Ridge Regression with Over-Parametrized Two-Layer Networks Converge to Ridgelet Spectrum

Sho Sonoda
RIKEN AIP

Isao Ishikawa
Ehime University & RIKEN AIP

Masahiro Ikeda
RIKEN AIP

Abstract

Characterization of local minima draws much attention in theoretical studies of deep learning. In this study, we investigate the distribution of parameters in an over-parametrized finite neural network trained by ridge regularized empirical square risk minimization (RERM). We develop a new theory of ridgelet transform, a wavelet-like integral transform that provides a powerful and general framework for the theoretical study of neural networks involving not only the ReLU but general activation functions. We show that the distribution of the parameters converges to a spectrum of the ridgelet transform. This result provides a new insight into the characterization of the local minima of neural networks, and the theoretical background of an inductive bias theory based on lazy regimes. We confirm the visual resemblance between the parameter distribution trained by SGD, and the ridgelet spectrum calculated by numerical integration through numerical experiments with finite models.

1 INTRODUCTION

Characterizing local minima is important in theoretical studies of neural networks. Despite the high-dimensionality of parameters, neural networks have become state-of-the-art in many application areas since the emergence of AlexNet (Krizhevsky et al., 2012). This has been a mystery of machine learning theory because several VC-based arguments have shown that the generalization error is upper bounded by the dimension of parameters, or the capacity of the hypothesis class (Neyshabur et al., 2015; Bartlett et al., 2017), but as Arora et al. (2018) pointed out, these bounds are not tight in practice. As Zhang et al. (2017) suggested, many researchers now consider that the typical solutions obtained via deep learning are concentrated in a much smaller class than expected from the algebraic dimension of parameters or any other data-independent capacities.

However, characterizing local minima is a challenging problem due to the nonlinearity of parameters and the non-convexity of learning problems. To tackle this problem, the *over-parametrization* is considered to be one of the promising assumption for theoretical analysis of neural networks, which assumes that the number of parameters in neural networks is sufficiently larger than the sample size. This assumption has revolutionized our understanding of the local minima. For example, the global convergence of deep learning is now proved in many ways, and some researchers further conjecture that the typical solutions are close to the initial parameters (see Section 5 for more details).

In this study, we provide an explicit expression for the *global minimizer* in the over-parametrized regime by means of the integral representation (Barron, 1993; Murata, 1996; Sonoda and Murata, 2017). The integral representation is an effective machinery to analyze the neural networks using harmonic analysis, a branch of mathematics. It is realized as a linear operator between function spaces (see Definition 2.1), and provides a principled approach to study over-parametrized neural networks with not only ReLU but also a wide range of activation functions. Recently, this has been recognized as an effective reparametrization in the *mean-field theory* (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018), which employs the integral representation to show the global convergence for finite two-layer networks.

To be precise, we develop a new theory of the *rigelet transform on the torus*, and prove for the first time that the parameter distributions of *finite two-layer neural networks trained by regularized empirical risk minimization (RERM)* converges to a ridgelet spec-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

trum as both the parameter number and sample size tend to infinity. By virtue of the over-parametrization, our theorem holds not only for strict global minima but also other suboptimal minima such as random features solutions. The ridgelet transform, which is a wavelet-like integral transform, is originally developed by Murata (1996), Candès (1998) and Rubin (1998), and has a remarkable application to analysis of neural networks (see eg., Starck et al., 2010 and Appendix A.5).

Numerical simulation confirms our main theoretical results. Namely, the scatter plot of parameter distributions learned by stochastic gradient descent (SGD) shows a similar pattern to the ridgelet spectrum. While our theory do not assume any specific training algorithm (but ERM), the empirical results further suggests that our theoretical findings hold for a more realistic settings.

To the present date, mean-field theories have not provided the explicit expression like ridgelet transform because they consider the integral- representation without ridgelet transform. If we know that the local minima tends to a ridgelet spectrum, then we can further understand the theoretical backgrounds behind the *lazy learning*, a recent trend of inductive bias theories, such as the *neural tangent kernel* (Jacot et al., 2018; Lee et al., 2019) and the *strong lottery ticket hypothesis* (Frankle and Carbin, 2019), claiming that the learned parameters are very close to the initial parameters. This is reasonable when the initial parameters cover the support of the ridgelet spectrum. As a consequence, this study develops a new direction of the theoretical studies of local minima. See Related Works (Section 5) for more discussions.

Contributions are summarized as follows: This study

- develops a complete set of the ridgelet transform on the torus including reconstruction formula, admissible condition, Plancherel formula, boundedness, density, and several formulas for calculus;
- mathematically proves (1) that the population risk minimizer of the ridge regression problem with integral representation NNs is expressed by the ridgelet transform, and (2) that the empirical risk minimizer (ERMer) of the ridge regression problem with finite two-layer NNs converges to the ridgelet transform in the over-parametrized regime, namely, when the parameter number and the sample size tend to infinity;
- empirically confirms that the parameter distributions in finite two-layer NNs trained by stochastic gradient descent (SGD) visually converge to the ridgelet spectrum obtained by numerical integration; and
- develops a new direction of the theoretical studies of local minima that would reinforce a wide range of recent global convergence theories including mean-field theories and lazy learning.

The structure of this paper is as follows: In Section 2, we develop the theory of the ridgelet transform on the torus. In Section 3, we give our main results. In Section 4, we conduct numerical simulation. In Sections 5, we discuss the relation to previous studies. In Section 6, we provides conclusions and further discussions.

Notations. The m is the dimension of the Euclidean space of the input data. We denote by $\mathrm{d}\boldsymbol{x}$ the Lebesgue measure on $\mathbb{R}^m$.

We denote by $\mathbb{T}$ the torus $\mathbb{R}/T\mathbb{Z}$ for a fixed $T > 0$, which is identified with the interval $[-T/2, T/2)$. We denote by $\mathrm{d}b$ the invariant measure on $\mathbb{T}$, that is identical with the Lebesgue measure on $[-T/2, T/2)$ via the above identification.

For $A > 0$, we denote by $\mathbb{I}_A$ the interval $[-A, A]$. We denote by $\mathrm{d}\boldsymbol{a}$ the Lebesgue masure on $\mathbb{I}_A^m$. We define $\mu_A := \mathrm{d}\boldsymbol{a}\mathrm{d}b$ a measure on $\mathbb{I}_A^m \times \mathbb{T}$.

For a measurable space X equipped with a measure μ, we denote by $L^p(X, \mu)$ the space of L^p integrable functions on X with respect to μ. For simplicity, we write $L^p(\mu)$ if X is obvious in context, or write $L^p(X)$ when μ is the Lebesgue measure or the invariant measure on $\mathbb{T}$.

For a topological space X, we denote by $C_b(X)$ the Banach space of bounded continuous functions on X equipped with the uniform norm.

For a periodic function $\sigma : \mathbb{T} \to \mathbb{R}$ and an integer n, we write the Fourier coefficient as $\widehat{\sigma}(n) := (1/T)\int_{-T/2}^{T/2} \sigma(t) e^{2\pi i n t/T} \mathrm{d}t$.

For a function $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma_{\boldsymbol{a},b}$ denotes a function $\boldsymbol{x} \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$, and $\sigma_{\boldsymbol{x}}$ denotes a function $(\boldsymbol{a}, b) \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$.

2 RIDGELET TRANSFORM ON THE TORUS

In this section, we establish the theory of the ridgelet transform on the torus, which is a basis of this study. For those who are not familiar with ridgelet analysis, we refer to the *Cheat Sheet* (Appendix A) including the list of handy formulas and the visualization of reconstruction formula with admissible and non-admissible functions.

The ridgelet transform on the torus is a complete set of new ridgelet transform, because periodic activation

functions cannot be *self-admissible* in the non-periodic context, and thus two theories are exclusive to each other. We need the self-admissibility for the Plancherel formula to hold.

2.1 Periodic Activation Function

In this study, we consider the activation function σ to be bounded and measurable function from $\mathbb{T}$ to $\mathbb{R}$, or equivalently, a bounded measurable periodic function σ on $\mathbb{R}$ with period T: $\sigma(t+T)=\sigma(t)$.

Originally, the ridgelet transform is defined on the real line $\mathbb{R}$ (Murata, 1996; Candès, 1998). However, the non-compactness of $\mathbb{R}$ gives rise to several technical difficulties in the proofs, especially, in establishing a connection between the ridgelet transform and finite neural networks. Moreover, the original definition excludes non-integrable activation functions such as the hyperbolic tangent function and the rectified linear unit (ReLU). Sonoda and Murata (2017) have extended the ridgelet transform to accept such non-integrable activation functions, by introducing an auxiliary dual activation function. However, their extension sacrifices the Plancherel formula, which we need in this study.

Although it might be possible to develop a truncated version of the ridgelet theory, such as the "ridgelet transform on a closed interval", it disables us from using fruitful results in Fourier analysis. In contrast, if we impose a periodicity on σ, we can use a quite powerful mathematical machinery, that is the theory of the Fourier transform on the torus $\mathbb{T}\simeq[-T/2,T/2)$. Since we may take arbitrarily large T, it is not so harmful as we often consider a finite dataset that is always contained in a (sufficiently large) compact domain. It is worth remarking that there exists a study (Sitzmann et al., 2020) that utilizes a periodicity of the activations, in which the authors report neural networks with periodic activations perform better in some machine learning tasks using real world data.

2.2 Integral Representation of Neural Networks

We give a definition of an integral representation.

Definition 2.1 (Integral Representation)**.** *Let $\sigma:\mathbb{T}\to\mathbb{R}$ be a bounded measurable function, and let P be a finite Borel measure on $\mathbb{R}^m$. For any finite Borel measure λ on $\mathbb{R}^m\times\mathbb{T}$, we define an integral representation of a neural network $S_\lambda:L^2(\lambda)\to L^2(P)$ by*

$$S_\lambda[\gamma](\boldsymbol{x}):=\int_{\mathbb{R}^m\times\mathbb{T}}\gamma(\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)\mathrm{d}\lambda(\boldsymbol{a},b). \quad (1)$$

In this study, we mainly consider two cases: $\lambda_d=\sum_{i=1}^d\delta_{\boldsymbol{a}_i,b_i}$, and $\lambda=\mu_A$. As for the first case, $S_{\lambda_d}[\gamma](\boldsymbol{x})=\sum_{i=1}^d\gamma(\boldsymbol{a}_i,b_i)\sigma(\boldsymbol{a}_i\cdot\boldsymbol{x}-b_i)$. Thus S_{λ_d} represents a finite two-layer neural network. As for the second case, the operator S_{μ_A} can be regarded as a continuum limit of neural networks whose hidden parameters $(\boldsymbol{a}_i,b_i)$ are contained in $\mathbb{I}_A^m\times\mathbb{T}$.

Here, we provide a remark on the space $L^2(P)$. As $L^2(\mathbb{R}^m)$ does not contain $\sigma_{\boldsymbol{a},b}$ and thus any finite neural networks, we cannot see the direct connection between finite neural networks and integral representations of neural networks in $L^2(\mathbb{R}^m)$. To circumvent this technical issue, we consider $L^2(P)$ since $\sigma_{\boldsymbol{a},b}\in L^2(P)$.

We note the boundedness of the integral representation:

Proposition 2.2. *The linear operator S_λ is bounded, namely, there exists a positive constant $C>0$ such that $\|S_\lambda[\gamma]\|_{L^2(P)}\leq C\|\gamma\|_{L^2(\lambda)}$ for all $\gamma\in L^2(\lambda)$.*

The boundedness is a sufficient condition to establish the unique existence of the global optimum in the learning problem (7) we will consider.

2.3 Ridgelet Transform

Let us introduce an assumption on the bounded measurable function σ.

Assumption 2.3 (Admissible Condition)**.** *The function $\sigma:\mathbb{T}\to\mathbb{R}$ is bounded and measurable, and satisfies the following two conditions: (1) $\widehat{\sigma}(0)=0$, and (2) $T^{m+1}\sum_{n\neq0}|\widehat{\sigma}(n)|^2/|n|^m=1$.*

We need the admissibility condition (AC) in the proof of the reconstruction formula (3) below. It is not at all strong. In fact, the infinite sum in the second condition always converge because σ is square integrable, thus, we may replace σ with a function satisfying these condition via only multiplying and subtracting constants. For example, restrictions of ReLU and hyperbolic tangent to $\mathbb{T}$ with slight modifications on the constants, namely, $\mathrm{ReLU}|_{[-T/2,T/2]}-T/8$ and $\tanh|_{[-T/2,T/2]}$ are admissible. Note that we can further eliminate the constant $-T/8$ in the ReLU by simply adding an offset b_0 to the model as $S[\gamma]+b_0$. It is a routine to extend our analysis for $S[\gamma]+b_0$ to have a parallel consequence. In this case, we do not need Item 1 of the AC.

We introduce the *ridgelet transform* and its reconstruction formula.

Definition 2.4 (Ridgelet Transform)**.** *Impose Assumption 2.3 on σ. Then, we define the ridgelet trans-*

form $R : L^2(\mathbb{R}^m) \to L^2(\mathbb{R}^m \times \mathbb{T})$ *by*

$$R[f](\boldsymbol{a}, b) := \int_{\mathbb{R}^m} f(\boldsymbol{x})\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\boldsymbol{x}. \tag{2}$$

For a rigorous treatment of the well-definedness of the ridgelet transform, see Remark 2.7 below.

Theorem 2.5 (Reconstruction Formula)**.** *Impose Assumption 2.3 on* σ*. Then for* $f, g \in L^2(\mathbb{R}^m)$*, we have*

$$\lim_{A\to\infty} S_{\mu_A}\left[R[f]\right] = f, \tag{3}$$

$$\left\langle R[f], R[g]\right\rangle_{L^2(\mathbb{R}^m\times\mathbb{T})} = \langle f, g\rangle_{L^2(\mathbb{R}^m)}. \tag{4}$$

By discretizing the integral in (3), we have a stronger result of a well-known universality of two-layer neural networks as a corollary of Theorem 2.5:

Corollary 2.6. *Impose Assumption 2.3 on* σ*, and assume* f *is a rapidly decreasing smooth function. Then, for an arbitrary* $\varepsilon > 0$ *and a compact domain* $K \subset \mathbb{R}^m$*, there exists* $A > 0$ *and* $d > 0$ *such that the following inequality almost surely holds:*

$$\left\| \frac{(2A)^m T}{d} \sum_{i=1}^{d} R[f](\boldsymbol{a}_i, b_i)\sigma_{\boldsymbol{a}_i, b_i} - f \right\|_{L^\infty(K)} < \varepsilon,$$

where $(\boldsymbol{a}_i, b_i)$*'s are i.i.d samples drawn from the uniform distribution over* $\mathbb{I}_A^m \times \mathbb{T}$*.*

Typical universality results only concern approximation power of neural networks. Such results guarantee the representation power of neural networks, however, their parameters could become too large to be realized in the real world. In contrast, Corollary 2.6 provides us not only the approximation power but also detailed information of the parameter distributions. Although there might be many candidates of neural networks that represent the target function, Corollary 2.6 shows that one of them are given by the ridgelet transform, a simple integral transform. Conversely, under over-parametrized condition, we will prove that the parameter distribution of an optimal neural network is closely related to the ridgelet transform.

Remark 2.7. For mathematical and logical accuracy, we need to define $R[f]$ for all the $f \in L^2(\mathbb{R}^m)$ with Theorem 2.5 via bounded extension, essentially the same arguments in the definition of the L^2-Fourier transform on the Eulidean space. More precisely, We first define $R[f]$ for $f \in L^1(\mathbb{R}^m)$, which is absolutely convergent because $\sigma \in L^\infty(\mathbb{T})$. Then, we show the Plancherel formula for $f \in L^1(\mathbb{R}^m) \cap L^2(\mathbb{R}^m)$ as in Theorem 2.5. Finally, we extend $R[f]$ for $f \in L^2(\mathbb{R}^m)$ as a common limit of $R[f_i]$, where f_i is any sequence in $L^1(\mathbb{R}^m) \cap L^2(\mathbb{R}^m)$ that converges to f in $L^2(\mathbb{R}^m)$.

3 MAIN RESULTS

In this section, we describe the formulation of our problem and main results (Theorems 3.3 and 3.4). We fix an activation function $\sigma : \mathbb{T} \to \mathbb{R}$. We assume that σ is continuous almost everywhere, equivalently, Riemann integrable, and satisfies Assumption 2.3. We also fix a square integrable function $f \in L^2(\mathbb{R}^m)$ as a data generating function, and an absolutely continuous probability measure P on $\mathbb{R}^m$ with bounded density function $p \in L^1(\mathbb{R}^m)$ as the input data distribution. We write an empirical measure corresponding to P by $P_N := 1/N \sum_{i=1}^N \delta_{\boldsymbol{x}_i}$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^m$ are i.i.d samples drawn from P. For $A, T > 0$ and $d \in \mathbb{N}$, let

$$\Lambda_{d,A} := \left\{ \frac{C_0}{d} \sum_{i=1}^{d} \delta_{(\boldsymbol{a}_i, b_i)} \mid (\boldsymbol{a}_i, b_i) \in \mathbb{I}_A^m \times \mathbb{T} \right\}, \tag{5}$$

where $C_0 := (2A)^m T$, be the collection of d-term *hidden parameter distributions* on $\mathbb{I}_A^m \times \mathbb{T}$.

Main Claim (Theorems 3.3 and 3.4) Our main results are summarized in the following formula, which is a converse of Corollary 2.6:

$$\lim_{N\to\infty} \lim_{d\to\infty} \gamma^*_{N,d} = R\left[\frac{pf}{\beta + p}\right] + \Delta_A,$$

where $\gamma^*_{N,d}$ represents the parameter distribution of d-term two-layer neural networks trained by regularized empirical risk minimization with N training examples, β is a regularization parameter, and Δ_A is a small residual term that tends to 0 as $A \to \infty$.

We call this inner limit $\lim_{d\to\infty} \gamma^*_{N,d}$ along the parameter number d the *over-parametrization*. In this sense, we show "over-parametrized networks converge to the ridgelet transform."

3.1 Square Loss Minimization

For an arbitrary $\beta > 0$, any finite Borel measures λ and P on $\mathbb{R}^m \times \mathbb{T}$ and $\mathbb{R}^m$, respectively, and any square integrable function $f \in L^2(P)$, we consider the following form of L^2-regularized square risk:

$$J(\gamma; f, P, \lambda, \beta) := \left\| f - S_\lambda[\gamma] \right\|^2_{L^2(P)} + \beta \left\| \gamma \right\|^2_{L^2(\lambda)}. \tag{6}$$

We denote by $\gamma^*[f; P, \lambda, \beta]$, the unique element that attains the minimum

$$\min_{\gamma \in L^2(\lambda)} J(\gamma; f, P, \lambda, \beta), \tag{7}$$

which always exists as long as S_λ is a densely defined closed operator (see Appendix B for the proof). As we

have already seen in Proposition 2.2, S_λ is bounded, and thus the minimizer always exists.

The minimizer of (7) behaves well under limit manipulations, namely, we have the following lemma:

Lemma 3.1. *For any finite Borel measure λ on $\mathbb{R}^m \times \mathbb{T}$, as $N \to \infty$, we have*

$$\|\gamma^*[f; P_N, \lambda, \beta] - \gamma^*[f; P, \lambda, \beta]\|_{L^2(\lambda)} \to 0 \quad P\text{-a.s.}$$

Now we consider two types of minimization problems. Our goal is to describe the relationship between the minimizers as well as investigate the properties of them.

Continuous Population Risk Minimizer γ^*. We denote by γ^* the population risk minimizer of

$$\min_{\gamma \in L^2(\mu_A)} J(\gamma; f, P, \mu_A, \beta). \tag{8}$$

The minimizer γ^* is equal to $\gamma^*[f; P, \mu_A, \beta]$, and referenced as a theoretically ideal object, which shows up as the global minimizer with over-parametrized neural networks.

Finite Empirical Risk Minimizer $\gamma^*_{N,d}$. We denote by $\gamma^*_{N,d}$ an empirical risk minimizer of

$$\min_{\lambda \in \Lambda_{d,A}} \min_{\gamma \in L^2(\lambda)} J(\gamma; f, P_N, \lambda, \beta_d). \tag{9}$$

By definition, the minimization problem (9) is equivalent to an ordinary learning problem of two-layer neural networks in term of the following empirical risk with respect to the parameters $(\boldsymbol{a}_j, b_j, c_j) \in \mathbb{I}_A^m \times \mathbb{T} \times \mathbb{R}$:

$$\frac{1}{N}\sum_{i=1}^N \left| f(\boldsymbol{x}_i) - \frac{C_0}{d}\sum_{j=1}^d c_j \sigma_{\boldsymbol{a}_j, b_j}(\boldsymbol{x}_i) \right|^2 + \beta_d \frac{C_0}{d}\sum_{j=1}^d |c_j|^2, \tag{10}$$

where we write $C_0 := (2A)^m T$. By definition, $\gamma_{N,d}$ attains the minumum of (7) for some $\lambda_d^* \in \Lambda_{d,A}$, namely, we have $\gamma^*_{N,d} = \gamma^*[f; P_N, \lambda_d^*, \beta_d]$. We call λ the *hidden parameter distribution* of the ERMer $\gamma^*_{N,d}$.

As we see soon later, our main theorem holds not only for the strict global minimizer but also for more general solutions that satisfy a very mild assumption:

Assumption 3.2. *A sequence of hidden parameter distributions $\{\lambda_d\}_{d=1}^\infty$ ($\lambda_d \in \Lambda_{d,A}$) weakly converges to the uniform distribution μ_A over the parameter domain $\mathbb{I}_A^m \times \mathbb{T}$, namely, for any bounded continuous function $h \in C_b(\mathbb{I}_A^m \times \mathbb{T})$, $\int h \mathrm{d}\lambda_d \to \int h \mathrm{d}\boldsymbol{a}\mathrm{d}b$ as $d \to \infty$.*

Here, we remark for potential confusions: The objective function (7) may remind some readers of the kernel ridge regression (KRR) with either $k(\boldsymbol{x}, \boldsymbol{y}) := \mathbb{E}_{\boldsymbol{a},b\sim\lambda}[\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\sigma(\boldsymbol{a}'\cdot\boldsymbol{x} - b')]$ on the data space, or $K((\boldsymbol{a}, b), (\boldsymbol{a}', b')) := \mathbb{E}_{\boldsymbol{x}\sim P}[\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\sigma(\boldsymbol{a}'\cdot\boldsymbol{x} - b')]$ on the parameter space. However, both KRRs cannot deal with our problem (7). Recall that our final goals are to specify the parameter distribution $\gamma^* \in L^2(\mu_A)$ and to show the convergence of the finite minimizers $\gamma_d = \sum_{i=1}^d c_i \delta_{(a_i, b_i)}$ to γ^*. In general, γ^* involves null component when $\beta > 0$, but H_K does not involve null components and thus the minimizer γ^* in $L^2(\mu_A)$ cannot always included in H_K.

3.2 Explicit Representation of Continuous Minimizer

The first main result is the explicit representation of the continuous minimizer γ^*, the solution of (8), in terms of the ridgelet transform.

Theorem 3.3. *Let $f \in L^2(\mathbb{R}^m)$ be a bounded square integrable function, and let P be an absolutely continuous probability measure on $\mathbb{R}^m$ with bounded density function $p \in L^1(\mathbb{R}^m)$. For $A > 0$ and $\beta > 0$, we have*

$$\gamma^* = R\left[\frac{pf}{\beta + p}\right] + \Delta_A, \tag{11}$$

where Δ_A is an element of $L^2(\mu_A)$ such that

$$\lim_{A\to\infty} \|\Delta_A\|_{L^2(\mu_A)} = 0. \tag{12}$$

By Corollary 2.6, it is reasonable to expect the minimizer γ^* and the ridgelet transform are intimately related to each other. However, since there exists a nonzero element $\gamma_0 \in L^2(\mu_A)$ satifying $S_{\mu_A}[\gamma_0] = 0$, $R[f] + \gamma_0$ also provides a parameter distribution that approximates the target f well. Theorem 3.3 shows that the regularization term removes the effect of γ_0, and the minimizer γ^* coincodes with the ridgelet transform except a small oscillation Δ_A.

The principal term $\gamma^*_{pri} := R[fp/(\beta + p)]$ of the obtained minimizer is understood as a *shrinkage estimator*, or a *biased estimator*, of $\gamma^\circ := R[f]$. Namely, while γ° exactly attains $S[\gamma^\circ] = f$, the obtained estimate $S[\gamma^*_{pri}] = fp/(\beta + p)$ is intentionally biased from f, and the norm $\|\gamma^*_{pri}\|$ is intentionally $p/(\beta+p)$-times smaller than $\|\gamma^\circ\|$. Recall that a regularized estimator is generally a biased estimator, and shrinkage is a natural consequence of ridge regression because the regularizer $\beta\|\gamma\|^2$ penalizes the norm of γ.

As described in Proposition B.2 for general settings, if $\beta \to +0$, then γ^* converges to the *minimum norm solution*. In our setting, by the continuity in β, it is simply given by $\lim_{\beta\to+0} \gamma^*_{pri} = \lim_{\beta\to 0} R[fp/(\beta + p)] = R[f] = \gamma^\circ$. However, we remark that this *does not*

mean that "If we minimize (6) without any regularization (by letting exactly $\beta = 0$), then $\gamma^* = R[f]$". In this case, the correct answer is $\gamma^* = R[f] + \ker S$. Namely, the minimizer will have a redundancy in null space $\ker S$.

3.3 Convergence of Finite Minimizers in the Over-parametrization Regime

The second main result is a convergence of parameter distributions of finite neural networks with over-parametrization.

Theorem 3.4. *Let $\{\gamma^*_{N,d}\}_{d=1}^{\infty}$ be a sequence of ERMers. Impose Assumption 3.2 on the hidden parameter distributions λ_d of $\gamma^*_{N,d}$, namely, λ_d weakly converges to μ_A. Assume $\beta_d \to \beta$ as $d \to \infty$. Then, for any bounded continuous function h on $\mathbb{I}_A^m \times \mathbb{T}$, we have*

$$\lim_{N\to\infty} \lim_{d\to\infty} \int h\gamma^*_{N,\lambda_d} \mathrm{d}\lambda_d = \int h\gamma^* \mathrm{d}\mu_A. \tag{13}$$

Here the limit with respect to N is in the sense of P-a.s. convergence.

Theorem 3.4 claims that the over-parametrized two-layer neural networks weakly converges to the population risk minimizers γ^* as the sample size gets increased. Combined with Theorem 3.3, we obtain the statement "an over-parametrized neural network converges to the ridgelet spectrum". The "weak convergence" is not much weak because if we take an arbitrary region of interest $K \subset \mathbb{R}^m$, and let the indicator function 1_K to be the test function h, then the parameter distribution eventually converges to $\int_K R[f](\boldsymbol{a}, b)\mathrm{d}\boldsymbol{a}\mathrm{d}b$. In Section 4 below, we will see that the parameters of finite neural networks trained by SGD accumulate the ridgelet spectrum.

We provide a remark on the assumption that λ_d converges to a measure λ. Since we consider the ridge regression, the support of ERMers cannot concentrate in a null set, for example, a lower dimensional submanifold, as the parameter number gets increased. More precisely, we have the following simple lemma:

Lemma 3.5. *Let λ be a finite Borel measure on $\mathbb{R}^m \times \mathbb{T}$. Let $\gamma \in L^2(\lambda)$. Assume $\|\gamma\|_{L^1(\lambda)} > C$ for some $C > 0$. Then we have $\|\gamma\|_{L^2(\lambda)} > C/\lambda(\mathrm{supp}(\gamma))$.*

This lemma implies if the the support of coefficient functions is collapsed to a null set, then its L^2-norm explodes. Therefore, the ridge regularization exclude such a coefficient function as a solution of the minimization problem in question.

Proof of Theorem 3.4. We provide a sketch of the proof. In fact, we prove a stronger convergence result as follows:

Lemma 3.6. *Let $\{\lambda_d\}_{d=1}^{\infty}$ ($\lambda_d \in \Lambda_{d,A}$) be a sequence of a finite Borel measure. Impose Assumption 3.2 on $\{\lambda_d\}_{d=1}^{\infty}$. Assume that $\beta_d \to \beta$ as $d \to \infty$. Then, as $d \to \infty$, we have*

$$\left\|\gamma^*[f; P_N, \lambda_d, \beta_d] - \gamma^*[f; P_N, \mu_A, \beta]\right\|_{L^2(\gamma_d)} \longrightarrow 0.$$

As a consequence, in the over-parametrized regime, the convergence occurs even when the hidden parameters are not optimized but at least they converge to μ_A. Combined with Lemma 3.1, the minimizer $\gamma^*[P_N, \mu_A]$ almost surely converges to γ^* as $N \to \infty$. □

4 NUMERICAL SIMULATION

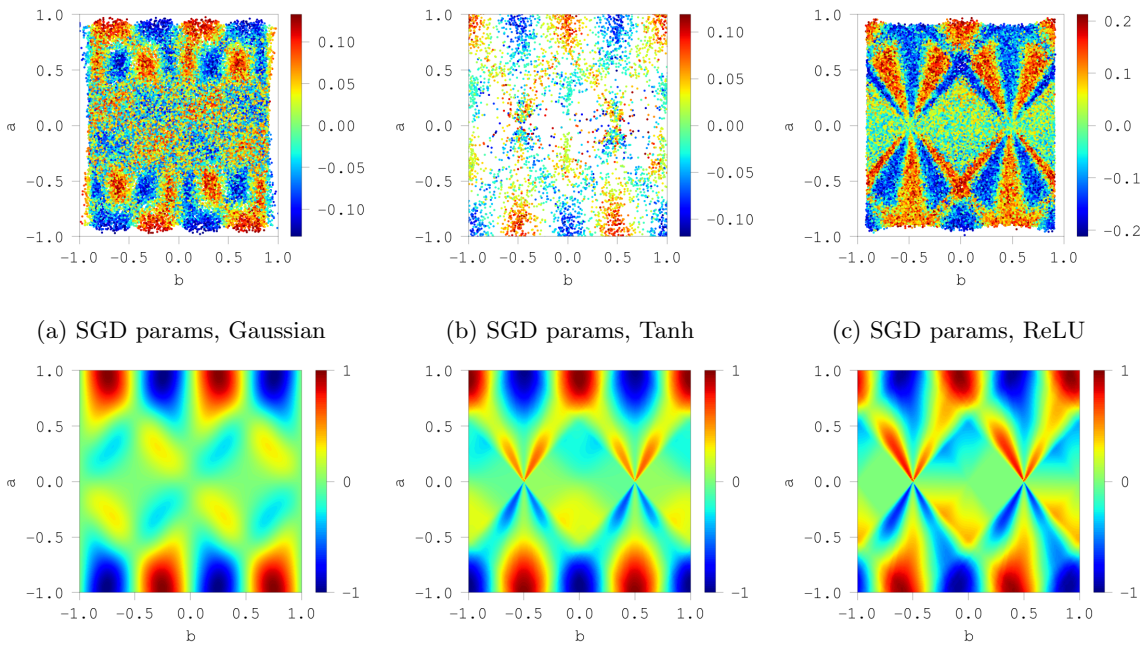
In order to verify the main results, we conducted numerical simulation with artificial datasets. Here, we only display the results of Experiment 1. The readers are also encouraged to refer Appendix D for further experimental results.

4.1 Data Generation

For the sake of visualization, all the datasets are 1-in-1-out, so that the scatter plot will be displayed in a three-dimensional manner: $(a, b) \in \mathbb{R}^2$ in position and $c \in \mathbb{R}$ in color. However, we remark that our theoretical results are valid for any dimension. We always consider the uniform distribution $x_i \sim U(-1, 1)$ for the input vectors, and generate $n = 1,000$ samples for training, except for the case of *Topologist's Sine Curve (TSC)* $y_i = \sin \frac{2\pi}{x_i}$. For the TSC, we generate $n = 10,000$ because the frequency tends to infinity as x tends to 0.

4.2 Scatter Plot of SGD Trained Parameters

Given a dataset $D_n = \{(x_i, y_i)\}_{i=1}^n$, we repeatedly train $s = 1,000$ neural networks $g(x; \theta^{(t)}) = \sum_{i=1}^d c_i^{(t)} \sigma(a_i^{(t)} x - b_i^{(t)}), (t \in [s])$ with activation function σ = periodic Gaussian, periodic Tanh and periodic ReLU. The training is conducted by minimizing the square loss: $L(\theta) = \frac{1}{n}\sum_{i=1}^n |y_i - g(x_i; \theta)|^2$ using stochastic gradient descent (SGD) with learning rate $\eta > 0$ and weight decay rate $\beta > 0$. Note that the weight decay has an equivalent effect to the L^2-regularization. In the main theory, only c is imposed L^2-regularization, and (a, b) are strictly restricted in a compact domain $\mathbb{I}_A^m \times \mathbb{T}$. However, in the experiments, all the parameters are imposed L^2-regularization for the sake of simplicity. The initial parameters are drawn from the uniform distribution $U(-1, 1)$. All the parameters are updated by SGD, so that this is *not* a random features method (Rahimi and Recht, 2008) in which hidden parameters (a, b) are frozen after initialization. After the training, we obtain sd sets



(a) SGD params, Gaussian (b) SGD params, Tanh (c) SGD params, ReLU

(d) R. spect, Gaussian (e) R. spect, Tanh (f) R. spect, ReLU

Figure 1: Parameter distributions $\gamma(\boldsymbol{a}, b)$ trained by SGD (top) and ridgelet spectra $R[f](\boldsymbol{a}, b)$ obtained by numerical integration (bottom) for the common data generating function $f(x) = \sin 2\pi x, (x \in [-1, 1])$.

of parameters $\{(a_i^{(t)}, b_i^{(t)}, c_i^{(t)})\}_{t\in[s], i\in[d]}$, and plot them in the (a, b, c)-space. (c is visualized in color.)

4.3 Heatmap of Ridgelet Spectrum

Given a dataset $D_n = \{(x_i, y_i)\}_{i=1}^n$, we approximately compute the ridgelet spectrum $R[f](a, b)$ of f at every sample points (a, b) by numerical integration:

$$R[f](a, b) \approx \frac{1}{n} \sum_{i=1}^{n} y_i \sigma(ax_i - b)\Delta x, \qquad (14)$$

where Δx is a normalizing constant, which is a constant because we assume that x_i be uniformly distributed. We remark that more sophisticated methods for the numerical computation of the ridgelet transform have been developed. See Do and Vetterli (2003) and Sonoda and Murata (2014) for example.

4.4 Results

In Figure 1, we compare the scatter plots of SGD trained parameters and the heatmaps of ridgelet spectra. All six figures are obtained from the common data generating function $f(x) = \sin 2\pi x$ on $[-1, 1]$. Despite the fact that the scatter plots and heatmaps are obtained from different procedures: numerical optimization and numerical integration, both figures share characteristics in common. For example, red and blue parameters in the scatter plots (a-c) concentrate in the area where the heatmaps (d-f) indicate the same colors. Due to the periodic assumption, the ridgelet spectrum spreads infinitely in b with period $T = 1$. On the other hand, due to the weight decay and initial locations of parameters, the SGD trained parameters gather around the origin. Here, we used the uniform distribution $U(-1, 1)$ for the initialization. We can understand that these differences between the scatter plot and ridgelet spectrum as the residual term $\Delta_{A,\beta}$ in the main theorem. Another remarkable fact is that the SGD trained parameters essentially did not change their positions in (a, b) from the initialized value. This is reasonable when the support of initial parameters overlap the ridgelet spectrum from the beginning. We can understand this phenomenon as the so-called lazy regime.

5 RELATED WORKS

A preprint by Sonoda et al. (2018) is the closest result with non-periodic σ. Compared to their result, we improved a lot. In their work, the function class of the data generating gunctions f remains to be an abstract RKHS $H^{\sigma\rho}$, the minimizer γ^* is given as an abstract projection of $R[f]$ onto a closed subspace, a hyper-parameter ρ in the ridgelet transform R remains not specified, and neither finite models nor finite sam-

ples are discussed. As far as we have noticed, we are the first to have revealed that the finite empirical minimizers do converge to the ridgelet spectrum.

Earlier Global Convergence Results. In the past, many authors have investigated the local minima of deep learning. However, these results have often posed strong assumptions such as that (A1) the activation function is limited to linear or ReLUs (Kawaguchi, 2016; Soudry and Carmon, 2016; Nguyen and Hein, 2017; Hardt and Ma, 2017; Lu and Kawaguchi, 2017; Yun et al., 2018); (A2) the parameters are random (Choromanska et al., 2015; Poole et al., 2016; Pennington et al., 2018; Jacot et al., 2018; Lee et al., 2019; Frankle and Carbin, 2019); (A3) the input is subject to normal distribution (Brutzkus and Globerson, 2017); or (A4) the target functions are low-degree polynomials or another sparse neural network (Yehudai and Shamir, 2019; Ghorbani et al., 2019). Due to these simplifying assumptions, we know very little about the minimizers themselves. In this study, from the perspective of harmonic analysis, we present a stronger characterization of the distribution of parameters in the over-parametrized regime. As a result, our theory (A1') accepts a wide range of activation functions, (A2') need not assume the randomness of parameter distributions, (A3') need not specify the data distribution, and (A4') preserves the universal approximation property of neural networks such as the density in L^2.

Mean-Field Theory. The *mean-field theory* (Rotskoff and Vanden-Eijnden, 2018; Mei et al., 2018; Sirignano and Spiliopoulos, 2020a,b) a.k.a. the *gradient flow* theory (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Arbel et al., 2019) has employed the integral representation and parameter distribution to prove the global convergence. These lines of studies claim that for the stochastic gradient descent learning of two-layer networks, the time evolution of a finite parameter distribution, say $\gamma_d(t)$, with parameter number d and continuous training time t, asymptotically converges to the time evolution of the continuous parameter distribution as $d \to \infty$. Here, the time evolution is described by a gradient flow, called *the partial differential equation, the Wasserstein gradient flow, or the McKean-Vlasov equation*, $\frac{\mathrm{d}}{\mathrm{d}t}\gamma_\infty(t) = -\frac{1}{2}\nabla_\gamma \|f - S[\gamma_\infty(t)]\|^2$ with initial condition $\gamma_\infty(0) = \gamma_{init}$. However, we should point out that these arguments oversights the null component in the parameter distributions. As we explained in Appendix A.4, the equation $f = S[\gamma]$ has an infinitely different solutions, say γ_1 and γ_2 that satisfy $S[\gamma_1] = S[\gamma_2]$ but $\gamma_1 \neq \gamma_2$. Hence, even though the convergence $S[\gamma_d] \to S[\gamma_\infty]$ in the function space $L^2(P)$ is established, in general, we *cannot* conclude the convergence $\gamma_d \to \gamma_\infty$ in the space of parameter distributions $L^2(\mu_A)$. This leaves the parameter distribution indeterminate. Nevertheless, our numerical simulation results have shown a "visual" convergence. By explicitly posing a regularization term on γ, we have specified the parameter distribution at the global minimum and have shown that the *weak convergence in the space of parameter distributions*: $\gamma_d \xrightarrow{w.} R[f]$. (We remark that some authors consider *noisy SGD*, which is equivalent to imposing the L^2-regularization.)

In order to avoid potential confusions, we provide supplementary explanations on the *trick* behind the mean-field theory. In the mean-field theory, the gradient flow $\mathrm{d}\gamma(t)/\mathrm{d}t = -\nabla\|f - S[\gamma(t)]\|^2$ is often explained as the *system of interacting particles* by identifying the parameters $\{(\boldsymbol{a}_i, b_i)\}_{i=1}^d$ as the coordinate system of d physical particles. The particles obeys a *non-linear equation of motion* with *interacting potential* $I[\gamma](\boldsymbol{a}, b) := \int K(\boldsymbol{a}, b; \boldsymbol{a}', b')\mathrm{d}\gamma(\boldsymbol{a}', b')$, where $K(\boldsymbol{a}, b; \boldsymbol{a}', b') := \int \sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\sigma(\boldsymbol{a}'\cdot\boldsymbol{x} - b')\mathrm{d}P(\boldsymbol{x})$, which is naturally derived by expanding the square loss function. Based on this physical analogy, this potential seems natural. However, here is the trick because in the potential I, the null space $\ker S$ is eliminated by *implicitly* applying S. Namely, since

$$I[\gamma](\boldsymbol{a}, b) = \int \sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)S[\gamma](\boldsymbol{x})dP(\boldsymbol{x}), \qquad (15)$$

we can verify that $I[\gamma + \ker S] = I[\gamma]$. This clearly indicates that the interactive potential is degenerate in γ, and thus the mean-field theory would only show a weaker convergence result than our main results.

Lazy Learning. The *lazy learning*, such as the *neural tangent kernel* (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019) and the *strong lottery ticket hypothesis* (Frankle and Carbin, 2019), employs a slightly different formulation of over-parametrization to investigate the inductive bias of deep learning. These lines of studies draw much attention by radically claiming that the minimizers are very close to the initialized state. In this study, we revealed that, in the (not lazy but) active regime, the shape of the parameter distribution converges to the ridgelet spectrum. According to our results, lazy learning is reasonable when the initial parameter distribution covers the ridgelet spectrum in its support, since the initial parameters need not to be *actively* updated. Furthermore, the lazy assumption can be reasonable when the data generating function f is a low frequency function, and thus the ridgelet spectrum $R[f]$ concentrates around the origin, because the initial parameter distribution is typically a normal (or sometimes a uniform) distribution centered at the origin $(\boldsymbol{a}, b) = (\boldsymbol{0}, 0)$ and thus eventually the initial parameters cover the ridgelet supectrum.

Implicit Regularization. Recently, gradient descent methods are said to impose *implicit regularization* (see eg. Zhang et al., 2017; Neyshabur, 2017; Gunasekar et al., 2018b,a), which often motivates the lazy learning. Although we have no unifying formulation of the implicit regularization to the present, and thus we have simply employed the L^2-regularization, we may formulate the implicitly regularized problem as the minimization problem of $J_{imp}[\gamma; f, \gamma_{init}] := \|f - S[\gamma]\|^2_{L^2(P)} + \beta\|\gamma - \gamma_{init}\|^2_{L^2(\mu_A)}$ for a given initial parameter distribution γ_{init} on $\mathbb{I}^m_A \times \mathbb{T}$. Then, immediately because $J_{imp}[\gamma; f, \gamma_{init}] = J[\gamma - \gamma_{init}; f]$, we can conclude that the minimizer γ^*_{imp} is given by $\gamma^* + \mathsf{Proj}_{\to \ker S}[\gamma_{init}]$, as $\beta \to 0$. Namely, the implicitly regularized solution γ_{imp} again meets a ridgelet spectrum γ^* but also holds a null component $\mathsf{Proj}_{\to \ker S}[\gamma_{init}]$. Investigation of the *role-of-null-space* would be an interesting future work.

6 CONCLUSION

In this study, we have derived the unique explicit expression—the ridgelet spectrum with residual—of over-parametrized two-layer neural networks trained by regularized empirical square risk minimization. To the present, many studies have proven the global convergence of deep learning. However, we know very little about the minimizer itself because the settings are typically very simplified. To investigate the minimizers, we develop the ridgelet transform on the torus, which is a complete set of new ridgelet transform. The scatter plots of learned parameters have shown a very similar pattern to the ridgelet spectra, which supports our theoretical result. Although we considered an idealized ERM, the visual convergence suggested much more. Extending our main theorem to a more realistic settings is our important future work. Moreover, although we assumed two-layer and ridge regression, as often assumed in recent over-parametrized theories, we conjecture that for a deep network, say $f_2 \circ f_1$ for example, each intermediate layer converges to ridgelet spectrums as $S[R[f_2]]$ and $S[R[f_1]]$; and that for a general loss function J, if it is continuous, namely $\|\gamma\| \leq CJ(\gamma)$, then the minimizer is given as a certain modified version of $R[f]$ (like $R[fp/(\beta + p)]$).

6.1 Further Discussions after Rebuttal

The Main Theorems mathematically rigorously show that finite ERMers eventually converge to the *unique closed-form solution* $R[fp/(\beta + p)]$. While conventional theories show the global convergence, our theory characterizes the limit point as the ridgelet transform, which complements the conventional theories. The uniqueness and the closed-form expression allow us to design theories at a higher resolution than, for example, those that simply assume and/or conclude a sub-Gaussian randomness of parameter distributions. For example, we can predict the shape of minimizers as presented in Sections A.5 and D. As for the quality of solutions, by the uniqueness of the minimizer and the continuity of integral representation operator S, if the loss value of a current solution γ_{local} is $\varepsilon \geq 0$, then the difference vector $\Delta\gamma := \gamma_{local} - \gamma_{global}$ is as small as $O(\varepsilon)$ in $L^2(\mathbb{R}^m \times \mathbb{R})$. Therefore, it is reasonable to say that regardless of the training process, a near-optimal solution also has a similar shape with ridgelet spectrum.

In the mean-field theory, it is known that the parameter distribution converges to a stable distribution, a.k.a. a *Gibbs distribution*, $\gamma_\infty \propto \exp(-\beta L)$ with regularization parameter β and loss function L, under certain convergence conditions (Mei et al., 2018; Tzen and Raginsky, 2020; Suzuki, 2020). The existence of such a distribution is a natural consequence of the fact that SGD is a stochastic gradient flow induced by a locally convex function. Note, however, that the Gibbs distribution contains an unknown loss function L, so in general the limit point itself cannot be given explicitly. In other words, the Gibbs distribution is an *equation* that encodes the sufficient conditions for a parameter distribution γ to be a limit point. In order to obtain the limit point in closed form, we need to solve this equation. The ridgelet transform can be understood as a closed-form solution for the Gibbs distribution. (To be exact, however, this study does not fully consider the convergence conditions proposed in mean-field theories, simply because these are still developing, and the current version of the convergence conditions are quite restrictive.) Again, closed-form solutions are more informative than equations.

Acknowledgements

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. We thank Taiji Suzuki and Atsushi Nitanda for productive comments on improving this study in many directions. This work was supported by JSPS KAKENHI 18K18113, JST CREST JPMJCR1913, JPMJCR2015, and JST ACTX JPMJAX2004.

References

Arbel, M., Korba, A., SALIM, A., and Gretton, A. (2019). Maximum Mean Discrepancy Gradient Flow. In *Advances in Neural Information Processing Systems 32*, pages 6481–6491.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On Exact Computation

with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems 32*, pages 8139–8148.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. (2018). Stronger Generalization Bounds for Deep Nets via a Compression Approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 254–263.

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

Bartlett, P., Foster, D. J., and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems 31*, pages 6240–6249.

Brutzkus, A. and Globerson, A. (2017). Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs. In *Proceedings of The 34th International Conference on Machine Learning*, volume 70, pages 605–614.

Candès, E. J. (1998). *Ridgelets: theory and applications*. PhD thesis, Standford University.

Chizat, L. and Bach, F. (2018). On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems 32*, pages 3036–3046.

Choromanska, A., LeCun, Y., and Ben Arous, G. (2015). Open Problem: The landscape of the loss surfaces of multilayer networks. In *The 28th Annual Conference of Learning Theory*, volume 40, pages 1–5.

Do, M. N. and Vetterli, M. (2003). The finite ridgelet transform for image representation. *Image Processing, IEEE Transactions on*, 12(1):16–28.

Frankle, J. and Carbin, M. (2019). The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations 2019*, pages 1–42.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. (2019). Limitations of Lazy Training of Two-layers Neural Network. In *Advances in Neural Information Processing Systems 32*, pages 9111–9121.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018a). Characterizing Implicit Bias in Terms of Optimization Geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1832–1841.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018b). Implicit Bias of Gradient Descent on Linear Convolutional Networks. In *Advances in Neural Information Processing Systems 31*, pages 9461–9471.

Hardt, M. and Ma, T. (2017). Identity Matters in Deep Learning. In *International Conference on Learning Representations 2017*, pages 1–14.

Helgason, S. (2011). *Integral Geometry and Radon Transforms*. Springer-Verlag New York.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pages 8571–8580.

Kawaguchi, K. (2016). Deep Learning without Poor Local Minima. In *Advances in Neural Information Processing Systems 29*, pages 586–594.

Kostadinova, S., Pilipović, S., Saneva, K., and Vindas, J. (2014). The ridgelet transform of distributions. *Integral Transforms and Special Functions*, 25(5):344–358.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer-Verlag Berlin Heidelberg.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. (2019). Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems 32*, pages 8572–8583.

Lu, H. and Kawaguchi, K. (2017). Depth Creates No Bad Local Minima. *arXiv preprint: 1702.08580*, pages 1–10.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

Murata, N. (1996). An integral representation of functions using three-layered betworks and their approximation bounds. *Neural Networks*, 9(6):947–956.

Neyshabur, B. (2017). *Implicit Regularization in Deep Learning*. PhD thesis, TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO.

Neyshabur, B., Tomioka, R., and Srebro, N. (2015). Norm-Based Capacity Control in Neural Networks. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1–26.

Nguyen, Q. and Hein, M. (2017). The Loss Surface of Deep and Wide Neural Networks. In *Proceedings of The 34th International Conference on Machine Learning*, volume 70, pages 2603–2612.

Nitanda, A. and Suzuki, T. (2017). Stochastic Particle Gradient Descent for Infinite Ensembles. *arXiv preprint: 1712.05438*.

Pennington, J., Schoenholz, S., and Ganguli, S. (2018). The emergence of spectral universality in deep networks. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1924–1932.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016). Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pages 3360–3368.

Rahimi, A. and Recht, B. (2008). Random Features for Large-Scale Kernel Machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Rotskoff, G. and Vanden-Eijnden, E. (2018). Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems 31*, pages 7146–7155.

Rubin, B. (1998). The Calderón reproducing formula, windowed X-ray transforms, and radon transforms in L^p-spaces. *Journal of Fourier Analysis and Applications*, 4(2):175–197.

Sirignano, J. and Spiliopoulos, K. (2020a). Mean Field Analysis of Neural Networks: A Law of Large Numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752.

Sirignano, J. and Spiliopoulos, K. (2020b). Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852.

Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., and Wetzstein, G. (2020). Implicit Neural Representations with Periodic Activation Functions. *arXiv preprint: 2006.09661*.

Sonoda, S., Ishikawa, I., Ikeda, M., Hagihara, K., Sawano, Y., Matsubara, T., and Murata, N. (2018). The global optimum of shallow neural network is attained by ridgelet transform. *arXiv preprint: 1805.07517*, pages 1–14.

Sonoda, S. and Murata, N. (2014). Sampling hidden parameters from oracle distribution. In *24th International Conference on Artificial Neural Networks (ICANN) 2014*, volume 8681, pages 539–546.

Sonoda, S. and Murata, N. (2017). Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268.

Soudry, D. and Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint: 1605.08361*, pages 1–12.

Starck, J.-L., Murtagh, F., and Fadili, J. M. (2010). The ridgelet and curvelet transforms. In *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, pages 89–118. Cambridge University Press.

Suzuki, T. (2020). Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics. In *Advances in Neural Information Processing Systems 33*, pages 19224–19237.

Tzen, B. and Raginsky, M. (2020). A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled McKean-Vlasov dynamics. *arXiv preprint: 2002.01987*.

Yehudai, G. and Shamir, O. (2019). On the Power and Limitations of Random Features for Understanding Neural Networks. In *Advances in Neural Information Processing Systems 32*, pages 6598–6608.

Yun, C., Sra, S., and Jadbabaie, A. (2018). Global Optimality Conditions for Deep Neural Networks. In *International Conference on Learning Representations 2018*, pages 1–14.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations 2017*, pages 1–15.