\bar{b}_t^{(2)} - \widehat{b}_t^{(2)}\|_2 = O(n^{-1}), \text{ as } n \rightarrow \infty. \quad (\text{A52})$$

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

$$\begin{aligned} \|\bar{W}_t^{(2)} - \widehat{W}_t^{(2)}\|_2 &= \|n^{-1} \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T \left(\hat{\Theta}_n^{-1} [I - (I - n\eta \hat{\Theta}_n)^t] - \right. \\ &\quad \left. \tilde{\Theta}_n^{-1} [I - (I - n\eta \tilde{\Theta}_n)^t] \right) (f(\mathcal{X}, \theta_0) - \mathcal{Y})\|_2 \\ &\leq \|n^{-1} \nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 \left(\|\hat{\Theta}_n^{-1} - \tilde{\Theta}_n^{-1}\|_2 \|I - (I - n\eta \hat{\Theta}_n)^t\|_2 + \right. \\ &\quad \left. \|\tilde{\Theta}_n^{-1}\|_2 \|I - (I - n\eta \hat{\Theta}_n)^t\|_2 - \|I - (I - n\eta \tilde{\Theta}_n)^t\|_2 \right) \|f(\mathcal{X}, \theta_0) - \mathcal{Y}\|_2 \\ &\leq n^{-1} \|\nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 (O(n^{-1})O(1) + O(1)O(n^{-1})) \cdot O(1) \\ &= O(n^{-2}) \|\nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2. \end{aligned} \quad (\text{A53})$$

Here $\nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T$ is a $n \times M$ matrix, the ij -th entry of the matrix is $\nabla_{W_i^{(2)}} f(x_j, \theta_0)$. According to (A29), we have $\nabla_{W_i^{(2)}} f(x_j, \theta_0) = O(1)$. Then $\|\nabla_{W^{(2)}} f(\mathcal{X}, \theta_0)^T\|_2 = O(n^{1/2})$. So we have $\|\bar{W}_t^{(2)} - \widehat{W}_t^{(2)}\|_2 = O(n^{-3/2})$, and $O(n^{-3/2})$ does not contain any constant factor which is related to t . Then

$$\sup_t \|\bar{W}_t^{(2)} - \widehat{W}_t^{(2)}\|_2 = O(n^{-3/2}), \text{ as } n \rightarrow \infty. \quad (\text{A54})$$

□

G TRAINING ONLY THE OUTPUT LAYER APPROXIMATES TRAINING A WIDE NETWORK

By combining Theorem 4 and the fact that training a linearized model approximates training a wide network (Lee et al., 2019, Theorem H.1), we obtain the following.

Corollary A5 (Training only output weights vs training all weights). *Consider the settings of Theorem 4, and assume that the joint distribution of $(\mathcal{W}, \mathcal{B})$ is sub-Gaussian. Then $\sup_t \|f^{\text{lin}}(x, \tilde{\omega}_t) - f(x, \theta_t)\|_2 = O(n^{-\frac{1}{2}})$ with arbitrarily high probability over the random initialization (2).*

Corollary A5 is obtained by combining Theorem 4 and the fact that training a linearized model approximates training a wide network (Lee et al., 2019, Theorem H.1). Although Lee et al. (2019, Theorem H.1) consider Gaussian initialization, the arguments extend to sub-Gaussian initialization.

Proof of Corollary A5. Using Theorem 4, we have that

$$\sup_t |f^{\text{lin}}(x, \tilde{\omega}_t) - f^{\text{lin}}(x, \omega_t)| = O(n^{-1}), \text{ as } n \rightarrow \infty. \quad (\text{A55})$$

According to Lee et al. (2019, Theorem H.1), in the case of Gaussian initialization, we have

$$\sup_t |f^{\text{lin}}(x, \omega_t) - f(x, \theta)| = O(n^{-\frac{1}{2}}), \text{ as } n \rightarrow \infty. \quad (\text{A56})$$

Under our neural network setting, which is a one-input network with a single hidden layer of n ReLUs and a linear output, we can generalize the above result to sub-Gaussian initialization. In the remark

of Theorem 4, we illustrate that the empirical NTK converges to analytic NTK for initialization with finite variance distribution. Then for sub-Gaussian initialization the empirical NTK still converges to analytic NTK. Then the only part we need to adapt in the proof of Lee et al. (2019, Theorem H.1) is the following theorem (Lee et al., 2019, Theorem G.3):

Theorem A6. *Let A be an $N \times n$ random matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2 \exp(-t^2/2)$ one has*

$$\|A\|_{\text{op}} \leq \sqrt{N} + \sqrt{n} + t. \quad (\text{A57})$$

Then Lee et al. (2019) applies the above theorem to weight matrices in the neural network. In our case, the weight matrices $W^{(1)}$ and $W^{(2)}$ are $1 \times n$ matrices, which can be regarded as vectors. So

$$\|W^{(1)}\|_{\text{op}} = \sqrt{\sum_{i=1}^n (W_i^{(1)})^2}. \quad (\text{A58})$$

Now we use sub-Gaussian initialization. Then $\mathbb{P}(|W_i^{(1)}| \geq t) \leq 2 \exp(-t^2/2\sigma^2)$ for some positive σ . Then $(W_i^{(1)})^2$ is sub-exponential. Using the property of sub-Gaussian exponential, we have $\mathbb{E} \exp(|W_i^{(1)}|^2/\lambda) \leq 2$ for some positive λ . Using Vershynin (2018, Theorem 1.4.1), we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n (W_i^{(1)})^2 - \mathbb{E} \sum_{i=1}^n (W_i^{(1)})^2\right| \geq t\right) \leq 2 \exp\left[-c \min\left(\frac{t^2}{n\lambda^2}, \frac{t}{\lambda}\right)\right]. \quad (\text{A59})$$

Let $t = n\lambda$, then we have

$$\mathbb{P}\left(\sum_{i=1}^n (W_i^{(1)})^2 \geq \mathbb{E} \sum_{i=1}^n (W_i^{(1)})^2 + n\lambda\right) \leq 2 \exp(-cn). \quad (\text{A60})$$

Since $2 \exp(-cn) \rightarrow 0$ as $n \rightarrow \infty$, the above equation means that with arbitrarily high probability,

$$\begin{aligned} \sum_{i=1}^n (W_i^{(1)})^2 &\leq \mathbb{E} \sum_{i=1}^n (W_i^{(1)})^2 + n\lambda \\ &= n\mathbb{E}(W_i^{(1)})^2 + n\lambda \\ &= O(n). \end{aligned} \quad (\text{A61})$$

So $\|W^{(1)}\|_{\text{op}} = O(\sqrt{n})$. For the same reason, $\|W^{(2)}\|_{\text{op}} = O(1)$. Then follow the remaining argument of Lee et al. (2019) we can show that

$$\sup_t |f^{\text{lin}}(x, \omega_t) - f(x, \theta)| = O(n^{-\frac{1}{2}}), \text{ as } n \rightarrow \infty. \quad (\text{A62})$$

Combine the above equation with (A55) then we finish the proof. \square

H PROOF OF THEOREM 5

We consider the continuous version of problem (15):

$$\begin{aligned} &\min_{\alpha \in C(\mathbb{R}^2)} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) \, d\mu(W^{(1)}, b) \\ &\text{subject to } \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)}x_j + b]_+ \, d\mu(W^{(1)}, b) = y_j, \quad j = 1, \dots, M. \end{aligned} \quad (\text{A63})$$

Here the only difference between (15) and (A63) is the difference of measures μ_n and μ .

Proof of Theorem 5. The Lagrangian of problem (15) is

$$L(\alpha_n, \lambda^n) = \int_{\mathbb{R}^2} \alpha_n^2(W^{(1)}, b) \, d\mu_n(W^{(1)}, b) + \sum_{j=1}^M \lambda_j^n (g_n(x_j, \alpha_n) - y_j) \quad (\text{A64})$$

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

$$\nabla_{\alpha_n} L = 2\alpha_n(W^{(1)}, b) + \sum_{j=1}^M \lambda_j^n [W^{(1)}x_j + b]_+ = 0 \text{ when } (W^{(1)}, b) = (W_i^{(1)}, b_i), i = 1, \dots, k. \quad (\text{A65})$$

Then

$$\bar{\alpha}_n(W^{(1)}, b) = -\frac{1}{2} \sum_{j=1}^M \lambda_j^n [W^{(1)}x_j + b]_+ \text{ when } (W^{(1)}, b) = (W_i^{(1)}, b_i), i = 1, \dots, k. \quad (\text{A66})$$

Since only function values on $(W_i^{(1)}, b_i)_{i=1}^M$ are really taken into account in problem (15), we can let

$$\bar{\alpha}_n(W^{(1)}, b) = -\frac{1}{2} \sum_{j=1}^M \lambda_j^n [W^{(1)}x_j + b]_+ \quad \forall (W^{(1)}, b) \in \mathbb{R}^2 \quad (\text{A67})$$

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

Here $\lambda_j^n, j = 1, \dots, M$ are chosen to make $g_n(x_i, \bar{\alpha}_n) = y_i, i = 1, \dots, M$. It means that

$$-\frac{1}{2} \sum_{j=1}^M \lambda_j^n \int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x_i + b]_+ d\mu_n(W^{(1)}, b) = y_i, i = 1, \dots, M. \quad (\text{A68})$$

Similarly the Lagrangian of problem (A63) is

$$\tilde{L}(\alpha, \lambda) = \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) d\mu(W^{(1)}, b) + \sum_{j=1}^M \lambda_j (g(x_j, \alpha) - y_j). \quad (\text{A69})$$

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

$$\nabla_{\alpha} \tilde{L} = 2\alpha(W^{(1)}, b) + \sum_{j=1}^M \lambda_j [W^{(1)}x_j + b]_+ = 0 \quad \forall (W^{(1)}, b) \in \mathbb{R}^2. \quad (\text{A70})$$

Then

$$\bar{\alpha}(W^{(1)}, b) = -\frac{1}{2} \sum_{j=1}^M \lambda_j [W^{(1)}x_j + b]_+ \quad \forall (W^{(1)}, b) \in \mathbb{R}^2. \quad (\text{A71})$$

Here $\lambda_j, j = 1, \dots, M$ are chosen to make $g(x, \alpha) = y_i, i = 1, \dots, M$. It means that

$$-\frac{1}{2} \sum_{j=1}^M \lambda_j \int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x_i + b]_+ d\mu(W^{(1)}, b) = y_i, i = 1, \dots, M. \quad (\text{A72})$$

Compare (A68) and (A72). Since the number of samples is finite, x_i is also bounded. Then by the assumption that \mathcal{W} and \mathcal{B} have finite fourth moments, we have that $[W^{(1)}x_j + b]_+ [W^{(1)}x_i + b]_+$ has finite variance. According to central limit theorem, as $n \rightarrow \infty$, $\int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x_i + b]_+ d\mu_n(W^{(1)}, b)$ tends to Gaussian distribution of variance $O(n^{-1})$. Then

$$\left| \int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x_i + b]_+ d\mu_n(W^{(1)}, b) - \int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x_i + b]_+ d\mu(W^{(1)}, b) \right| = O(n^{-1/2}) \quad (\text{A73})$$

$\forall i = 1, \dots, M, \forall j = 1, \dots, M$ with high probability. Since (A68) and (A72) are systems of linear equations and coefficients of (A68) converge to coefficients of (A72) at the rate of $O(n^{-1/2})$, then

$$|\lambda_j^n - \lambda_j| = O(n^{-1/2}), \quad j = 1, \dots, M. \quad (\text{A74})$$

Compare (A67) and (A71). Given $(W^{(1)}, b)$, we have

$$|\bar{\alpha}_n(W^{(1)}, b) - \bar{\alpha}(W^{(1)}, b)| = O(n^{-1/2}) \quad (\text{A75})$$

Next we want to prove that $\sup_{x \in [-L, L]} |g_n(x, \bar{\alpha}_n) - g(x, \bar{\alpha})| = O(n^{-1/2})$. Firstly, we prove that $\sup_{x \in [-L, L]} |g_n(x, \bar{\alpha}) - g(x, \bar{\alpha})| = O(n^{-1/2})$. Since

$$\begin{aligned} g_n(x, \bar{\alpha}) &= \int_{\mathbb{R}^2} \bar{\alpha}(W^{(1)}, b) [W^{(1)}x + b]_+ d\mu_n(W^{(1)}, b) \\ g(x, \bar{\alpha}) &= \int_{\mathbb{R}^2} \bar{\alpha}(W^{(1)}, b) [W^{(1)}x + b]_+ d\mu(W^{(1)}, b) \end{aligned} \quad (\text{A76})$$

So

$$\begin{aligned} \mathbb{E}g_n(x, \bar{\alpha}) &= g(x, \bar{\alpha}) \\ \text{Var } g_n(x, \bar{\alpha}) &= \frac{1}{n} \int_{\mathbb{R}^2} [\bar{\alpha}(W^{(1)}, b) [W^{(1)}x + b]_+ - g(x, \bar{\alpha})]^2 d\mu(W^{(1)}, b). \end{aligned} \quad (\text{A77})$$

Here the expectation and the variance are with respect to $(W_i^{(1)}, b_i)_{i=1}^n$. According to (A71) and the assumption that \mathcal{W} and \mathcal{B} have finite fourth moments, the integral in (A77) is bounded on $[-L, L]$. So $\sup_{x \in [-L, L]} \text{Var } g_n(x, \bar{\alpha}) = O(n^{-1})$. According to central limit theorem, as $n \rightarrow \infty$, $g_n(x, \bar{\alpha})$ tends to Gaussian distribution of variance $O(n^{-1})$ for any $x \in [-L, L]$. Then $|g_n(x, \bar{\alpha}) - g(x, \bar{\alpha})| = O(n^{-1/2})$ pointwise on $[-L, L]$ with high probability. Then we only need to prove that the sequence of functions $\{g_n(x, \bar{\alpha})\}_{n=1}^\infty$ is uniformly equicontinuous. Actually, $\forall x_1, x_2 \in [-L, L]$

$$\begin{aligned} &|g_n(x_1, \bar{\alpha}) - g_n(x_2, \bar{\alpha})| \\ &\leq \int_{\mathbb{R}^2} |\bar{\alpha}(W^{(1)}, b) [W^{(1)}x_1 + b]_+ - \bar{\alpha}(W^{(1)}, b) [W^{(1)}x_2 + b]_+| d\mu_n(W^{(1)}, b) \\ &\leq \int_{\mathbb{R}^2} |\bar{\alpha}(W^{(1)}, b)| |W_i^{(1)}| |x_1 - x_2| d\mu_n(W^{(1)}, b) \\ &\leq \int_{\mathbb{R}^2} |\bar{\alpha}(W^{(1)}, b)| |W_i^{(1)}| d\mu_n(W^{(1)}, b) |x_1 - x_2|. \end{aligned} \quad (\text{A78})$$

Because $\int_{\mathbb{R}^2} |\bar{\alpha}(W^{(1)}, b)| |W_i^{(1)}| d\mu_n(W^{(1)}, b) \rightarrow \int_{\mathbb{R}^2} |\bar{\alpha}(W^{(1)}, b)| |W_i^{(1)}| d\mu(W^{(1)}, b)$ with probability 1 according to the law of large numbers. So $\int_{\mathbb{R}^2} |\bar{\alpha}(W^{(1)}, b)| |W_i^{(1)}| d\mu_n(W^{(1)}, b)$ is bounded and the bound is independent of n . So $\{g_n(x, \bar{\alpha})\}_{n=1}^\infty$ is uniformly equicontinuous. Then by the argument similar to Arzela-Ascoli theorem, with high probability,

$$\sup_{x \in [-L, L]} |g_n(x, \bar{\alpha}) - g(x, \bar{\alpha})| = O(n^{-1/2}). \quad (\text{A79})$$

Finally, we prove that $\sup_{x \in [-L, L]} |g_n(x, \bar{\alpha}_n) - g_n(x, \bar{\alpha})| = O(n^{-1/2})$. Since $\forall x \in [-L, L]$

$$\begin{aligned} &|g_n(x, \bar{\alpha}_n) - g_n(x, \bar{\alpha})| \\ &\leq \int_{\mathbb{R}^2} |\bar{\alpha}_n(W^{(1)}, b) [W^{(1)}x + b]_+ - \bar{\alpha}(W^{(1)}, b) [W^{(1)}x + b]_+| d\mu_n(W^{(1)}, b) \\ &\leq \int_{\mathbb{R}^2} |\bar{\alpha}_n(W^{(1)}, b) - \bar{\alpha}(W^{(1)}, b)| [W^{(1)}x + b]_+ d\mu_n(W^{(1)}, b) \\ &\leq \int_{\mathbb{R}^2} \left| -\frac{1}{2} \sum_{j=1}^M (\lambda_j^n - \lambda_j) [W^{(1)}x_j + b]_+ \right| [W^{(1)}x + b]_+ d\mu_n(W^{(1)}, b) \\ &\leq \frac{1}{2} \sum_{j=1}^M |\lambda_j^n - \lambda_j| \int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x + b]_+ d\mu_n(W^{(1)}, b) \\ &\leq \frac{1}{2} \left(\max_{x \in [-L, L]} \int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x + b]_+ d\mu_n(W^{(1)}, b) \right) \sum_{j=1}^M |\lambda_j^n - \lambda_j|. \end{aligned} \quad (\text{A80})$$

Because $[-L, L]$ is compact and $\int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x + b]_+ d\mu_n(W^{(1)}, b)$ converges according to the law of large numbers, $\max_{x \in [-L, L]} \int_{\mathbb{R}^2} [W^{(1)}x_j + b]_+ [W^{(1)}x + b]_+ d\mu_n(W^{(1)}, b)$ is a finite number independent of n . Then according to (A74), $\max_{(W^{(1)}, b) \in \text{supp}(\mu)} |\bar{\alpha}_n(W^{(1)}, b) - \bar{\alpha}(W^{(1)}, b)| \rightarrow$

0. Then

$$\sup_{x \in [-L, L]} |g_n(x, \bar{\alpha}_n) - g_n(x, \bar{\alpha})| = O(n^{-1/2}). \quad (\text{A81})$$

Combined with (A79), we have

$$\sup_{x \in [-L, L]} |g_n(x, \bar{\alpha}_n) - g(x, \bar{\alpha})| = O(n^{-1/2}). \quad (\text{A82})$$

This concludes the proof. \square

I PROOF OF THEOREM 6

The second derivative g'' is given by

$$g''(x, \gamma) = p_{\mathcal{C}}(x) \int_{\mathbb{R}} \gamma(W^{(1)}, x) |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}). \quad (\text{A83})$$

The detailed calculation of (A83) is as follows:

$$\begin{aligned} g''(x, \gamma) &= \int_{\mathbb{R}^2} \gamma(W^{(1)}, c) |W^{(1)}| \delta(x - c) d\nu(W^{(1)}, c) \\ &= \int_{\text{supp}(\nu_{\mathcal{C}})} \left(\int_{\mathbb{R}} \gamma(W^{(1)}, c) |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) \right) \delta(x - c) d\nu_{\mathcal{C}}(c) \\ &= \int_{\text{supp}(\nu_{\mathcal{C}})} \left(\int_{\mathbb{R}} \gamma(W^{(1)}, c) |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) \right) \delta(x - c) p_{\mathcal{C}}(c) dc \\ &= p_{\mathcal{C}}(x) \int_{\mathbb{R}} \gamma(W^{(1)}, x) |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}). \end{aligned} \quad (\text{A84})$$

Proof of Theorem 6. First, if $x \notin \text{supp}(\zeta)$, similar to (A83), we have

$$\begin{aligned} g(x, (\bar{\gamma}, \bar{u}, \bar{v})) &= p_{\mathcal{C}}(x) \int_{\mathbb{R}} \gamma(W^{(1)}, x) |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\ &= 0. \end{aligned} \quad (\text{A85})$$

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

$$\begin{aligned} \min_{h \in C^2(\text{supp}(\zeta))} \quad & \int_{\text{supp}(\zeta)} \frac{(h''(x))^2}{\zeta(x)} dx \\ \text{subject to} \quad & h(x_j) = y_j, \quad j = 1, \dots, m. \end{aligned} \quad (\text{A86})$$

Let $L(f) = \int_{\text{supp}(\zeta)} \frac{(f''(x))^2}{p(x)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)} dx$. Then the functional $L(f)$ is strictly convex on space $\{f \in C^2(\mathbb{R}^2) | f(x_i) = y_i, i = 1, \dots, m\}$ when $m \geq 2$. It means that the minimizer of problem (A86) is unique.

Suppose $h(x)$ is the minimizer of problem (A86) and $h(x)$ is different from $g(x, (\bar{\gamma}, \bar{u}, \bar{v}))$ restricted on $\text{supp}(\zeta)$. Then by uniqueness of the solution,

$$L(h) < L(g(\cdot, (\bar{\gamma}, \bar{u}, \bar{v}))). \quad (\text{A87})$$

Now our goal is to find out another (γ, u, v) with smaller cost in problem (17). Then $(\bar{\gamma}, \bar{u}, \bar{v})$ is not the solution of (17), which is a contradiction. We set

$$\gamma(W^{(1)}, c) = \frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)}, \quad c \in \text{supp}(\zeta). \quad (\text{A88})$$

Then according to (A83),

$$\begin{aligned}
g''(x, \gamma) &= p(x) \int_{\mathbb{R}} \gamma(W^{(1)}, x) |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\
&= p(x) \int_{\mathbb{R}} \frac{h''(x)|W^{(1)}|}{p(x)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)} |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\
&= \frac{h''(x)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)} \int_{\mathbb{R}} |W^{(1)}|^2 d\nu_{\mathcal{W}|\mathcal{C}=x}(W^{(1)}) \\
&= \frac{h''(x)}{\mathbb{E}(\mathcal{W}^2|\mathcal{C}=x)} \mathbb{E}(\mathcal{W}^2|\mathcal{C}=x) \\
&= h''(x), \quad x \in \text{supp}(\zeta).
\end{aligned} \tag{A89}$$

It means that we can find $u, v \in \mathbb{R}$ such that $ux + v + g(x, \gamma) \equiv h(x)$. Then we find out (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 (17). Next we compute the cost of (γ, u, v) :

$$\begin{aligned}
&\int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c) d\nu(W^{(1)}, c) \\
&= \int_{\mathbb{R}^2} \left(\frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)} \right)^2 d\nu(W^{(1)}, c) \\
&= \int_{\text{supp}(\zeta)} \left(\int_{\mathbb{R}} \left(\frac{h''(c)|W^{(1)}|}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)} \right)^2 d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) \right) d\nu_{\mathcal{C}}(c) \\
&= \int_{\text{supp}(\zeta)} \left(\frac{h''(c)}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)} \right)^2 \left(\int_{\mathbb{R}} |W^{(1)}|^2 d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) \right) p_{\mathcal{C}}(c) dc \\
&= \int_{\text{supp}(\zeta)} \left(\frac{h''(c)}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)} \right)^2 \left(\int_{\mathbb{R}} |W^{(1)}|^2 d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) \right) p_{\mathcal{C}}(c) dc \\
&= \int_{\text{supp}(\zeta)} \frac{(h''(c))^2}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)} dx \\
&= L(h).
\end{aligned} \tag{A90}$$

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

$$\begin{aligned}
&\int_{\mathbb{R}^2} \bar{\gamma}^2(W^{(1)}, c) d\nu(W^{(1)}, c) \\
&= \int_{\text{supp}(\zeta)} \left(\int_{\mathbb{R}} \bar{\gamma}^2(W^{(1)}, c) d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) \right) p_{\mathcal{C}}(c) dc \\
&\geq \int_{\text{supp}(\zeta)} \frac{\left(\int_{\mathbb{R}} \bar{\gamma}(W^{(1)}, c) |W^{(1)}| d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) \right)^2}{\int_{\mathbb{R}} |W^{(1)}|^2 d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})} p_{\mathcal{C}}(c) dc \quad (\text{Cauchy-Schwarz inequality}) \\
&= \int_{\text{supp}(\zeta)} \frac{(g''(c, \bar{\gamma})/p_{\mathcal{C}}(c))^2}{\int_{\mathbb{R}} |W^{(1)}|^2 d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)})} p_{\mathcal{C}}(c) dc \quad (\text{according to (A83)}) \\
&= \int_{\text{supp}(\zeta)} \frac{(g''(c, \bar{\gamma}))^2}{p_{\mathcal{C}}(c)\mathbb{E}(\mathcal{W}^2|\mathcal{C}=c)} dc \\
&= L(g(\cdot, \bar{\gamma})) \\
&= L(g(\cdot, (\bar{\gamma}, \bar{u}, \bar{v}))) \quad (g(\cdot, (\bar{\gamma}, \bar{u}, \bar{v}))) \text{ has the same second derivative as } g(\cdot, \bar{\gamma}).
\end{aligned} \tag{A91}$$

Then

$$\begin{aligned}
&\int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c) d\nu(W^{(1)}, c) = L(h) \quad (\text{according to (A90)}) \\
&< L(g(\cdot, (\bar{\gamma}, \bar{u}, \bar{v}))) \quad (\text{according to (A87)}) \\
&\leq \int_{\mathbb{R}^2} \bar{\gamma}^2(W^{(1)}, c) d\nu(W^{(1)}, c) \quad (\text{according to (A91)}).
\end{aligned} \tag{A92}$$

It means that the cost of (γ, u, v) is smaller than the cost of $(\bar{\gamma}, \bar{u}, \bar{v})$. So $(\bar{\gamma}, \bar{u}, \bar{v})$ is not the solution of (17), which is a contradiction. So our assumption is wrong. So $h(x) \equiv g(x, (\bar{\gamma}, \bar{u}, \bar{v}))$ on $\text{supp}(\zeta)$, and $g(x, (\bar{\gamma}, \bar{u}, \bar{v}))$ is the solution of problem (A86). In the last step we prove that $g''(x, (\bar{\gamma}, \bar{u}, \bar{v})) = 0$ when $x \notin [\min_i x_i, \max_i x_i]$ and $g(x, (\bar{\gamma}, \bar{u}, \bar{v}))$ restricted on $\text{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$ is the solution of (19). We only need to prove these statements for $h(x)$, which is the solution of (A86).

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 (19), so $h(x)$ can be replaced by following $\tilde{h}(x)$ which also satisfies the constraints of problem (19):

$$\tilde{h}(x) = \begin{cases} h(x) & x \in [\min_i x_i, \max_i x_i] \\ h'(\min_i x_i)(x - \min_i x_i) + h(\min_i x_i) & x \in (-\infty, \min_i x_i) \\ h'(\max_i x_i)(x - \max_i x_i) + h(\max_i x_i) & x \in (\max_i x_i, \infty) \end{cases} \quad (\text{A93})$$

Then

$$\tilde{h}''(x) = \begin{cases} h''(x) & x \in [\min_i x_i, \max_i x_i] \\ 0 & x \in (-\infty, \min_i x_i) \\ 0 & x \in (\max_i x_i, \infty) \end{cases} \quad (\text{A94})$$

So the cost of $\tilde{h}(x)$ is less than that of $h(x)$. Then the fact $h(x)$ is the minimizer of (A86) 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 $S = \text{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$. Since $h(x)$ is the solution of the problem (A86), we get $h(x)|_S$ is the solution of the problem (19). \square

In the case of not using ASI, problem (17) becomes:

$$\begin{aligned} \min_{\gamma \in C(\mathbb{R}^2), u \in \mathbb{R}, v \in \mathbb{R}} \quad & \int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c) \, d\nu(W^{(1)}, c) \\ \text{subject to} \quad & ux_j + v + \int_{\mathbb{R}^2} \gamma(W^{(1)}, c)[W^{(1)}(x_j - c)]_+ \, d\nu(W^{(1)}, c) = y_j - f(x_j, \theta_0), \\ & j = 1, \dots, M. \end{aligned} \quad (\text{A95})$$

Then Theorem 6 without ASI is stated as follows.

Theorem A7 (Theorem 6 without ASI). *Suppose $(\bar{\gamma}, \bar{u}, \bar{v})$ is the solution of (A95), and consider the corresponding output function*

$$g(x, (\bar{\gamma}, \bar{u}, \bar{v})) = \bar{u}x + \bar{v} + \int_{\mathbb{R}^2} \bar{\gamma}(W^{(1)}, c)[W^{(1)}(x - c)]_+ \, d\nu(W^{(1)}, c) + f(x, \theta_0). \quad (\text{A96})$$

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

$$\begin{aligned} \min_{h \in C^2(S)} \quad & \int_S \frac{(h''(x) - f''(x, \theta_0))^2}{\zeta(x)} \, dx \\ \text{subject to} \quad & h(x_j) = y_j, \quad j = 1, \dots, M. \end{aligned} \quad (\text{A97})$$

J PROOF OF PROPOSITION 7 AND REMARKS ON PROPOSITION 8

Proof of Proposition 7. 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

$$p_{\mathcal{W}, \mathcal{C}}(W, C) = \left| \frac{\partial(W, -WC)}{\partial(W, C)} \right| p_{\mathcal{W}, \mathcal{B}}(W, -WC) = |W| p_{\mathcal{W}, \mathcal{B}}(W, -WC), \quad (\text{A98})$$

and

$$\begin{aligned} \mathbb{E}(W^2 | C = x) p_C(x) &= \int_{\mathbb{R}} W^2 p_{\mathcal{W}, \mathcal{C}}(W | C = x) \, dW p_C(x) \\ &= \int_{\mathbb{R}} W^2 p_{\mathcal{W}, \mathcal{C}}(W, x) \, dW \\ &= \int_{\mathbb{R}} |W|^3 p_{\mathcal{W}, \mathcal{B}}(W, -Wx) \, dW. \end{aligned} \quad (\text{A99})$$

□

Proof of Proposition 8. The construction is given in the statement of the proposition. □

Remark A8 (Remark to Proposition 8, sampling the initial parameters). The variables $(\mathcal{W}, \mathcal{B})$ can be sampled by first sampling C from $p_C(x) = \frac{1}{Z} \frac{1}{\varrho(x)}$, 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_C is the breakpoint density at initialization. The form of this has been studied for uniform initialization by Sahs et al. (2020). We provide the explicit form of the smoothness penalty function for several types of initialization in Appendix K.

Remark A9 (Remark to Proposition 8, 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, (A99) gives

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

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.

K PROOF OF THEOREM 9

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

Proof of Theorem 9. Gaussian initialization. Using (A99), we have

$$\begin{aligned} \mathbb{E}(W^2 | C = x) p_C(x) &= \int_{\mathbb{R}} |W|^3 p_{\mathcal{W}}(W) p_{\mathcal{B}}(-Wx) dW \\ &= \int_{\mathbb{R}} |W|^3 \frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{W^2}{2\sigma_w^2}} \frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{W^2 x^2}{2\sigma_b^2}} dW \\ &= \frac{1}{2\pi\sigma_w\sigma_b} \int_{\mathbb{R}} |W|^3 e^{-\left(\frac{1}{2\sigma_w^2} + \frac{x^2}{2\sigma_b^2}\right)W^2} dW \\ &= \frac{1}{2\pi\sigma_w\sigma_b} \int_{\mathbb{R}} |W|^3 e^{-\left(\frac{1}{2\sigma_w^2} + \frac{x^2}{2\sigma_b^2}\right)W^2} dW \end{aligned} \tag{A100}$$

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

$$\begin{aligned} \mathbb{E}(W^2 | C = x) p_C(x) &= \frac{\sigma}{\sqrt{2\pi}\sigma_w\sigma_b} \int_{\mathbb{R}} |W|^3 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{W^2}{2\sigma^2}} dW \\ &= \frac{\sigma}{\sqrt{2\pi}\sigma_w\sigma_b} \sigma^3 \cdot 2 \cdot \sqrt{\frac{2}{\pi}} \\ &= \frac{\sigma}{\sqrt{2\pi}\sigma_w\sigma_b} \sigma^3 \cdot 2 \cdot \sqrt{\frac{2}{\pi}} \\ &= \frac{2\sigma^4}{\pi\sigma_w\sigma_b} \\ &= \frac{2\sigma_w^3\sigma_b^3}{\pi(\sigma_b^2 + x^2\sigma_w^2)^2}. \end{aligned} \tag{A101}$$

Then

$$\begin{aligned} \zeta(x) &= \mathbb{E}(W^2 | C = x) p_C(x) \\ &= \frac{2\sigma_w^3\sigma_b^3}{\pi(\sigma_b^2 + x^2\sigma_w^2)^2}. \end{aligned} \tag{A102}$$

□

Proof of Theorem 9. Binary-uniform initialization. Since \mathcal{W} is either -1 or 1 , $\mathbb{E}(\mathcal{W}^2|\mathcal{C} = x) = 1$ for any $x \in \text{supp}(\nu_{\mathcal{C}})$. Since $\mathcal{B} \sim \mathcal{U}(-a_b, a_b)$, it is easy to check $-\mathcal{B}/\mathcal{W} \sim \mathcal{U}(-a_b, a_b)$. So $\zeta(x) = 1/2a_b$, $x \in [-a_b, a_b]$. □

Proof of Theorem 9. Uniform initialization. According to Theorem 1 in Sahs et al. (2020), the density function $p_{\mathcal{C}}(c)$ of $\nu_{\mathcal{C}}$ is

$$p_{\mathcal{C}}(c) = \frac{1}{4a_w a_b} \left(\min \left\{ \frac{a_b}{|c|}, a_w \right\} \right)^2, \quad c \in \text{supp}(\nu_{\mathcal{C}}). \quad (\text{A103})$$

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

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

$$\begin{aligned} p_{\mathcal{W}, \mathcal{C}}(W^{(1)}, c) &= p_{\mathcal{W}, \mathcal{B}}(W^{(1)}, -cW^{(1)}) \frac{\partial b}{\partial c} \\ &= \frac{1}{4a_w a_b} \mathbb{1}_{W^{(1)} \in [-a_w, a_w]} \cdot \mathbb{1}_{-cW^{(1)} \in [-a_b, a_b]} \cdot (-W^{(1)}) \end{aligned} \quad (\text{A104})$$

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

$$\begin{aligned} p_{\mathcal{W}|\mathcal{C}}(W^{(1)}|c) &= \frac{p_{\mathcal{W}, \mathcal{C}}(W^{(1)}, c)}{p_{\mathcal{C}}(c)} \\ &= \frac{\frac{1}{4a_w a_b} \mathbb{1}_{W^{(1)} \in [-a_w, a_w]} \cdot \mathbb{1}_{-cW^{(1)} \in [-a_b, a_b]} \cdot (-W^{(1)})}{p_{\mathcal{C}}(c)} \end{aligned} \quad (\text{A105})$$

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 $\frac{a_b}{a_w} \geq L$, $\mathbb{E}(\mathcal{W}^2|\mathcal{C} = c)$ and $p_{\mathcal{C}}(c)$ are constant when $c \in [-L, L]$. Then $\zeta(x) = \mathbb{E}(\mathcal{W}^2|\mathcal{C} = x)p_{\mathcal{C}}(x)$ is constant when $c \in [-L, L]$. □

L DIFFERENCE BETWEEN SOLUTIONS OF VARIATIONAL PROBLEMS (15) AND (17)

In this section, we show that the solution of the variational problem with linearly adjusted training data (17) is close to the solution of training with the original training data (15). 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 (15):

$$\begin{aligned} &\min_{\alpha \in C(\mathbb{R}^2)} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) d\mu(W^{(1)}, b) \\ &\text{subject to} \quad \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)}x_j + b]_+ d\mu(W^{(1)}, b) = y_j, \quad j = 1, \dots, M, \end{aligned} \quad (\text{A106})$$

and the linearly adjusted variational problem (here we don't replace α by γ to make comparison between (A106) and (A107) more clear):

$$\begin{aligned} &\min_{\alpha \in C(\mathbb{R}^2), u \in \mathbb{R}, v \in \mathbb{R}} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) d\mu(W^{(1)}, b) \\ &\text{subject to} \quad ux_j + v + \int_{\mathbb{R}^2} \alpha(W^{(1)}, b)[W^{(1)}x_j + b]_+ d\mu(W^{(1)}, b) = y_j, \quad j = 1, \dots, M. \end{aligned} \quad (\text{A107})$$

In this paper, our main focus is on the variational problem (A107), thus we derive our main result Theorem 1 which is a statement on linearly adjusted training data. In this section, we try to analyze the difference between solutions of variational problems (A106) and (A107), and thus show that to what extent the variational problem (4) in Theorem 4 describes the implicit bias of gradient descent on original training data.

Suppose the solution of problem (A106) is $\bar{\alpha}_1$, and the corresponding output function is

$$g(x, \bar{\alpha}_1) = \int_{\mathbb{R}^2} \bar{\alpha}_1(W^{(1)}, b) [W^{(1)}x + b]_+ d\mu(W^{(1)}, b). \quad (\text{A108})$$

The solution of problem (A107) is $(\bar{\alpha}_2, \bar{u}, \bar{v})$ and the corresponding output function is.

$$g(x, (\bar{\alpha}_2, \bar{u}, \bar{v})) = \bar{u}x + \bar{v} + \int_{\mathbb{R}^2} \bar{\alpha}_2(W^{(1)}, b) [W^{(1)}x + b]_+ d\mu(W^{(1)}, b). \quad (\text{A109})$$

Our goal is to show that $g(x, \bar{\alpha}_1)$ and $g(x, (\bar{\alpha}_2, \bar{u}, \bar{v}))$ are close to each other. Since $(\bar{\alpha}_2, \bar{u}, \bar{v})$ is the minimizer of (A106), we have that $(\bar{\alpha}_1, 0, 0)$ is a feasible solution of (A106) but not optimal, which means

$$\int_{\mathbb{R}^2} \bar{\alpha}_1^2(W^{(1)}, b) d\mu(W^{(1)}, b) \geq \int_{\mathbb{R}^2} \bar{\alpha}_2^2(W^{(1)}, b) d\mu(W^{(1)}, b) \quad (\text{A110})$$

Suppose that the straight line $\bar{u}x + \bar{v}$ can be fitted by an infinite width network with parameters α_s , i.e.

$$\int_{\mathbb{R}^2} \alpha_s(W^{(1)}, b) [W^{(1)}x + b]_+ d\mu(W^{(1)}, b) = \bar{u}x + \bar{v} \quad (\text{A111})$$

Then $\bar{\alpha}_2 + \alpha_s$ is a feasible solution of the problem (A106). It is easy to show that $g(x, \bar{\alpha}_2 + \alpha_s) = g(x, (\bar{\alpha}_2, \bar{u}, \bar{v}))$. So we only need to measure the difference between $g(x, \bar{\alpha}_1)$ and $g(x, \bar{\alpha}_2 + \alpha_s)$, which can be characterized by the difference between $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$. From the optimality of α_1 , we have

$$\int_{\mathbb{R}^2} \bar{\alpha}_1^2(W^{(1)}, b) d\mu(W^{(1)}, b) \leq \int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2(W^{(1)}, b) d\mu(W^{(1)}, b) \quad (\text{A112})$$

Using the first order optimality condition on the problem (A106), we have that there exist $\lambda_j \in \mathbb{R}$ such that

$$\alpha_1(W^{(1)}, b) = \sum_{j=1}^M \lambda_j [W^{(1)}x + b]_+. \quad (\text{A113})$$

Since both $\bar{\alpha}_1$ and $\bar{\alpha}_2 + \alpha_s$ are the feasible solutions of the problem (A192),

$$\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \cdot [W^{(1)}x_j + b]_+ d\mu(W^{(1)}, b) = 0, \quad j = 1, \dots, M. \quad (\text{A114})$$

Using (A113) and (A114), we have

$$\begin{aligned} & \int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \bar{\alpha}_1 d\mu(W^{(1)}, b) \\ &= \int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \sum_{j=1}^M \lambda_j [W^{(1)}x + b]_+ d\mu(W^{(1)}, b) \\ &= \sum_{j=1}^M \lambda_j \int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s) \cdot [W^{(1)}x + b]_+ d\mu(W^{(1)}, b) \\ &= 0. \end{aligned} \quad (\text{A115})$$

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

$$\begin{aligned}
& \int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s)^2 d\mu(W^{(1)}, b) \\
&= \int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - (2\bar{\alpha}_2 + 2\alpha_s - \bar{\alpha}_1)\bar{\alpha}_1 d\mu(W^{(1)}, b) \\
&= \int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - \bar{\alpha}_1^2 + (2\bar{\alpha}_2 + 2\alpha_s - 2\bar{\alpha}_1)\bar{\alpha}_1 d\mu(W^{(1)}, b) \\
&= \int_{\mathbb{R}^2} (\bar{\alpha}_2 + \alpha_s)^2 - \bar{\alpha}_1^2 d\mu(W^{(1)}, b) \quad (\text{use (A115)}) \\
&= \int_{\mathbb{R}^2} (\bar{\alpha}_2^2 + 2\bar{\alpha}_2\alpha_s + \alpha_s^2) - \bar{\alpha}_1^2 d\mu(W^{(1)}, b) \\
&\leq \int_{\mathbb{R}^2} (\bar{\alpha}_1^2 + 2\bar{\alpha}_2\alpha_s + \alpha_s^2) - \bar{\alpha}_1^2 d\mu(W^{(1)}, b) \quad (\text{use (A110)}) \\
&\leq \int_{\mathbb{R}^2} 2\bar{\alpha}_2\alpha_s + \alpha_s^2 d\mu(W^{(1)}, b) \\
&\leq 2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_2^2 d\mu(W^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b) \\
&\leq 2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b) \quad (\text{use (A110)}).
\end{aligned} \tag{A116}$$

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

$$\begin{aligned}
& \frac{\int_{\mathbb{R}^2} (\bar{\alpha}_1 - \bar{\alpha}_2 - \alpha_s)^2 d\mu(W^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)} \\
&\leq \frac{2\sqrt{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b) \cdot \int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)} + \int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)} \\
&= 2\sqrt{\frac{\int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)}} + \frac{\int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)}{\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)}.
\end{aligned} \tag{A117}$$

It means that if $\int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)$ is much smaller than $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)$, the relative distance 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 $\int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)$ is indeed much smaller than $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)$ when the training data is not highly linearly correlated.

Here we compute the value of $\int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)$ and $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)$ in a concrete example. The training data we use are $\{(-2, 1.5), (-1.6, 0.5), (0.3, 1.5), (0.6, 0.5), (2, 1.5)\}$. We use uniform initialization $W \sim U(-1, 1)$, $B \sim U(-2, 2)$, the same as in Figure 1. By solving the variational problem (A107), we get $\bar{u} = -0.2228$ and $\bar{v} = -4.0838$. We choose $\alpha_s(W^{(1)}, b) = 6\bar{u}W^{(1)} + 1.5\bar{v}b$. It is easy to verify that the above construction of α_s satisfies (A111). Then we compute the integral of α_s^2 and get $\int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b) = 50.9652$. By solving the variational problem (A106), we can get the numerical solution of $\bar{\alpha}_1$. Then we compute the integral of $\bar{\alpha}_1^2$ and get $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b) = 1462.85$. We can see that $\int_{\mathbb{R}^2} \alpha_s^2 d\mu(W^{(1)}, b)$ is much smaller than $\int_{\mathbb{R}^2} \bar{\alpha}_1^2 d\mu(W^{(1)}, b)$.

M EQUIVALENCE OF OUR CHARACTERIZATION AND NTK NORM MINIMIZATION

In this section we demonstrate that NTK norm minimization (Zhang et al., 2019), which characterizes the implicit bias of training a linearized model by gradient descent, is equivalent to our characterization in Section 5. Following Jacot et al. (2018), Zhang et al. (2019) 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.

$$\begin{aligned}\tilde{\Theta}_n(x_1, x_2) &= \frac{1}{n} \nabla_{W^{(2)}} f(x_1, \theta_0) \nabla_{W^{(2)}} f(x_2, \theta_0)^T \\ &= \frac{1}{n} \sum_{i=1}^n \nabla_{W_i^{(2)}} f(x_1, \theta_0) \nabla_{W_i^{(2)}} f(x_2, \theta_0) \\ &= \frac{1}{n} \sum_{i=1}^n [W_i^{(1)} x_1 + b_i^{(1)}]_+ [W_i^{(1)} x_2 + b_i^{(1)}]_+.\end{aligned}\tag{A118}$$

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

$$\tilde{\Theta}(x_1, x_2) = \int_{\mathbb{R}^2} [W^{(1)} x_1 + b^{(1)}]_+ [W^{(1)} x_2 + b^{(1)}]_+ d\mu(W^{(1)}, b).\tag{A119}$$

Equivalently, using the notation in Section 5.2, we have

$$\tilde{\Theta}(x_1, x_2) = \int_{\mathbb{R}^2} [W^{(1)}(x_1 - c)]_+ [W^{(1)}(x_2 - c)]_+ d\nu(W^{(1)}, c).\tag{A120}$$

Next, Zhang et al. (2019) 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:

$$(i) \quad \forall x \in S, \tilde{\Theta}(\cdot, x) \in \mathcal{H}_{\tilde{\Theta}}(S);\tag{A121}$$

$$(ii) \quad \forall x \in S, \forall f \in \mathcal{H}_{\tilde{\Theta}}, \langle f(\cdot), \tilde{\Theta}(\cdot, x) \rangle_{\tilde{\Theta}} = f(x);\tag{A122}$$

$$(iii) \quad \forall x, y \in S, \langle \tilde{\Theta}(\cdot, x), \tilde{\Theta}(\cdot, y) \rangle_{\tilde{\Theta}} = \tilde{\Theta}(x, y).\tag{A123}$$

Here the domain is $S = \text{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$, which is the same as in Theorem 1 and Theorem 6. Using the reproducing kernel Hilbert space, Zhang et al. (2019) prove that $f^{\text{lin}}(x, \tilde{\omega}_\infty)$ (defined in Section 4.2) is the solution of the following optimization problem:

$$\min_{g \in \mathcal{H}_{\tilde{\Theta}}(S)} \|g\|_{\tilde{\Theta}_n} \quad \text{s.t. } g(x_j) = y_j, \quad j = 1, \dots, M.\tag{A124}$$

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

$$\min_{g \in \mathcal{H}_{\tilde{\Theta}}(S)} \|g\|_{\tilde{\Theta}} \quad \text{s.t. } g(x_j) = y_j, \quad j = 1, \dots, M.\tag{A125}$$

In Section 5, we show that $f^{\text{lin}}(x, \tilde{\omega}_\infty)$ is the solution of the optimization problem (15) in function space. As width n tends to infinity, the optimization problem (15) becomes (A63), which we repeat below:

$$\begin{aligned}\min_{\alpha \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) d\mu(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha(W^{(1)}, b) [W^{(1)} x_j + b]_+ d\mu(W^{(1)}, b) = y_j, \quad j = 1, \dots, M.\end{aligned}\tag{A126}$$

Since optimization problems (A125) and (A126) 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 A10 (Equivalence of our variational problem and NTK norm minimization). *Assume that optimization problems (A125) and (A126) are both feasible. Suppose $\bar{\alpha}$ is the solution of (A126), and consider the corresponding output function:*

$$\bar{g}(x) = \int_{\mathbb{R}^2} \bar{\alpha}(W^{(1)}, b) [W^{(1)} x + b]_+ d\mu(W^{(1)}, b).\tag{A127}$$

Then $\bar{g}(x)$ restricted on S is the solution of the optimization problem (A125).

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

Proof of Theorem A10. Since $\bar{\alpha}(W^{(1)}, b)$ is the solution of (A126), according to (A71) in the proof of Theorem 5,

$$\bar{\alpha}(W^{(1)}, b) = -\frac{1}{2} \sum_{j=1}^M \lambda_j [W^{(1)} x_j + b]_+ \quad \forall (W^{(1)}, b) \in \mathbb{R}^2 \quad (\text{A128})$$

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

$$\bar{\alpha}(W^{(1)}, b) = \int_S h(x) [W^{(1)} x + b]_+ dx, \quad (\text{A129})$$

where $h(x)$ can be a combination of Dirac delta functions. Then substitute (A129) into the expression of $\bar{g}(x)$ (A127) to obtain

$$\begin{aligned} \bar{g}(x) &= \int_{\mathbb{R}^2 \times S} h(\tilde{x}) [W^{(1)} \tilde{x} + b]_+ [W^{(1)} x + b]_+ d\mu(W^{(1)}, b) d\tilde{x} \\ &= \int_S h(\tilde{x}) \tilde{\Theta}(x, \tilde{x}) d\tilde{x}, \end{aligned} \quad (\text{A130})$$

where we use the expression of the NTK in equation (A119). Then

$$\begin{aligned} \langle g(x), g(x) \rangle_{\bar{\Theta}} &= \langle g(x), \int_S h(\tilde{x}) \tilde{\Theta}(x, \tilde{x}) d\tilde{x} \rangle_{\bar{\Theta}} \\ &= \int_S h(\tilde{x}) \langle g(x), \tilde{\Theta}(x, \tilde{x}) \rangle_{\bar{\Theta}} d\tilde{x} \\ &= \int_S h(\tilde{x}) g(\tilde{x}) d\tilde{x} \quad (\text{here we use the property of RKHS norm (A122)}) \\ &= \int_{S \times S} h(\tilde{x}) h(\bar{x}) \tilde{\Theta}(\tilde{x}, \bar{x}) d\tilde{x} d\bar{x} \quad (\text{use (A130)}). \end{aligned} \quad (\text{A131})$$

On the other hand, using (A129), the objective of (A126) becomes

$$\begin{aligned} &\int_{S^2} \bar{\alpha}^2(W^{(1)}, b) d\mu(W^{(1)}, b) \\ &= \int_{S \times S \times \mathbb{R}^2} h(\tilde{x}) [W^{(1)} \tilde{x} + b]_+ h(\bar{x}) [W^{(1)} \bar{x} + b]_+ d\tilde{x} d\bar{x} d\mu(W^{(1)}, b) \\ &= \int_{S \times S} h(\tilde{x}) h(\bar{x}) \int_{\mathbb{R}^2} [W^{(1)} \tilde{x} + b]_+ [W^{(1)} \bar{x} + b]_+ d\mu(W^{(1)}, b) d\tilde{x} d\bar{x} \\ &= \int_{S \times S} h(\tilde{x}) h(\bar{x}) \tilde{\Theta}(\tilde{x}, \bar{x}) d\tilde{x} d\bar{x} \quad (\text{use (A119)}). \end{aligned} \quad (\text{A132})$$

Comparing (A131) and (A132), we have that optimization problems (A125) and (A126) are equivalent if $\alpha(W^{(1)}, b)$ has the form (A129) and $g(x)$ has the form (A130). Moreover, if every function $g \in \mathcal{H}_{\bar{\Theta}}(S)$ can be approximated by the shallow network, we can find $\alpha(W^{(1)}, b)$ in form of (A129) such that $g(x)$ is expressed in the form of (A130). In this sense we show that optimization problems (A125) and (A126) are equivalent. \square

In Section 5.2, we relax the optimization problem (16) to (17) 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 (A125) to the following problem:

$$\min_{g \in \mathcal{H}_{\bar{\Theta}}(S), u \in \mathbb{R}, v \in \mathbb{R}} \|g - ux - v\|_{\bar{\Theta}} \quad \text{s.t. } g(x_j) = y_j, \quad j = 1, \dots, M. \quad (\text{A133})$$

Then the optimization problems (17) and (A133) are equivalent. Theorem 6 shows that (17) and (19) have the same solution on the set $S = \text{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$. Then we have that optimization problems (A133) and (19) are equivalent, which means that

$$\min_{u \in \mathbb{R}, v \in \mathbb{R}} \|g - ux - v\|_{\tilde{\Theta}} = \int_S \frac{(g''(x))^2}{\zeta(x)} dx, \quad \forall g \in \mathcal{H}_{\tilde{\Theta}}(S). \quad (\text{A134})$$

Next, we directly prove the above equation (A134). Given function $g \in \mathcal{H}_{\tilde{\Theta}}(S)$, let $h = \text{argmin}_{h \in \mathcal{H}_{\tilde{\Theta}}(S)} \|h\|_{\tilde{\Theta}}$, 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 $W = \{h \in \mathcal{H}_{\tilde{\Theta}}(S) : \langle h, x \rangle_{\tilde{\Theta}} = 0, \langle h, 1 \rangle_{\tilde{\Theta}} = 0\}$, which is the orthogonal complement of $\text{span}\{1, x\}$. Then h is the projection of g on W . Since $h = g - ux - v$, $h'' = g''$. So we can reformulate the equation (A134) which we want to prove in the following theorem:

Theorem A11 (Explicit form of the kernel norm). *The kernel norm on the space $W = \{h \in \mathcal{H}_{\tilde{\Theta}}(S) : \langle h, x \rangle_{\tilde{\Theta}} = 0, \langle h, 1 \rangle_{\tilde{\Theta}} = 0\}$ is given as follows:*

$$\|h\|_{\tilde{\Theta}}^2 = \int_S \frac{(h''(x))^2}{\zeta(x)} dx, \quad \forall h \in W. \quad (\text{A135})$$

This theorem gives the explicit form of the kernel norm in a subspace of $\mathcal{H}_{\tilde{\Theta}}(S)$. Next we prove the above theorem using the property of kernel norm.

Proof of Theorem A11. Let $\tilde{\Theta}_x(\cdot) = \tilde{\Theta}(\cdot, x)$. We can find the orthogonal projection of $\tilde{\Theta}_x$ on space W , which we denote by $\tilde{\Theta}_{x,W}$. Then we only need to prove that $\langle h, \tilde{\Theta}_{x,W} \rangle_{\tilde{\Theta}} = \int_S \frac{h''(y)\tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy$ for any $h \in W$ and $x \in S$.

First, $\tilde{\Theta}_{x,W} = \tilde{\Theta}_x - ux - v$ for some constant $u, v \in \mathbb{R}$. Since $h \in W$, $\langle h, 1 \rangle_{\tilde{\Theta}} = 0$ and $\langle h, x \rangle_{\tilde{\Theta}} = 0$. Then

$$\begin{aligned} \langle h, \tilde{\Theta}_{x,W} \rangle_{\tilde{\Theta}} &= \langle h, \tilde{\Theta}_x - ux - v \rangle_{\tilde{\Theta}} \\ &= \langle h, \tilde{\Theta}_x \rangle_{\tilde{\Theta}} - u \langle h, x \rangle_{\tilde{\Theta}} - v \langle h, 1 \rangle_{\tilde{\Theta}} \\ &= \langle h, \tilde{\Theta}_x \rangle_{\tilde{\Theta}} \\ &= h(x) \quad (\text{use the reproducing property of the kernel (A122)}). \end{aligned} \quad (\text{A136})$$

Next, using the notation from Section 5.2 we have

$$\begin{aligned}
\tilde{\Theta}_{x,W}''(y) &= (\tilde{\Theta}_x(y) - uy - v)'' \\
&= \tilde{\Theta}_x(y)'' \\
&= \frac{\partial^2}{\partial y^2} \tilde{\Theta}(x, y) \\
&= \frac{\partial^2}{\partial y^2} \int_{\mathbb{R}^2} [W^{(1)}(x - c)]_+ [W^{(1)}(y - c)]_+ d\nu(W^{(1)}, c) \quad (\text{use (A120)}) \\
&= \frac{\partial^2}{\partial y^2} \int_{\mathbb{R}^2} (W^{(1)})^2 [\text{sign}(W^{(1)})(x - c)]_+ [\text{sign}(W^{(1)})(y - c)]_+ d\nu_{\mathcal{W}|\mathcal{C}=c}(W^{(1)}) d\nu_{\mathcal{C}}(c) \\
&= \frac{\partial^2}{\partial y^2} \int_{\mathbb{R}} (\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = c) [x - c]_+ [y - c]_+ \\
&\quad + \mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = c) [c - x]_+ [c - y]_+) p_{\mathcal{C}}(c) dc \\
&= \int_{\mathbb{R}} \left(\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = c) [x - c]_+ \frac{\partial^2}{\partial y^2} [y - c]_+ \right. \\
&\quad \left. + \mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = c) [c - x]_+ \frac{\partial^2}{\partial y^2} [c - y]_+ \right) p_{\mathcal{C}}(c) dc \\
&= \int_{\mathbb{R}} (\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = c) [x - c]_+ \delta(y - c) \\
&\quad + \mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = c) [c - x]_+ \delta(y - c)) p_{\mathcal{C}}(c) dc \\
&= (\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = y) [x - y]_+ + \mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = y) [y - x]_+) p_{\mathcal{C}}(y). \tag{A137}
\end{aligned}$$

Then

$$\begin{aligned}
&\int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy \\
&= \int_S \frac{h''(y) (\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = y) [x - y]_+ + \mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = y) [y - x]_+) p_{\mathcal{C}}(y)}{\zeta(y)} dy \\
&= \int_S \frac{h''(y) (\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = y) [x - y]_+ + \mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = y) [y - x]_+)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = y)} dy \\
&= \int_S \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = y)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = y)} h''(y) [x - y]_+ + \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = y)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = y)} h''(y) [y - x]_+ dy. \tag{A138}
\end{aligned}$$

Now we regard $\int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy$ as a function of x , then

$$\begin{aligned}
&\frac{\partial^2}{\partial x^2} \int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy \\
&= \frac{\partial^2}{\partial x^2} \int_S \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = y)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = y)} h''(y) [x - y]_+ + \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = y)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = y)} h''(y) [y - x]_+ dy \\
&= \int_S \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = y)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = y)} h''(y) \delta(x - y) + \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = y)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = y)} h''(y) \delta(y - x) dy \\
&= \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} \geq 0) | \mathcal{C} = x)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = x)} h''(x) + \frac{\mathbb{E}(\mathcal{W}^2 \mathbb{1}(\mathcal{W} < 0) | \mathcal{C} = x)}{\mathbb{E}(\mathcal{W}^2 | \mathcal{C} = x)} h''(x) \\
&= h''(x). \tag{A139}
\end{aligned}$$

From the definition of the space W , we see that the second derivative uniquely determines the element in W . Since $h \in W$, in order to show that $\int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy = h(x)$, we only need to show $\int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy \in W$, i.e. $\langle \int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy, 1 \rangle_{\tilde{\Theta}} = 0$ and $\langle \int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy, x \rangle_{\tilde{\Theta}} = 0$.

Then

$$\begin{aligned}
\langle \int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy, 1 \rangle_{\tilde{\Theta}} &= \langle \int_S \frac{h''(y) \frac{\partial^2}{\partial y^2} \tilde{\Theta}(x, y)}{\zeta(y)} dy, 1 \rangle_{\tilde{\Theta}} \\
&= \langle \int_S \frac{h''(y) \lim_{h \rightarrow 0} \frac{\tilde{\Theta}(x, y+h) - 2\tilde{\Theta}(x, y) + \tilde{\Theta}(x, y-h)}{h^2}}{\zeta(y)} dy, 1 \rangle_{\tilde{\Theta}} \\
&= \lim_{h \rightarrow 0} \langle \int_S \frac{h''(y) \frac{\tilde{\Theta}(x, y+h) - 2\tilde{\Theta}(x, y) + \tilde{\Theta}(x, y-h)}{h^2}}{\zeta(y)} dy, 1 \rangle_{\tilde{\Theta}} \\
&= \lim_{h \rightarrow 0} \int_S \frac{h''(y) \frac{\langle \tilde{\Theta}(x, y+h), 1 \rangle_{\tilde{\Theta}} - 2\langle \tilde{\Theta}(x, y), 1 \rangle_{\tilde{\Theta}} + \langle \tilde{\Theta}(x, y-h), 1 \rangle_{\tilde{\Theta}}}{h^2}}{\zeta(y)} dy \\
&= \lim_{h \rightarrow 0} \int_S \frac{h''(y) \frac{y+h-2y+y-h}{h^2}}{\zeta(y)} dy \\
&= 0.
\end{aligned} \tag{A140}$$

Similarly we can show that $\langle \int_S \frac{h''(y) \tilde{\Theta}_{x,W}''(y)}{\zeta(y)} dy, x \rangle_{\tilde{\Theta}} = 0$. This concludes the proof. \square

N GRADIENT DESCENT TRAJECTORY AND TRAJECTORY OF SMOOTHING SPLINES

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 $E(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$, where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M]^T$ 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:

$$\begin{aligned}
E(\mathbf{w}) &= \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \\
&= \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{y}^T \mathbf{X} \mathbf{w} + \mathbf{y}^T \mathbf{y} \\
&= \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{y}^T \mathbf{X} \mathbf{w} + \mathbf{y}^T \mathbf{y} \\
&= \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T H (\mathbf{w} - \mathbf{w}^*) + E_0,
\end{aligned} \tag{A141}$$

where $H = 2\mathbf{X}^T \mathbf{X}$, E_0 is the minimum of the loss function, and \mathbf{w}^* is the minimizer. The eigenvalues and eigenvectors of H are as follows:

$$H \mathbf{u}_j = \lambda_j \mathbf{u}_j. \tag{A142}$$

Then expand \mathbf{w} and \mathbf{w}^* in terms of the eigenvectors of H :

$$\mathbf{w} = \sum_j w_j \mathbf{u}_j, \quad \mathbf{w}^* = \sum_j w_j^* \mathbf{u}_j. \tag{A143}$$

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

$$\tilde{w}_j = \frac{\lambda_j}{\lambda_j + c} w_j^*. \tag{A144}$$

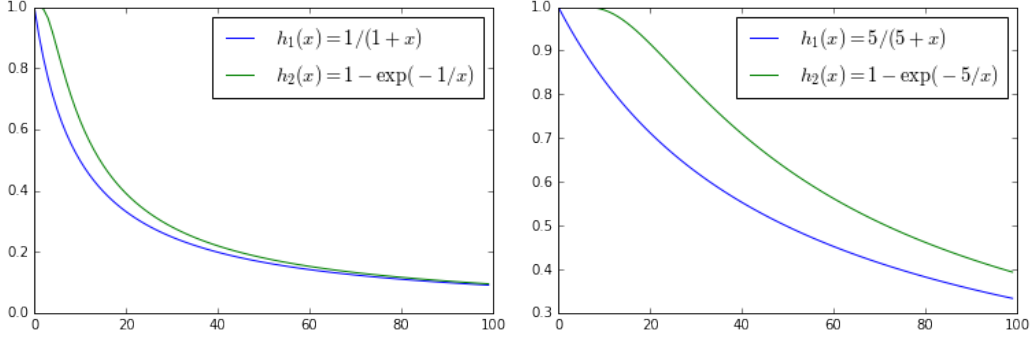


Figure A7: Plot of functions $h_1(x)$ and $h_2(x)$. The left panel plots the two function when $\lambda_j = 1$. The right panel plots the two function when $\lambda_j = 5$.

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

$$\begin{aligned}\mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)} - \eta \nabla E \\ &= \mathbf{w}^{(\tau-1)} - \eta H(\mathbf{w}^{(\tau-1)} - \mathbf{w}^*), \\ \mathbf{w}^{(0)} &= \mathbf{0}.\end{aligned}\tag{A145}$$

Writing $\mathbf{w}^{(\tau)} = \sum_j w_j^{(\tau)} \mathbf{u}_j$, then

$$w_j^{(\tau)} = (1 - (1 - \eta \lambda_j)^\tau) w_j^*.\tag{A146}$$

Note that $1 - (1 - \eta \lambda_j)^\tau \rightarrow 1 - e^{-\eta \tau \lambda_j}$ as $\eta \rightarrow 0$. Hence choosing a sufficiently small learning rate, approximately we have

$$w_j^{(\tau)} = (1 - e^{-\eta \tau \lambda_j}) w_j^*.\tag{A147}$$

From (A144) and (A147), 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, $c \|\mathbf{w} - \bar{\mathbf{w}}\|_2^2$, and the gradient descent iteration is initialized at a point different from zero, $\mathbf{w}^{(0)} = \bar{\mathbf{w}}$.

Now we want to take a closer look at the trajectories. Consider the following two functions:

$$h_1(x) = \frac{\lambda_j}{\lambda_j + x}, \quad h_2(x) = 1 - e^{-\lambda_j/x}.\tag{A148}$$

Actually we can verify that $h_1(0) = h_2(0) = 1$ and $\lim_{x \rightarrow \infty} \frac{h_1(x)}{h_2(x)} = 1$. It implies that these two functions are close to each other on $[0, \infty)$. Figure A7 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 (A144) and (A147), 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 (12) here:

$$\tilde{W}_0^{(2)} = \bar{W}^{(2)}, \quad \tilde{W}_{t+1}^{(2)} = \tilde{W}_t^{(2)} - \eta \nabla_{W^{(2)}} L^{\text{lin}}(\tilde{\omega}_t).\tag{A149}$$

It is actually minimizing the following loss function of $W^{(2)} - \bar{W}$:

$$E(W^{(2)} - \bar{W}) = \sum_{j=1}^M \left(\sum_{i=1}^n (W_i^{(2)} - \bar{W}_i^{(2)}) [W_i^{(1)} x_j + b_i]_+ - (y_j - f(x_j, \theta_0)) \right)^2.\tag{A150}$$

Here we change the variable from $W^{(2)}$ to $W^{(2)} - \bar{W}$. Then $W_t^{(2)} - \bar{W} = 0$ when $t = 0$, so that gradient descent start from the zero initial weight vector. Since the above model is linear with respect

to $W^{(2)} - \bar{W}$, 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 $\bar{W}_t^{(2)}$ at iteration t is close to the minimizer of the following regularized optimization problem:

$$\min_{W^{(2)}} \sum_{j=1}^M \left(\sum_{i=1}^n (W_i^{(2)} - \bar{W}_i^{(2)}) [W_i^{(1)} x_j + b_i]_+ - (y_j - f(x_j, \theta_0)) \right)^2 + c \|W^{(2)} - \bar{W}\|_2^2, \quad (\text{A151})$$

where $c = \frac{1}{\eta_n t}$. Using the same approach and notation as in Section 5, the optimization problem (A151) is equivalent to

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \quad & \sum_{j=1}^M \left(\int_{\mathbb{R}^2} \alpha_n(W^{(1)}, b) [W^{(1)} x_j + b]_+ d\mu_n(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{n\eta_n t} \int_{\mathbb{R}^2} \alpha_n^2(W^{(1)}, b) d\mu_n(W^{(1)}, b), \end{aligned} \quad (\text{A152})$$

where we use the ASI trick (see Appendix C.2). Here (A152) has an extra factor $\frac{1}{n}$ compared to (A151). This is because we define $\alpha_n(W_i^{(1)}, b_i) = n(W_i^{(2)} - \bar{W}_i^{(2)})$. According to Theorem A1, $\eta_n \leq \frac{1}{Kn\sqrt{M}\lambda_{\max}(\bar{\Theta}_n)}$ 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 A1 is satisfied. The limit of the optimization problem (A152) as the width n tends to infinity is:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \sum_{j=1}^M \left(\int_{\mathbb{R}^2} \alpha(W^{(1)}, b) [W^{(1)} x_j + b]_+ d\mu(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{\bar{\eta}t} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) d\mu(W^{(1)}, b). \end{aligned} \quad (\text{A153})$$

Following the same reasoning of Section 5.2, we relax the optimization problem (A153) to the following one:

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2), u \in \mathbb{R}, v \in \mathbb{R}} \quad & \sum_{j=1}^M \left(u x_j + v + \int_{\mathbb{R}^2} \alpha(W^{(1)}, b) [W^{(1)} x_j + b]_+ d\mu(W^{(1)}, b) - y_j \right)^2 \\ & + \frac{1}{\bar{\eta}t} \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) d\mu(W^{(1)}, b). \end{aligned} \quad (\text{A154})$$

Using the same technique and notation as in Theorem 6, we can prove that the solution of (A154) actually solves the following optimization problem:

$$\min_{h \in C^2(S)} \sum_{j=1}^M [h(x_j) - y_j]^2 + \frac{1}{\bar{\eta}t} \int_S \frac{(h''(x))^2}{\zeta(x)} dx. \quad (\text{A155})$$

Then in order to study the trajectory of gradient descent, we can study the optimization problem (A155) with varying t . Figure A8 illustrates smoothing spline and gradient descent trajectories. The solution of (A155) is called spatially adaptive smoothing spline. Here the curvature penalty function is $\frac{1}{\bar{\eta}t} \frac{1}{\zeta(x)}$, with time dependent smoothness regularization coefficient $\frac{1}{\bar{\eta}t}$. Next, we give out the solution of (A155) 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 A12 (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.

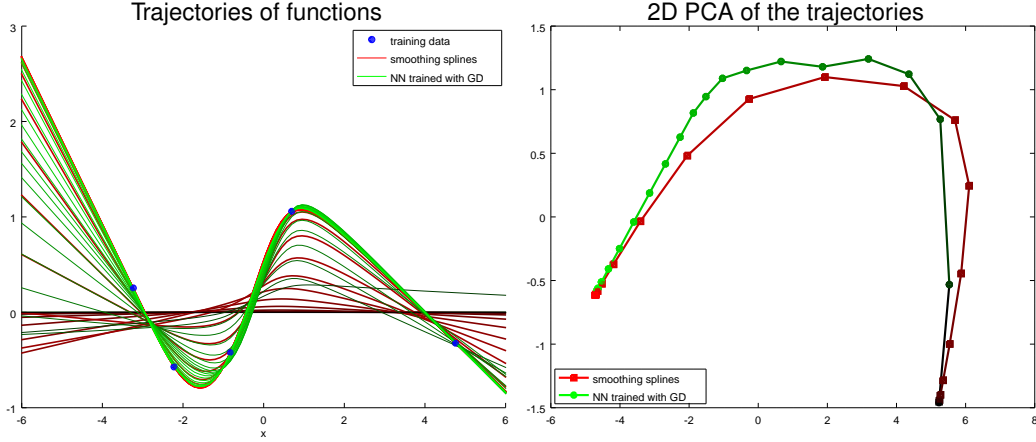


Figure A8: Trajectories of functions obtained by gradient descent training a neural network and by smoothing splines of the training data with decreasing regularization strength (from dark to bright). The left panel plots 20 functions along each trajectory. The right panel shows the same functions in 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 $\mathcal{W} \sim N(0, 1)$, $\mathcal{B} \sim N(0, 1)$. The results are similar for smaller networks and different initializations.

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 $\lambda = \frac{1}{\eta t z}$. Then (A155) becomes the following optimization problem:

$$\min_{h \in C^2(S)} \sum_{j=1}^M [h(x_j) - y_j]^2 + \lambda \int_S (h''(x))^2 dx. \quad (\text{A156})$$

German (2001) gives the explicit form of the minimizer \hat{h} of (A156), which is called a smoothing spline. The minimizer \hat{h} is a natural cubic spline with knots at the sample points x_1, \dots, 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

$$(\hat{h}(x_1), \dots, \hat{h}(x_M))^T = (I + \lambda A)^{-1} Y. \quad (\text{A157})$$

The matrix A has entries $A_{ij} = \int_S h_i''(x) h_j''(x) dx$, 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 out a rather explicit form of matrix A , which is an $M \times M$ matrix given by $A = \Delta^T W^{-1} \Delta$. Here Δ is an $(M-2) \times M$ matrix of second differences with elements:

$$\Delta_{ii} = \frac{1}{h_i}, \quad \Delta_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}}, \quad \Delta_{i,i+2} = \frac{1}{h_{i+1}}.$$

And W is an $(M-2) \times (M-2)$ symmetric tri-diagonal matrix with elements:

$$W_{i-1,i} = W_{i,i-1} = \frac{h_i}{6}, \quad W_{i,i} = \frac{h_i + h_{i+1}}{3}, \text{ here } h_i = x_{i+1} - x_i.$$

As $\lambda \rightarrow 0$, the smoothing spline converges to the interpolating spline, and as $\lambda \rightarrow \infty$, it converges to the linear least squares estimate.

N.3 TRAJECTORY OF SPATIALLY ADAPTIVE SMOOTHING SPLINES

Let the curvature penalty $\rho(x) = \frac{1}{\eta t} \frac{1}{\zeta(x)} \frac{1}{M}$. Then (A155) can be written as

$$\min_{h \in W_2(S)} \frac{1}{M} \sum_{i=1}^M [h(x_i) - y_i]^2 + \int_S \rho(x) (h''(x))^2 dx, \quad (\text{A158})$$

where $W_2(S) = \{f: f, f' \text{ absolutely continuous and } f'' \in L^2(S)\}$, with $L^2(S)$ the square integrable functions over the domain S . Abramovich and Steinberg (1996); Pintore et al. (2006) give out the solution of (A158) 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 $\langle f, g \rangle_\rho = \int f''(x)g''(x)\rho(x) dx$. 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

$$K_\rho(x_1, x_2) = \int_S \rho(u)^{-1} [x_1 - u]_+ [x_2 - u]_+ du. \quad (\text{A159})$$

Then the minimizer \hat{h} of (A158) is given by

$$\hat{h}(x) = \sum_{j=1}^M c_j K_\rho(x_j, x) + a + bx. \quad (\text{A160})$$

Now define the $M \times M$ matrix

$$\Sigma_\rho = \{K_\rho(x_i, x_j)\}_{i,j=1,\dots,M}, \quad (\text{A161})$$

and the $M \times 2$ matrix

$$T = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_M \end{bmatrix}. \quad (\text{A162})$$

Denote the vector of coefficients $\mathbf{c} = (c_1, \dots, c_M)^T$ and the vector of output values $\mathbf{y} = (y_1, \dots, y_M)^T$. Then the coefficients in (A160) satisfy the following conditions:

$$\Sigma_\rho \left[(\Sigma_\rho + MI)\mathbf{c} + T \begin{pmatrix} a \\ b \end{pmatrix} \right] = \Sigma_\rho \mathbf{y} \quad \text{and} \quad T^T \left[\Sigma_\rho \mathbf{c} + T \begin{pmatrix} a \\ b \end{pmatrix} \right] = T^T \mathbf{y}. \quad (\text{A163})$$

After solving for (A163), we get the values of \mathbf{c} , a and b . Plug them into (A160), then we get the exact form of the minimizer of (A158).

O SOLUTION TO THE VARIATIONAL PROBLEMS AFTER TRAINING

O.1 INTERPOLATING SPLINES WITH UNIFORM CURVATURE PENALTY

Theorem 9(b) and (c) show that for certain distributions of $(\mathcal{W}, \mathcal{B})$, ζ is constant. In this case problem (19) is solved by the cubic spline interpolation of the data with natural boundary conditions (Ahlberg et al., 1967).

Theorem A13 (Ahlberg et al. 1967). *For training samples $\{(x_i, y_i)\}_{i=1}^M$, suppose $x_j \in S$, $j = 1, \dots, M$. Then cubic spline interpolation of data $\{(x_i, y_i)\}_{i=1}^M$ with natural boundary condition is the solution of*

$$\begin{aligned} & \min_{h \in C^2(S)} \int_S (h''(x))^2 dx \\ & \text{subject to } h(x_j) = y_j, \quad j = 1, \dots, m. \end{aligned} \quad (\text{A164})$$

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 out the form of the solution to the variational problem (19) by using the result in Appendix N. We add a coefficient λ before the regularization term in the optimization problem (A158) and choose $\rho(x) = \frac{1}{\zeta(x)}$. Then we get

$$\min_{h \in W_2(S)} \frac{1}{M} \sum_{i=1}^M [h(x_j) - y_j]^2 + \lambda \int_S \frac{1}{\zeta(x)} (h''(x))^2 dx. \quad (\text{A165})$$

As $\lambda \rightarrow 0$, the minimizer of (A165) converges to the solution of the following optimization problem:

$$\min_{h \in W^2(S)} \int_S \frac{(h''(x))^2}{\zeta(x)} dx \quad \text{s.t.} \quad h(x_j) = y_j, \quad j = 1, \dots, m, \quad (\text{A166})$$

which is exactly the variational problem (19). According to Appendix N, the solution of (A165) is given by:

$$\hat{h}^{(\lambda)}(x) = \sum_{j=1}^M c_j^{(\lambda)} K_{\frac{\lambda}{\zeta}}(x_j, x) + a^{(\lambda)} + b^{(\lambda)}x. \quad (\text{A167})$$

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

$$\Sigma_{\frac{\lambda}{\zeta}} \left[(\Sigma_{\frac{\lambda}{\zeta}} + MI) \mathbf{c}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = \Sigma_{\frac{\lambda}{\zeta}} \mathbf{y} \quad \text{and} \quad T^T \left[\Sigma_{\frac{\lambda}{\zeta}} \mathbf{c}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = T^T \mathbf{y}, \quad (\text{A168})$$

where $K_{\frac{\lambda}{\zeta}}$, $\Sigma_{\frac{\lambda}{\zeta}}$ and T are defined in (A159), (A161) and (A162). Since

$$\begin{aligned} K_{\frac{\lambda}{\zeta}}(x_1, x_2) &= \int_S \left(\frac{\lambda}{\zeta} \right)^{-1} [x_1 - u]_+ [x_2 - u]_+ du \\ &= \lambda^{-1} \int_S \left(\frac{1}{\zeta} \right)^{-1} [x_1 - u]_+ [x_2 - u]_+ du \\ &= \lambda^{-1} K_{\frac{1}{\zeta}}(x_1, x_2) \end{aligned} \quad (\text{A169})$$

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 (A167) and (A168) as

$$\hat{h}^{(\lambda)}(x) = \sum_{j=1}^M \bar{c}_j^{(\lambda)} K_{\frac{1}{\zeta}}(x_j, x) + a^{(\lambda)} + b^{(\lambda)}x, \quad (\text{A170})$$

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

$$\Sigma_{\frac{1}{\zeta}} \left[(\Sigma_{\frac{1}{\zeta}} + \lambda MI) \bar{\mathbf{c}}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = \Sigma_{\frac{1}{\zeta}} \mathbf{y} \quad \text{and} \quad T^T \left[\Sigma_{\frac{1}{\zeta}} \bar{\mathbf{c}}^{(\lambda)} + T \begin{pmatrix} a^{(\lambda)} \\ b^{(\lambda)} \end{pmatrix} \right] = T^T \mathbf{y}, \quad (\text{A171})$$

Now as $\lambda \rightarrow 0$, (A170) and (A171) become:

$$\hat{h}^{(0^+)}(x) = \sum_{j=1}^M \bar{c}_j^{(0^+)} K_{\frac{1}{\zeta}}(x_j, x) + a^{(0^+)} + b^{(0^+)}x, \quad (\text{A172})$$

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

$$\Sigma_{\frac{1}{\zeta}} \left[\Sigma_{\frac{1}{\zeta}} \bar{\mathbf{c}}^{(0^+)} + T \begin{pmatrix} a^{(0^+)} \\ b^{(0^+)} \end{pmatrix} \right] = \Sigma_{\frac{1}{\zeta}} \mathbf{y} \quad \text{and} \quad T^T \left[\Sigma_{\frac{1}{\zeta}} \bar{\mathbf{c}}^{(0^+)} + T \begin{pmatrix} a^{(0^+)} \\ b^{(0^+)} \end{pmatrix} \right] = T^T \mathbf{y}, \quad (\text{A173})$$

(A172) and (A173) give out the solution of (A165) as $\lambda \rightarrow 0$, which is also the solution to the variational problem (19).

P POSSIBLE GENERALIZATIONS

P.1 MULTI-DIMENSIONAL INPUTS

We have focused on 1D regression problems, but of course we are also interested in describing the implicit bias of gradient descent for multi-dimensional regression problems. Some of our results are independent of the input space dimension, and others can be generalized as we discuss in the following.

Consider a shallow neural network with d inputs. Use the same notation as in Section 5, and let $\mathbf{W}^{(1)}$ be the d -dimensional vector sampled from a d -dimensional random vector \mathbf{W} . Then optimization problem (15) becomes:

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^d \times \mathbb{R})} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha_n^2(\mathbf{W}^{(1)}, b) d\mu_n(\mathbf{W}^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha_n(\mathbf{W}^{(1)}, b) [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ d\mu_n(\mathbf{W}^{(1)}, b) = y_j, \quad j = 1, \dots, M, \end{aligned} \quad (\text{A174})$$

where $\mathbf{W}^{(1)}$ becomes a d -dimensional vector. The limit of the problem (A174) as width $n \rightarrow \infty$ is

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^d \times \mathbb{R})} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) d\mu(\mathbf{W}^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha(\mathbf{W}^{(1)}, b) [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ d\mu(\mathbf{W}^{(1)}, b) = y_j, \quad j = 1, \dots, M, \end{aligned} \quad (\text{A175})$$

Similar to Section 5, we can relax the optimization problem (A175) to

$$\begin{aligned} \min_{\substack{\alpha \in C(\mathbb{R}^d \times \mathbb{R}), \\ \mathbf{u} \in \mathbb{R}^{d+1}, v \in \mathbb{R}}} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) d\mu(\mathbf{W}^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^d \times \mathbb{R}} \alpha(\mathbf{W}^{(1)}, b) [\langle \mathbf{W}^{(1)}, \mathbf{x}_j \rangle + b]_+ d\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x}_j \rangle + v = y_j, \quad j = 1, \dots, M \end{aligned} \quad (\text{A176})$$

Let $g(\mathbf{x}, \alpha) = \int_{\mathbb{R}^d \times \mathbb{R}} \alpha(\mathbf{W}^{(1)}, b) [\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b]_+ d\mu(\mathbf{W}^{(1)}, b) + \langle \mathbf{u}, \mathbf{x} \rangle + v$ be the function represented by the infinite-width network. Then the Laplacian $\Delta g(\mathbf{x}, \alpha) = \sum_{i=1}^d \partial_{x_i}^2 g(\mathbf{x}, \alpha)$ is given by

$$\Delta g(\mathbf{x}, \alpha) = \int_{\mathbb{R}^{d+1}} \alpha(\mathbf{W}^{(1)}, b) \|\mathbf{W}^{(1)}\|_2^2 \delta(\langle \mathbf{W}^{(1)}, \mathbf{x} \rangle + b) d\mu(\mathbf{W}^{(1)}, b), \quad (\text{A177})$$

where δ denotes the Dirac delta function. To deal with the above integral, we change the variables in the following way. Let $\mathcal{U} = \|\mathbf{W}\|_2$, $\mathbf{V} = \mathbf{W}/\|\mathbf{W}\|_2$ and $\mathcal{C} = -\mathcal{B}/\|\mathbf{W}\|_2$. Let ν denote the distribution of $(\mathcal{U}, \mathbf{V}, \mathcal{C})$ and $\gamma(u, \mathbf{V}, c) = \alpha(u\mathbf{V}, -cu)$. Then (A177) becomes:

$$\begin{aligned} \Delta g(\mathbf{x}, \alpha) &= \int_{\mathbb{R} \times \mathbb{S}^{d-1} \times \mathbb{R}} \gamma(u, \mathbf{V}, c) \cdot u^2 \cdot \delta(u\langle \mathbf{V}, \mathbf{x} \rangle - cu) d\nu(u, \mathbf{V}, c) \\ &= \int_{\mathbb{R} \times \mathbb{S}^{d-1} \times \mathbb{R}} \gamma(u, \mathbf{V}, c) \cdot u \cdot \delta(\langle \mathbf{V}, \mathbf{x} \rangle - c) d\nu(u, \mathbf{V}, c) \\ &= \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \left(\int_{\mathbb{R}} \gamma(u, \mathbf{V}, c) \cdot u d\nu_{\mathcal{U}|\mathbf{V}=\mathbf{V}, \mathcal{C}=c}(u) \right) \delta(\langle \mathbf{V}, \mathbf{x} \rangle - c) d\nu_{\mathbf{V}, \mathcal{C}}(\mathbf{V}, c) \end{aligned} \quad (\text{A178})$$

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

$$\kappa(\mathbf{V}, c) = \int_{\mathbb{R}} \gamma(u, \mathbf{V}, c) \cdot u d\nu_{\mathcal{U}|\mathbf{V}=\mathbf{V}, \mathcal{C}=c}(u). \quad (\text{A179})$$

Then (A178) becomes:

$$\begin{aligned}\Delta g(\mathbf{x}, \alpha) &= \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \kappa(\mathbf{V}, c) \delta(\langle \mathbf{V}, \mathbf{x} \rangle - c) d\nu_{\mathbf{V}, c}(\mathbf{V}, c) \\ &= \int_{\mathbb{S}^{d-1}} \left(\int_{\mathbb{R}} \kappa(\mathbf{V}, c) \delta(\langle \mathbf{V}, \mathbf{x} \rangle - c) p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c) dc \right) d\nu_{\mathbf{V}}(\mathbf{V}) \\ &= \int_{\mathbb{S}^{d-1}} \kappa(\mathbf{V}, \langle \mathbf{V}, \mathbf{x} \rangle) p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(\langle \mathbf{V}, \mathbf{x} \rangle) d\nu_{\mathbf{V}}(\mathbf{V}),\end{aligned}\quad (\text{A180})$$

where $\nu_{\mathbf{V}}$ denote the distribution of \mathbf{V} . Assume that $\nu_{\mathbf{V}}$ has a density function $p_{\mathbf{V}}(\mathbf{V})$ on \mathbb{S}^{d-1} . Then (A180) becomes

$$\Delta g(\mathbf{x}, \alpha) = \int_{\mathbb{S}^{d-1}} \kappa(\mathbf{V}, \langle \mathbf{V}, \mathbf{x} \rangle) p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(\langle \mathbf{V}, \mathbf{x} \rangle) p_{\mathbf{V}}(\mathbf{V}) d\mathbf{V}, \quad (\text{A181})$$

Let $\beta(\mathbf{V}, c) = \kappa(\mathbf{V}, c) p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c) p_{\mathbf{V}}(\mathbf{V})$, then

$$\Delta g(\mathbf{x}, \alpha) = \int_{\mathbb{S}^{d-1}} \beta(\mathbf{V}, \langle \mathbf{V}, \mathbf{x} \rangle) d\mathbf{V}, \quad (\text{A182})$$

Actually the right-hand side of (A182) is precisely the dual Radon transform of β . According to (Ongie et al., 2020, Lemma 3),

$$\beta = -\frac{1}{2(2\pi)^{d-1}} \mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, \alpha)\}, \quad (\text{A183})$$

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

$$\mathcal{R}\{f\}(\omega, b) := \int_{\langle \omega, \mathbf{x} \rangle = b} f(\mathbf{x}) ds(\mathbf{x}), \quad (\omega, b) \in \mathbb{S}^{d-1} \times \mathbb{R}, \quad (\text{A184})$$

where $ds(\mathbf{x})$ represents integration with respect to $(d-1)$ -dimensional surface measure on the hyperplane $\langle \omega, \mathbf{x} \rangle = b$. The power of the negative Laplacian $(-\Delta)^{(d+1)/2}$ in (A183) is the operator defined in Fourier domain by

$$(-\Delta)^{(d+1)/2} f(\xi) = \|\xi\|^{d+1} \widehat{f}(\xi) \quad (\text{A185})$$

When $d+1$ is a even number, $(-\Delta)^{(d+1)/2}$ is the same as applying the negative Laplacian $(d+1)/2$ times, while if $d+1$ is odd it is a pseudo-differential operator given by convolution with a singular kernel.

Then according to (A183) and the definition of β , we have

$$\kappa(\mathbf{V}, c) = -\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2} g(\cdot, \alpha)\}(\mathbf{V}, c)}{2(2\pi)^{d-1} p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c) p_{\mathbf{V}}(\mathbf{V})} \quad (\text{A186})$$

According to the definition of κ (A179), we have

$$\begin{aligned}\kappa^2(\mathbf{V}, c) &= \left(\int_{\mathbb{R}} \gamma(u, \mathbf{V}, c) \cdot u d\nu_{\mathcal{U}|\mathbf{V}=\mathbf{V}, \mathcal{C}=c}(u) \right)^2 \\ &\leq \left(\int_{\mathbb{R}} \gamma^2(u, \mathbf{V}, c) d\nu_{\mathcal{U}|\mathbf{V}=\mathbf{V}, \mathcal{C}=c}(u) \right) \left(\int_{\mathbb{R}} u^2 d\nu_{\mathcal{U}|\mathbf{V}=\mathbf{V}, \mathcal{C}=c}(u) \right) \\ &= \left(\int_{\mathbb{R}} \gamma^2(u, \mathbf{V}, c) d\nu_{\mathcal{U}|\mathbf{V}=\mathbf{V}, \mathcal{C}=c}(u) \right) \mathbb{E}(\mathcal{U}^2 | \mathbf{V} = \mathbf{V}, \mathcal{C} = c)\end{aligned}\quad (\text{A187})$$

Given the function $\kappa(\mathbf{V}, c)$, the two sides of the above inequality are equal if

$$\gamma(u, \mathbf{V}, c) = \frac{\kappa(\mathbf{V}, c)u}{\mathbb{E}(\mathcal{U}^2 | \mathbf{V} = \mathbf{V}, \mathcal{C} = c)}. \quad (\text{A188})$$

Then the objective of (A176) has the following lower bound according to (A187):

$$\begin{aligned}
& \int_{\mathbb{R}^d \times \mathbb{R}} \alpha^2(\mathbf{W}^{(1)}, b) d\mu(\mathbf{W}^{(1)}, b) \\
&= \int_{\mathbb{R} \times \mathbb{S}^{d-1} \times \mathbb{R}} \gamma^2(u, \mathbf{V}, c) d\nu(u, \mathbf{V}, c) \\
&= \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \left(\int_{\mathbb{R}} \gamma^2(u, \mathbf{V}, c) d\nu_{\mathcal{U}|\mathbf{V}=\mathbf{V}, \mathcal{C}=c}(u) \right) d\nu_{\mathbf{V}, \mathcal{C}}(\mathbf{V}, c) \\
&\geq \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \frac{\kappa^2(\mathbf{V}, c)}{\mathbb{E}(\mathcal{U}^2|\mathbf{V}=\mathbf{V}, \mathcal{C}=c)} d\nu_{\mathbf{V}, \mathcal{C}}(\mathbf{V}, c) \tag{A189} \\
&= \int_{\mathbb{R}^d \times \mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot, \alpha)\}(\mathbf{V}, c)}{2(2\pi)^{d-1}p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c)p_{\mathbf{V}}(\mathbf{V})} \right)^2 \frac{1}{\mathbb{E}(\mathcal{U}^2|\mathbf{V}=\mathbf{V}, \mathcal{C}=c)} d\nu_{\mathbf{V}, \mathcal{C}}(\mathbf{V}, c) \\
&= \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \left(\frac{\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot, \alpha)\}(\mathbf{V}, c)}{2(2\pi)^{d-1}p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c)p_{\mathbf{V}}(\mathbf{V})} \right)^2 \frac{p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c)p_{\mathbf{V}}(\mathbf{V})}{\mathbb{E}(\mathcal{U}^2|\mathbf{V}=\mathbf{V}, \mathcal{C}=c)} d\mathbf{V}dc \\
&= \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \frac{(\mathcal{R}\{(-\Delta)^{(d+1)/2}g(\cdot, \alpha)\}(\mathbf{V}, c))^2}{4(2\pi)^{2(d-1)}p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c)p_{\mathbf{V}}(\mathbf{V})\mathbb{E}(\mathcal{U}^2|\mathbf{V}=\mathbf{V}, \mathcal{C}=c)} d\mathbf{V}dc
\end{aligned}$$

The two sides of the above inequality are equal if (A188) is satisfied. Then similar to Theorem 6, we show that under the assumptions that: (1) $\nu_{\mathbf{V}}$ has a density function $p_{\mathbf{V}}(\mathbf{V})$ on \mathbb{S}^{d-1} ; (2) $\nu_{\mathcal{C}|\mathbf{V}=\mathbf{V}}$ has a density function $p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c)$, the solution of (A176) in function space actually solves the following optimization problem:

$$\begin{aligned}
& \min_{g \in C(\mathbb{R}^d)} \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \frac{(\mathcal{R}\{(-\Delta)^{(d+1)/2}g\}(\mathbf{V}, c))^2}{4(2\pi)^{2(d-1)}p_{\mathcal{C}|\mathbf{V}=\mathbf{V}}(c)p_{\mathbf{V}}(\mathbf{V})\mathbb{E}(\mathcal{U}^2|\mathbf{V}=\mathbf{V}, \mathcal{C}=c)} d\mathbf{V}dc \tag{A190} \\
& \text{subject to } g(\mathbf{x}_j) = y_j, \quad j = 1, \dots, M,
\end{aligned}$$

The optimization problem (A190) characterizes the implicit bias of the gradient descent in function space for multi-dimensional setting. The details of Radon transform and how to make it well-defined are shown in the work of Ongie et al. (2020). We omit the details here.

Zhang et al. (2019) obtained a characterization in terms of the minimization of a kernel norm in function space. This result is also valid for multidimensional inputs. In Appendix M we proved the equivalence between kernel norm minimization and our results in the one-dimensional setting. In the multi-dimensional setting, it will be interesting to show that the kernel norm is equivalent to the objective in (A190) under some conditions.

Lastly, in the multi-dimensional setting the breakpoint density in one-dimensional setting is replaced by a density of the locus of non-linearity of the represented functions, which has been studied by Hanin and Rolnick (2019).

P.2 OTHER ACTIVATION FUNCTIONS

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 $L\sigma = \delta$, 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"¹ of two-layer neural networks with activation function σ is actually minimizing 1-norm of Lf 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

¹Here the form of "norm" depends on the degree of homogeneity of the activation σ . We use quotation marks here because the weight "norm" is a generalized notion of norm. It may not satisfy the property of norm.

get an analogous result. 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 (15) becomes:

$$\begin{aligned} \min_{\alpha_n \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha_n^2(W^{(1)}, b) \, d\mu_n(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha_n(W^{(1)}, b) \sigma(W^{(1)}x_j + b) \, d\mu_n(W^{(1)}, b) = y_j, \quad j = 1, \dots, M. \end{aligned} \quad (\text{A191})$$

The limit of the problem (A191) as width $n \rightarrow \infty$ is

$$\begin{aligned} \min_{\alpha \in C(\mathbb{R}^2)} \quad & \int_{\mathbb{R}^2} \alpha^2(W^{(1)}, b) \, d\mu(W^{(1)}, b) \\ \text{subject to} \quad & \int_{\mathbb{R}^2} \alpha(W^{(1)}, b) \sigma(W^{(1)}x_j + b) \, d\mu(W^{(1)}, b) = y_j, \quad j = 1, \dots, M. \end{aligned} \quad (\text{A192})$$

As in Section 5.2, we can change the variables and relax the optimization problem (A192) to

$$\begin{aligned} \min_{\substack{\gamma \in C(\mathbb{R}^2), \\ p \in C(\mathbb{R})}} \quad & \int_{\mathbb{R}^2} \gamma^2(W^{(1)}, c) \, d\nu(W^{(1)}, c) \\ \text{subject to} \quad & p(x_j) + \int_{\mathbb{R}^2} \gamma(W^{(1)}, c) \sigma(W^{(1)}(x_j - c)) \, d\nu(W^{(1)}, c) = y_j, \quad j = 1, \dots, M \\ & L p \equiv 0. \end{aligned} \quad (\text{A193})$$

If the activation function σ is ReLU, p is a linear function. Then (A193) becomes the optimization problem (17). Define the output function g of the neural network by

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

Assume that the activation function σ is homogeneous of degree k , i.e. $\sigma(ax) = a^k \sigma(x)$ for all $a > 0$. Similar to (A84), we have

$$\begin{aligned} (Lg)(x, (\gamma, p)) &= L \left(\int_{\mathbb{R}^2} \gamma(W^{(1)}, c) |W^{(1)}|^k \sigma(\text{sign}(W^{(1)}) \cdot (x - c)) \, d\nu(W^{(1)}, c) \right) \\ &= \int_{\mathbb{R}^2} \gamma(W^{(1)}, c) |W^{(1)}|^k \delta(x - c) \, d\nu(W^{(1)}, c) \\ &= \int_{\text{supp}(\nu_c)} \left(\int_{\mathbb{R}} \gamma(W^{(1)}, c) |W^{(1)}|^k \, d\nu_{\mathcal{W}|C=c}(W^{(1)}) \right) \delta(x - c) \, d\nu_c(c) \quad (\text{A194}) \\ &= \int_{\text{supp}(\nu_c)} \left(\int_{\mathbb{R}} \gamma(W^{(1)}, c) |W^{(1)}|^k \, d\nu_{\mathcal{W}|C=c}(W^{(1)}) \right) \delta(x - c) p_c(c) \, dc \\ &= p_c(x) \int_{\mathbb{R}} \gamma(W^{(1)}, x) |W^{(1)}|^k \, d\nu_{\mathcal{W}|C=x}(W^{(1)}). \end{aligned}$$

Then similar to Theorem 6, we show that the solution of (A193) in function space actually solves the following optimization problem:

$$\min_{h \in C^2(S)} \int_S \frac{((Lh)(x))^2}{\zeta(x)} \, dx \quad \text{s.t.} \quad h(x_j) = y_j, \quad j = 1, \dots, m, \quad (\text{A195})$$

where $\zeta(x) = p_c(x) \mathbb{E}(\mathcal{W}^{2k} | C = x)$ and $S = \text{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$.

P.3 DEEP NETWORKS AND OTHER ARCHITECTURES

For deep networks of L layers, if we only train the output layer, then we actually train a linear model. We can construct a two-layer neural network with activation σ which is a $(L - 1)$ -layer neural network. Then training this two-layer network is equivalent to training only the output layer of

the deep network. So we can use the arguments in Section P.2 about the other activation functions. However, it remains unclear how we can find out the operator L corresponding to this activation σ .

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) show that the model still is approximated by a linearized model. The result on kernel norm minimization (Zhang et al., 2019) 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.4 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). 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 4, 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 result will generalize. 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. Theorem 5 and Theorem 6 are about a variational problem that is derived from Theorem A1, in relation to the minimization of $\|\theta - \theta_2\|_2$. Theorem A1 remains valid for other loss functions beside MSE. To sum up, if we can generalize the Theorem 4 and the result of Lee et al. (2018) to other loss functions, we can generalize our main result in Theorem 1 to other loss functions.

P.5 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. (2018) 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.

The implicit bias of SGD has been studied in a series of articles, taking a different perspective to the implicit bias of gradient descent. This has been linked to the shape of the optimization landscape, with smaller mini-batch or larger step size leading to a bias towards wider minima of the objective function (Keskar et al., 2017; Wu et al., 2017; Dinh et al., 2017). Further, this is related to stability and robustness. Here again, it will be interesting to take an NTK perspective to analyze SGD (Allen-Zhu et al., 2019) and explore kernel norm minimization and explicit forms of the regularization in function space under SGD.

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. ISSN 0378-3758. doi: [https://doi.org/10.1016/0378-3758\(95\)00021-6](https://doi.org/10.1016/0378-3758(95)00021-6). 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. ISBN 9780080955452. URL <https://books.google.com/books?id=S7dlpjJHsRgC>.
- 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-zhul9a.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/>.
- Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.
- 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 Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.
- 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 2018. PMLR. URL <http://proceedings.mlr.press/v80/gunasekar18a.html>.
- Boris Hanin and David Rolnick. Complexity of linear regions in deep 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 2596–2604, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL <http://proceedings.mlr.press/v97/hanin19a.html>.
- 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.
- 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>.
- 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.
- C. Nasim. The solution of an integral equation. *Proceedings of the American Mathematical Society*, 40(1):95–101, 1973. ISSN 00029939, 10886826. 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.

- 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.
- Alexandre Pintore, Paul Speckman, and Chris C. Holmes. Spatially adaptive smoothing splines. *Biometrika*, 93(1):113–125, 03 2006. ISSN 0006-3444. doi: 10.1093/biomet/93.1.113. URL <https://doi.org/10.1093/biomet/93.1.113>.
- 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, 2020. URL <https://openreview.net/forum?id=BJl9PRVKDS>.
- 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>.
- 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.
- Roman Vershynin. Four lectures on probabilistic methods for data science. In M.W. Mahoney, J.C. Duchi, and A.C. Gilbert, editors, *The Mathematics of Data*, IAS/Park City Mathematics Series, pages 231–271. American Mathematical Society, 2018. ISBN 9781470435752. URL <https://books.google.de/books?id=7HJ6DwAAQBAJ>.
- Lei Wu, Zhanxing Zhu, and E Weinan. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.
- Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. A type of generalization error induced by initialization in deep neural networks. *arXiv preprint arXiv:1905.07777*, 2019.