On Learnability via Gradient Method for Two-Layer ReLU Neural Networks in Teacher-Student Setting

Shunta Akiyama [1] **Taiji Suzuki** [1 2]

Abstract

Deep learning empirically achieves high performance in many applications, but its training dynamics has not been fully understood theoretically. In this paper, we explore theoretical analysis on training two-layer ReLU neural networks in a teacher-student regression model, in which a student network learns an unknown teacher network through its outputs. We show that with a specific regularization and sufficient over-parameterization, the student network can identify the parameters of the teacher network with high probability via gradient descent with a norm dependent stepsize even though the objective function is highly non-convex. The key theoretical tool is the measure representation of the neural networks and a novel application of a dual certificate argument for sparse estimation on a measure space. We analyze the global minima and global convergence property in the measure space.

1. Introduction

Deep learning empirically achieves high performance in many applications, such as computer vision and speech recognition. To explain its success from the theoretical view point, we need to reveal its optimization dynamics and the generalization ability of the solution that is obtained by a particular optimization method such as gradient descent. However, its training dynamics has not been fully understood theoretically and thus the generalization ability of the solution is still an open question. One of the difficulties of this problem is non-convexity of the associated optimization problem (Li et al., 2018) for the optimization aspect, and the high dimensionality induced by over-parameterization for the generalization aspect. In this study, we tackle these two problems in a teacher-student problem with the ReLU activation under an over-parameterized setting. In this setting, we need to take care of the non-differentiability of the ReLU activation and the over-specification problem due to the over-parameterization which potentially causes difficulty to show favorable generalization ability such as exact recovery.

The teacher-student setting is one of the most common settings for theoretical studies, e.g., Tian (2017); Yehudai & Shamir (2020); Goldt et al. (2019); Safran & Shamir (2018); Safran et al. (2020); Tian (2020); Suzuki & Akiyama (2021); Zhang et al. (2019); Zhou et al. (2021) to name a few. Zhong et al. (2017) studied the case where the teacher and student have the same width, showed that the strong convexity holds around the parameters of the teacher network and proposed a special tensor method for initialization to achieve the global convergence to the global optimal. However, its global convergence is guaranteed only for a special initialization which excludes a pure gradient descent method. Moreover, the over-parameterized setting is not included in their analysis. Safran & Shamir (2018) empirically showed that gradient descent is likely to converge to non-global optimal local minima, even if we prepare a student that has the same size as the teacher. More recently, Yehudai & Shamir (2020) showed that even in the simplest case where the teacher and student have the width *one*, there exists distributions and activations in which gradient descent fails to learn. Safran et al. (2020) showed the strong convexity around the parameters of the teacher network in the case where the teacher and student have the same width for Gaussian inputs. They also studied the effect of over-parameterization and showed that over-parameterization will change the spurious local minima into the saddle points. However, it should be noted that this does not imply that a gradient descent can reach the global optima.

To alleviate the non-convexity of neural network optimization, over-parameterization is one of the promising approaches. Indeed, it is fully exploited by (i) Neural Tangent Kernel (NTK) (Allen-Zhu et al., 2019; Arora et al., 2019; Jacot et al., 2018; Du et al., 2019; Weinan et al., 2020) and (ii) mean field analysis (Nitanda & Suzuki, 2017; Chizat &

[1] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan [2] Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Shunta Akiyama <shunta_akiyama@mist.i-tokyo.ac.jp>, Taiji Suzuki <taiji@mist.i.u-tokyo.ac.jp>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Bach, 2018; Chizat, 2019; Suzuki & Akiyama, 2021; Mei et al., 2019; Tzen & Raginsky, 2020). (i) In the setting of NTK, the gradient descent of neural networks can be seen as the convex optimization in RKHS, and thus it is easier to analyze. On the other hand, in this regime, it is hard to explain the superiority of deep learning, because the estimation ability of the obtained estimator is reduced to that of the corresponding kernel. (ii) In the setting of the mean field analysis, a kind of continuous limit of neural network is considered and its convergence to some specific target functions has been analyzed. This regime is more suitable in terms of a "beyond kernel" perspective, but it essentially deals with a continuous limit and hence is difficult to show convergence to a teacher network with a *finite width*.

In this paper, we make full use of the "measure representation" of two-layer ReLU networks as in the mean field analysis, while our approach employs a *sparse* regularization on the measure of parameters to show the convergence of a gradient descent method to the global optimum where the teacher network has a finite width. The sparse regularization on a measure space is well studied in a so-called *BLASSO* problem (De Castro & Gamboa, 2012). Indeed, Chizat (2019) analyzed the gradient descent for two layer neural networks from the view point of BLASSO analyses, and showed the convergence to the global optimal. However they assumed several assumptions which are hard to clarify, and excluded a non-smooth activation such as the ReLU activation. On the other hand, we explicitly present a realistic condition under which a gradient descent converges to the global optimum. More specifically, our contributions can be summarized as follows:

- We show that with an appropriate sparse regularization, the optimal solution of a regularized empirical risk can be arbitrarily close to the true teacher-parameters for a sufficiently small regularization parameter. This implies effectiveness of a sparsity inducing regularization in deep learning.

- We prove that a gradient descent with a norm-dependent step size can converge to the global optimum of the regularized learning problem if the student network is appropriately over-parameterized.

- Combining the above results, we show that a gradient descent method with an over-parameterized initialization can find a network which is arbitrary close to the true teacher network. In particular, the size of the estimated network becomes "narrow" even though the initial solution is over-parameterized, which explains the feature learning ability of neural networks leading a better performance than shallow methods such as kernel methods.

1.1. Other Related Works

BLASSO problem The BLASSO problem (De Castro & Gamboa, 2012) is a regression problem with total variation regularization on a measure space, which is an extension of the LASSO problem to the measure space. One of the main theoretical interests of BLASSO studies (Bredies & Pikkarainen, 2013; Candès & Fernandez-Granda, 2013; Duval & Peyré, 2015; Poon et al., 2018; 2019) is to clarify whether the global minima of BLASSO can recover the "true" measure in the setting where the true measure is sparse, i.e., given by a sum of Dirac measures. Duval & Peyré (2015) showed that for a sufficiently small sample noise and an appropriate regularization, the global minimum will also be sparse and close to the true measure. A key theoretical tool is a dual certificate, which is motivated by the Fenchel duality. However, their analysis assumes smoothness on the objective function and thus is not directly applied to our setting because of the non-differentiability of the ReLU activation.

Sparse regularization It has been shown that explicit or implicit sparse regularization such as L_1-regularization is beneficial to obtain better performances of deep learning under certain situations (Chizat & Bach, 2020; Gunasekar et al., 2018; Woodworth et al., 2020; Klusowski & Barron, 2016). However, it is still an open question that a gradient descent can find the teacher model in a regression setting with the ReLU non-linear activation. Bach (2017) analyzed a neural network model with a sparse regularization (L_1-regularization) which can be regarded as an extension of Barron class (Barron, 1993), and derived its model capacity. It was shown that the Frank-Wolfe type method can estimate a target function in the neural network model, but unfortunately this does not imply that a gradient descent method can estimate the target function. Moreover, it is not clear that each update of the Frank-Wolfe method is computationally tractable.

Langevin dynamics approach The gradient Langevin dynamics (GLD) is a useful approach to obtain a global optimum of a non-convex objective function (Welling & Teh, 2011; Raginsky et al., 2017; Erdogdu et al., 2018; Suzuki & Akiyama, 2021). This approach can be also applied to neural network optimization but such analysis would not give any information about the landscape of the neural network training. Among them, Suzuki & Akiyama (2021) considered an infinite dimensional Langevin dynamics, but they excluded a non-differentiable activation such as ReLU and did not give any landscape analysis.

1.2. Notations

Here we give some notations used in the paper. Let $\mathcal{M}(\mathcal{C})$ be the set of the Radon measures on a topological space

$\mathcal{C}$ (we consider the Borel algebra of $\mathcal{C}$ as the σ-field on which the Radon measures are defined). Let $\delta_w(\cdot)$ be the Dirac measure on $w \in \mathbb{R}^d$, i.e., $\int f(x)\delta_w(\mathrm{d}x) = f(w)$. Let $[m] := \{1, \dots, m\}$ for a positive integer m. Let the inner product between $x, y \in \mathbb{R}^d$ be $\langle x, y \rangle := \sum_{j=1}^d x_i y_i$.

2. Problem Settings

In this section, we give the problem setting and the model that we consider in this paper. We focus on a regression problem where we observe n training examples $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ generated by the following model:

$$y_i = f^\circ(x_i), \tag{1}$$

where $f^\circ : \mathbb{R}^d \to \mathbb{R}$ is the unknown true function that we want to estimate, $(x_i)_{i=1}^n$ are independently identically distributed from $P_{\mathcal{X}}$. Later on we assume that $P_{\mathcal{X}}$ is the uniform distribution on the unit ball $\mathbb{S}^{d-1}$ (Assumption 3.1).

Based on the observed data D_n, we construct an estimator $\hat{f}$ which is supposed to be "close" to the true function f°. As its performance measure, we employ the mean squared error defined by $\|\widehat{f} - f^\circ\|_{L_2(P_{\mathcal{X}})}^2 := \mathbb{E}_{X \sim P_{\mathcal{X}}}[(\widehat{f}(X) - f^\circ(X))^2]$. Its empirical version is defined by $\|\widehat{f} - f^\circ\|_n^2 := \frac{1}{n}\sum_{i=1}^n (\widehat{f}(x_i) - f^\circ(x_i))^2$.

Teacher-Student Model In this section, we prepare the teacher-student model that we consider in this paper. The student model is the two-layer neural network with the ReLU-activation $\sigma(u) = \max\{x, 0\}$ (Glorot et al., 2011) and width M, which is defined as

$$f(x; \Theta) = \sum_{j=1}^M a_j \sigma(\langle w_j, x \rangle), \tag{2}$$

where $\Theta = ((a_1, w_1), \dots, (a_M, w_M)) \in (\mathbb{R} \times \mathbb{R}^d)^M$ is the trainable parameter. The teacher model is assumed to be included in the student model but the width could be smaller than M:

$$f^\circ(x) = \sum_{j=1}^m a_j^\circ \sigma(\langle w_j^\circ, x \rangle), \tag{3}$$

where m is the width of the teacher model and $(a_j^\circ, w_j^\circ) \in \mathbb{R} \times \mathbb{R}^d$ $(j \in [m])$. We consider an over-parameterized setting where $m \leq M$ is assumed to be satisfied. Hence, the teacher model can be regarded as an element of the student model by setting $a_j = 0$ for $j = m+1, \dots, M$. For notational simplicity, we denote by $\Theta^\circ := (a_j^\circ, w_j^\circ)_{j=1}^m \in (\mathbb{R} \times \mathbb{R}^d)^m$.

For a neural network model, it is generally difficult to write the close form of the (regularized) empirical risk minimizer. Therefore, we typically optimize Θ via the gradient descent technique, but due to the non-convexity of the objective function, it is far from trivial that the global minima can be obtained by gradient descent.

Sparse Regularized Empirical Risk To estimate the true parameter Θ°, we define the following regularized empirical risk minimization problem on the parameter space $(\mathbb{R} \times \mathbb{R}^d)^M$:

$$\min_{\Theta \in (\mathbb{R} \times \mathbb{R}^d)^M} \frac{1}{2n}\sum_{i=1}^n (y_i - f(x_i; \Theta))^2 + \lambda \sum_{j=1}^M |a_j|\|w_j\|, \tag{4}$$

where $\lambda \geq 0$ is a regularization parameter. The regularization term $\lambda \sum_{j=1}^M |a_j|\|w_j\|$ can be seen as an L_1-regularization which induces sparsity. Indeed, by the scale homogeneity of ReLU ($a_j\sigma(\langle w_j, x\rangle) = a_j\|w_j\|\sigma(\langle w_j/\|w_j\|, x\rangle)$), we may reset the parameter as $a_j' = a_j\|w_j\|$ and $w_j' = w_j/\|w_j\|$ and then the regularization term can be rewritten as $\lambda \sum_{j=1}^M |a_j'|$. Apparently, this is the L_1-norm of $(a_j')_{j=1}^M$.

In practice, we typically use the L_2-regularization $\frac{\lambda}{2}\sum_{j=1}^M (a_j^2 + \|w_j\|^2)$ instead of the L_1-regularization as induced above. However, the arithmetic-geometric mean relation yields that

$$|a_j|\|w_j\| = \min_{\substack{(a_j', w_j') \in \mathbb{R} \times \mathbb{R}^d: \\ |a_j|\|w_j\| = |a_j'|\|w_j'\|}} \tfrac{1}{2}(|a_j'|^2 + \|w_j'\|^2). \tag{5}$$

Therefore, our sparse regularization can be replaced by the L_2-regularization. In this paper, we directly consider the sparse regularization instead just for simplicity.

Remark 2.1. *We will see that the regularization term $\lambda \sum_{j=1}^M |a_j|\|w_j\|$ corresponds to the total-variation norm regularization for the measure representation of the network which we refer to in the next section. The same type of regularization has been considered in several studies, e.g., E et al. (2019); Neyshabur et al. (2015). In those studies, it plays an important role to show a better performance of deep learning compared with kernel methods. We further make full use of the sparsity to show the exact recovery of the true parameter Θ° even under the over-parameterized setting.*

3. Global Minima in the Teacher-Student Setting

In this section, we show that the minimizer of the regularized empirical risk (4) is arbitrarily close to the teacher network f° for a sufficiently large sample size n. Note that we are not arguing here that the optimal solution can be obtained by the gradient descent, but the computational issue will be addressed in the next section. We make the following assumptions for our analysis.

Assumption 3.1. $(x_i)_{i=1}^n$ *are i.i.d. observations from the uniform distribution on* $\mathbb{S}^{d-1}$, *that is,* $P_{\mathcal{X}} = \mathrm{Unif}(\mathbb{S}^{d-1})$.

Assumption 3.2. *The teacher network* $f^\circ = \sum_{j=1}^m a_j^\circ \sigma(\langle w_j^\circ, \cdot \rangle)$ *satisfies the following conditions:*

1. $a_j^\circ > 0 \quad (\forall j \in [m])$.
2. $\langle w_{j_1}^\circ, w_{j_2}^\circ \rangle = 0 \quad (\forall j_1, j_2 \in [m],\ j_1 \neq j_2)$.

The second assumption could be a bit strong, but the same assumption has been considered in several previous researches (Zhong et al., 2017; Tian, 2017; Safran & Shamir, 2018; Safran et al., 2020; Li et al., 2020). For example, Safran et al. (2020) analyzed the landscape of the objective under this assumption and showed a negative result that the loss landscape around the global minima is not even *locally convex*. On the other hand, they also showed that an over-parameterization turns a non-global optimal point into a saddle-point. However, they have not shown that a gradient descent can reach the optimal solution. Li et al. (2020) showed a global optimality of gradient descent in a specific teacher student setting under this condition. They consider a specific teacher model $f^\circ(x) = \sum_{j=1}^M a_j^\circ |\langle x, \theta_j^\circ \rangle|$ for $a_j^\circ > 0$ and a student model $f(x; W) = \frac{1}{M}\sum_{j=1}^M \|w_j\| \sigma(\langle w_j, x \rangle)$. This is relevant to ours, but specification of the teacher network is quite different from our setting.

The main ingredient of our analysis is the *measure representation* of the two layer ReLU-neural network. Using this representation, one can regard the neural network training as a sparse regularized learning on the measure space. This enables us to show (near) exact recovery. In particular, the *Beurling-LASSO (BLASSO)* analysis (De Castro & Gamboa, 2012) which could be seen as an infinite dimensional extension of sparse regularization theory is helpful.

3.1. Mesure Representation of Two-Layer Neural Networks and BLASSO Problem

We introduce the measure representation of the two-layer ReLU neural network. By using 1-homogeneity of the ReLU activation, it holds that

$$\begin{aligned}\sum_{j=1}^M a_j \sigma(\langle w_j, x \rangle) &= \sum_{j=1}^M a_j \|w_j\| \sigma\left(\left\langle \frac{w_j}{\|w_j\|}, x \right\rangle\right) \\ &= \int_{\mathbb{S}^{d-1}} \sigma(\langle \theta, x \rangle) \mathrm{d}\nu(\theta)\end{aligned} \tag{6}$$

with $\nu = \sum_{j=1}^m a_j \|w_j\| \delta_{w_j/\|w_j\|} \in \mathcal{M}(\mathbb{S}^{d-1})$. We call this ν a measure representation of the two-layer ReLU neural network. In the following, we write

$$f(x; \nu) = \int_{\mathbb{S}^{d-1}} \sigma(\langle \theta, x \rangle) \mathrm{d}\nu(\theta). \tag{7}$$

Under this representation, the teacher network is represented as $\nu^\circ = \sum_{j=1}^m r_j^\circ \delta_{\theta_j^\circ}$ with $r_j = a_j^\circ \|w_j^\circ\|$ and $\theta_j^\circ = w_j^\circ / \|w_j^\circ\|$.

Remark 3.3. *For a more general activation* σ, *we need to consider a measure on the product space* $\mathbb{R} \times \mathbb{R}^d$. *However, thanks to the 1-homogeneity of ReLU, we only need to consider a measure on* $\mathbb{S}^{d-1}$ *which is a compact metric space.*

With this measure representation, we may consider the following regression problem on the measure space instead:

$$\min_{\nu \in \mathcal{M}(\mathbb{S}^{d-1})} \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i; \nu))^2 + \lambda \|\nu\|_{\mathrm{TV}}, \tag{8}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the *total variation* norm of $\nu \in \mathcal{M}(\mathbb{S}^{d-1})$ that is defined by $\|\nu\|_{\mathrm{TV}} = \nu_+(\mathbb{S}^{d-1}) + \nu_-(\mathbb{S}^{d-1})$ for the Hahn–Jordan decomposition $\nu(\cdot) = \nu_+(\cdot) - \nu_-(\cdot)$. This can be seen as the continuous version of the original problem (4), which is called a BLASSO problem (De Castro & Gamboa, 2012). Since the measure representation covers any finite-width neural network, the following proposition holds.

Proposition 3.4. *Assume that a global minimum of* (8) *is obtained by a measure which is represented as a finite sum of Dirac measures:*

$$\nu^* = \textstyle\sum_{j=1}^{m^*} r_j^* \delta_{\theta_j^*},$$

then for the student network satisfying $M \geq m^*$, *the global minima of (4) can be obtained by the form whose measure representation is written by* ν^*.

There have been several studies that focused on the global minimum of the BLASSO problem (8). Duval & Peyré (2015) analyzed this problem in the context of sparse spike deconvolution, in which f is a Gaussian convolution filter and is an element of $L_2(\mathbb{T})$ (where $\mathbb{T}$ denotes the 1-dimensional torus), and showed that under the so-called *NDSC condition*, the global minima can be close to underlying measure. Poon et al. (2018; 2019) analyzed a more general setting and derived a sufficient condition for the NDSC condition. However, these analyses have required smoothness on the objective. Therefore, they can not be applied directly to our setting because of non-differentiability of the ReLU activation. We overcome this difficulty by directly deriving the *dual certificate* of the optimization problem.

3.2. Main Result 1: Global Minima of Regularized Empirical Risk

We prove that with a sufficiently small regularization parameter, the global minimizer of (8) is close to the teacher network with an arbitrarily small gap. We state this as the following theorem.

Theorem 3.5. *Assume that Assumptions 3.1 and 3.2 are satisfied. Suppose that $n > \mathrm{poly}(m, d, \log 1/\delta)$ for $\delta > 0$. Then, with probability at least $1 - \delta$, we have that, for any $\epsilon > 0$, with sufficiently small $\lambda > 0$, the optimal solution of (8) is uniquely determined and written by the form $\nu^* = \sum_{j=1}^m r_j^* \delta_{\theta_j^*}$ where $(r_j^*, \theta_j^*)_{j=1}^m \subset \mathbb{R} \times \mathbb{S}^{d-1}$ satisfy*

$$\begin{cases} \sum_{j=1}^m |r_j^\circ - r_j^*|^2 \leq \mathrm{O}(m\lambda^2) \\ \sum_{j=1}^m \mathrm{dist}^2(\theta_j^*, \theta_j^\circ) \leq \mathrm{O}(m\lambda^2) \end{cases} . \tag{9}$$

The proof can be found in Appendix A. From this theorem and Proposition 3.4, we immediately obtain the following corollary.

Corollary 3.6. *Under the same assumption with Theorem 3.5, for the student network model with more than m nodes, the optimal solution of (4) achieves the same property with Theorem 3.5, i.e., the measure representation of the optimal network satisfies (9).*

Therefore, as long as the network size M is sufficiently large such that $M \geq m$, we can recover the true network with arbitrarily small error by tuning the regularization parameter. The event of this property is uniform over the choice of the accuracy ϵ and corresponding regularization parameter λ. Hence, by decreasing λ gradually, we can finally recover the teacher model exactly. This result only characterizes the globally optimal solution and it does not say anything about the algorithmic convergence of a gradient descent method. In the next section, we address this issue.

Proof Strategy: Dual Certificate Theorem 3.5 can be shown through a dual certificate characterization of the optimal solution. Let the optimization problem (8) be P_λ. By the Fenchel's duality theorem (Rockafeller, 1967; Borwein & Zhu, 2005; Duval & Peyré, 2015), its dual problem D_λ is given by

$$(D_\lambda) \qquad \max_{p \in \mathbb{R}^n : \|f^*(p)\|_\infty \leq 1} \frac{1}{n^2} \sum_{i=1}^n y_i p_i - \frac{\lambda}{2n^2} \|p\|^2,$$

where $f^*(p)(\cdot) \in \mathcal{C}(\mathbb{S}^{d-1})$[1] that is defined by $f^*(p)(\theta) := \frac{1}{n} \sum_{i=1}^n \sigma(\langle \theta, x_i \rangle)$, and the strong duality holds, that is, ν_λ^* is the optimal solution of P_λ if the following optimality condition is satisfied for the *unique* solution p_λ of D_λ (the uniqueness of the dual solution follows from the strong convexity of the dual problem):

$$\begin{cases} f^*(p_\lambda) \in \partial \|\nu_\lambda^*\|_{\mathrm{TV}}, \\ p_{\lambda,i} = -\frac{1}{\lambda}(f(x_i; \nu_\lambda^*) - y_i) \quad (\forall i \in [n]). \end{cases}$$

We call $f^*(p_\lambda)$ a *dual certificate* for ν_λ^*. Conversely, if this condition is satisfied by $(\nu_\lambda^*, p_\lambda) \in \mathcal{M}(\mathbb{S}^{d-1}) \times \mathbb{R}^n$, then the pair is the optimal solution of both P_λ and D_λ. Therefore, our strategy is to show that the dual certificate $f^*(p_\lambda)$ admits only a primal optimal solution ν_λ^* that satisfies the condition in the theorem, i.e., the support of ν_λ^* consists of only m distinct points each of which is close to the true parameters $(\theta_j^\circ)_{j=1}^m$. To prove this, we show that there exist $(\theta_j^*)_{j=1}^m$ such that $(\mathrm{dist}(\theta_j^*, \theta_j^\circ))_{j=1}^m$ are sufficiently small and satisfy

$$\begin{cases} f^*(p_\lambda)(\theta_j^*) = 1 \quad (\forall j \in [m]), \\ |f^*(p_\lambda)(\theta)| < 1 \quad (\forall \theta \in \mathbb{S}^{d-1}/\{\theta_1^*, \ldots, \theta_m^*\}) \end{cases} \tag{10}$$

for sufficiently small λ. From this inequality, we can show that $(|r_j^* - r_j^\circ|)_{j=1}^m$ will also be sufficiently small. Finally by using the form $\nu^* = \sum_{j=1}^m r_j^* \delta_{\theta_j^*}$ and strong convexity of the empirical risk term in P_λ w.r.t. r_j^* and θ_j^* around the teacher parameters $(r_j^\circ, \theta_j^\circ)_{j=1}^m$, we get the quantitative bound as Eq. (9).

For that purpose, we particularly consider a setting where $\lambda = 0$, and consider the minimal norm certificate:

$$p_0 := \min\{\|p\| \mid p \in \mathbb{R}^n \text{ is a feasible solution of } D_0\}.$$

The most difficult pint in our analysis is to show the property (10) for the minimal norm certificate p_0. This is accomplished by carefully evaluating the analytic form of $f^*(p_0)$. Indeed, by using the orthogonality of $(\theta_j^\circ)_{j=1}^m$ and the fact that the input distribution is the uniform distribution, we can write down the minimal norm certificate and analyze it.

4. Global Convergence of Gradient Method

In this section, we investigate a gradient descent method for the optimization problem (4). We show that under some assumptions, a gradient descent with a norm-dependent step size converges to the global optimum of the problem. We also show that these assumptions for the global convergence are satisfied under the conditions we made in the previous section, which implies the identifiability of the teacher parameters through the gradient descent method.

4.1. Norm-Dependent Gradient Descent

We consider a standard gradient descent for optimizing the objective (4). To incorporate the 1-homogeneity of the ReLU activation function, we employ a step size that can be dependent on the norm of each parameter. As we see in proof of the global convergence, this norm dependency is helpful to describe an update in the measure space. Let F be the regularized empirical risk given in (4), that is, $F(\Theta) := \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i; \Theta))^2 + \lambda \sum_{j=1}^M |a_j| \|w_j\|$. Then, the update rule of the norm-dependent gradient descent can be

[1] $\mathcal{C}(S)$ is the set of continuous functions on a topological space S.

written as

$$a_{j,k+1} = a_{j,k} - \eta_{j,k} g_j(\Theta_k) \text{ for } g_j(\Theta_k) \in \partial_{a_j} F(\Theta_k),$$
$$w_{j,k+1} = w_{j,k} - \eta_{j,k} h_j(\Theta_k) \text{ for } h_j(\Theta_k) \in \partial_{w_j} F(\Theta_k),$$

where $\Theta_k = ((a_{1,k}, w_{j,k}), \dots, (a_{M,k}, w_{M,k}))$ is the parameter after k iterations, $\eta_{j,k} > 0$ is the norm-dependent step size which will be specified below. $\partial_a F(\Theta)$ denotes the sub-gradient of $F(\Theta)$ as a function of a. The sub-gradient is not always a singleton, but we employ the following one as g, h:

$$g_j(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\Theta) - y_i)\sigma(\langle w_j, x_i\rangle) + \lambda \operatorname{sgn}(a_j)\|w_j\|,$$
$$h_j(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i;\Theta) - y_i) a_j x_i \mathbb{1}\{\langle w_j, x_i\rangle \geq 0\} + \lambda \frac{|a_j| w_j}{\|w_j\|},$$

As for the norm-dependent step size $\eta_{j,k}$, we employ the following representation:

$$\eta_{j,k} = \alpha \frac{|a_{j,k}|\|w_{j,k}\|}{a_{j,k}^2 + \|w_{j,k}\|^2}, \tag{11}$$

where $\alpha > 0$ is a fixed constant. For the initialization, we consider the mean-field setting where each $a_{j,0} = \mathrm{O}(1/M)$:

$$a_{j,0} = \frac{2}{M} \quad (1 \leq j \leq M/2),$$
$$a_{j,0} = -\frac{2}{M} \quad (M/2 + 1 \leq j \leq M),$$
$$w_{j,0} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}).$$

With the norm-dependent step size, the sign of $a_{j,k}$ will not be changed during the optimization, and thus we need the both positive and negative sign initializations for $(a_{j,0})_{j=1}^{M}$. As pointed out by several authors (Chizat & Bach, 2018; Chizat, 2019; Suzuki & Akiyama, 2021; Mei et al., 2019), it is essentially important to analyze the dynamics of "feature learning" in the mean field regime where each node is adaptively updated to represent the target function efficiently. This is in contrast to NTK analysis (a.k.a., lazy training regime) where the basis functions are almost fixed during the optimization. The algorithm is summarized in Algorithm 1.

The global optimality of the gradient descent can be shown through the measure representation of the neural network. Indeed, we have seen in the previous section that the optimization problem of a neural network model can be generalized to the BLASSO problem on the measure space as presented in Eq. (8). Let J be the BLASSO objective function on the measure space: $J(\nu) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i;\nu))^2 + \lambda\|\nu\|_{\mathrm{TV}}$. Note that in the over-parameterized setting, we cannot formally define the convergence of the parameter Θ_k to the true one Θ° because they have different dimensionality. Therefore, we consider convergence of the measure corresponding to the parameter Θ instead. We assume "sparsity" of the global minima of J on the measure space to ensure the convergence of the measure representation as follows.

Algorithm 1 Norm-Dependent Gradient Descent

```
Input: student width $M$ (even), max iteration $K$, stepsize parameter $\alpha > 0$.
    Initialization : $a_{j,0} = 2/M, 1 \leq j \leq M/2$, $a_{j,0} = -2/M, M/2+1 \leq j \leq M$, $w_{j,0} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$
for $k = 1, 2, \dots, K$ do
    for $j = 1, \dots, M$ do
        $\eta_{j,k} = \alpha \frac{|a_{j,k}|\|w_{j,k}\|}{a_{j,k}^2 + \|w_{j,k}\|^2}$
        choose $g_j(\Theta_k) \in \partial_{a_j} F(\Theta_k)$
        choose $h_j(\Theta_k) \in \partial_{w_j} F(\Theta_k)$
        $a_{j,k+1} = a_{j,k} - \eta_{j,k} g_j(\Theta_k)$
        $w_{j,k+1} = w_{j,k} - \eta_{j,k} h_j(\Theta_k)$
    end for
end for
```

Assumption 4.1. *ar The global minimum of J is uniquely attained by a sum of Dirac measures:*

$$\nu^* := \sum_{j=1}^{m^*} r_j^* \delta_{\theta_j^*}, \tag{12}$$

where m^* is a positive integer, $r_j^* \neq 0$, $\theta_j^* \in \mathbb{S}^{d-1}$ ($j \in [m^*]$) and $\theta_j^* \neq \theta_{j'}^*$ for any $j \neq j'$.

Remark 4.2. *Note that this condition can be satisfied under Assumptions 3.1 and 3.2 by Theorem 3.5.*

By the same argument as Proposition 3.4, if we set $M \geq m^*$, the sparsity and uniqueness of the global minimum of J leads to the existence of the global minimum of F, which is essentially represented by m^* nodes. Even in this case, by the non-convexity of F, it is far from trivial to show the convergence of the gradient method to the global optimal solution. As we have stated, we show this through the measure representation of the network.

To show the result, we prepare some additional notations. For the intermediate solution $\Theta_k = \{(a_{j,k}, w_{j,k})\}_{k=1}^{M}$, we define $r_{j,k} = a_{j,k}\|w_{j,k}\|, \theta_{j,k} = \frac{w_{j,k}}{\|w_{j,k}\|}$ (if $\|w_{j,k}\| = 0$, we set $\theta_{j,k}$ be arbitrary fixed point in $\mathbb{S}^{d-1}$). Accordingly, the measure representation corresponds to Θ_k be

$$\nu_k := \textstyle\sum_{j=1}^{M} r_{j,k}\delta_{\theta_{j,k}}.$$

For two Radon measures $\mu_1, \mu_2 \in \mathcal{M}(\mathbb{S}^{d-1})$, $W_\infty(\mu_1, \mu_2)$ denotes the Wasserstein distance between them: $W_\infty(\mu_1, \mu_2) := \inf_{\gamma \in \Pi(\mu_1,\mu_2)} \sup_{(\theta_1,\theta_2)\in\mathrm{supp}(\gamma)} \mathrm{dist}(\theta_1, \theta_2)$, where $\Pi(\mu_1, \mu_2)$ is a set of product measures with marginals μ_1 and μ_2, $\mathrm{supp}(\gamma)$ is the support of γ, and $\mathrm{dist}(\theta_1, \theta_2) := \arccos(\langle \theta_1, \theta_2 \rangle)$ for $\theta_1, \theta_2 \in \mathbb{S}^{d-1}$.

Since $f(x; \nu)$ is a linear model with respect to ν and the squared loss is differentiable, the Fréchet subdifferential of $J(\nu)$ on $\mathcal{M}(\mathbb{S}^{d-1})$ can be defined and be represented as a set of functions $G(\cdot) : \mathbb{S}^{d-1} \to \mathbb{R}$ defined by

$$G(\theta) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i; \nu) - y_i) \sigma(\langle \theta, x_i \rangle) + \lambda \eta(\theta),$$

where $\eta \in \mathcal{C}(\mathbb{S}^{d-1})$ satisfies $\|\eta\|_\infty \leq 1$ and $\int \eta \mathrm{d}\nu = \|\nu\|_{\mathrm{TV}}$. Note that we have that $\partial J(\nu_k) := \{G \in \mathcal{C}(\mathbb{S}^{d-1}) \mid J(\mu) - J(\nu_k) \geq \int G(\theta) \mathrm{d}(\mu - \nu_k)$ for any $\mu \in \mathcal{M}(\mathbb{S}^{d-1})\}$ which is well defined because $J(\cdot)$ is a convex function on the measure space $\mathcal{M}(\mathbb{S}^{d-1})$.

4.2. Main Result 2: Global Optimality of Gradient Method

Here, we give the global convergence property of the norm-dependent gradient descent under a bit milder conditions than those assumed in the previous section. The analysis basically follows that of Chizat (2019), but they assumed smoothness on the activation and excluded the ReLU activation. To overcome this difficulty, our norm-dependent step size (Eq. (11)) plays the important role. Moreover, we carefully divide the parameter space into "smooth region" and "non-smooth irrelevant-region" to show a descent property of the objective. The assumptions below are made under a condition of a training data observation $D_n = (x_i, y_i)_{i=1}^n$.

Assumption 4.3 (Non-orthogonality between x and θ). *For any $i \in [n], j \in [m]^*$, we have $\langle x_i, \theta_j^* \rangle \neq 0$.*

Assumption 4.4 (Strong convexity w.r.t. r). *There exists a constant $\kappa > 0$ such that for any $r_1, \ldots, r_m \in \mathbb{R}$, $\| \sum_{j=1}^m r_j \sigma(\langle \theta_j^*, \cdot \rangle)\|_n^2 \geq \kappa(r_1^2 + \cdots + r_m^2)$.*

Assumption 4.5 (Non-degeneracy). *There exists no $\theta \notin \mathrm{supp}(\nu^*)$ such that $J'(\nu^*)(\theta) = 0$.*

Assumption 4.6 (Boundedness). *There exists a constant $C_F > 0$ such that, for any k, it holds that $F(\Theta_k) \leq C_F$.*

Assumption 4.7 (Boundedness of input). *$\|x_i\| \leq 1$ for all $i \in [n]$.*

Assumption 4.3 is satisfied almost surely if $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. This is required to ensure the smoothness of the objective around the optimal parameter $(r_j^*, \theta_j^*)_{j=1}^{m^*}$. Otherwise the objective function F is non-differentiable at the global optimal with respect to θ_j, which causes difficulty to show the local convergence around the global optimal. Assumption 4.4 is also almost surely satisfied if the nodes $x \mapsto \sigma(\langle x, \theta_j^* \rangle)$ ($j \in [m^*]$) are linearly independent in $L_2(P_{\mathcal{X}})$. Assumption 4.5 is a bit tricky but is assumed in several existing work (Duval & Peyré, 2015; Chizat, 2019; Flinth et al., 2020) ensures that the true parameters $(\theta_j^*)_{j=1}^{m^*}$ are uniquely determined. Assumption 4.5 is also needed to ensure that in a local convergence phase, which we describe in Theorem 4.8, ν_k vanishes rapidly far away from $(\theta_j)_{j=1}^{m^*}$. This assumption can be verified under the same setting as Theorem 3.5 by utilizing a dual certificate argument. Assumption 4.7 is just fixing the scaling factor and is satisfied under the setting $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ (Assumption 3.1).

Theorem 4.8. *Assume that Assumptions 4.1, 4.3–4.7 hold. Let $\tau = \mathrm{Unif}(\mathbb{S}^{d-1})$, $\nu_0^+ = 2/M \sum_{j=1}^{M/2} \delta_{w_{j,0}}, \nu_0^- = 2/M \sum_{j=M/2+1}^{M} \delta_{w_{j,0}}$ and $J^* = J(\nu^*)$. Then, for any $0 < \epsilon < 1/2$, there exist constants $\rho, C, C', C_M > 0$, $J_0 > J^*$, $\kappa_0 > 0$ such that if $\alpha > 0$ satisfies*

$$\alpha < \min\{(J_0 - J^*)^{1+\epsilon/2}/C, 1/8C_1, \\ 1/10C_2, \rho/C_2, \lambda^2/8C_F^2\}$$

with $C_1 = 2\sqrt{n}C_F + \lambda$ and $C_2 = 2\sqrt{n}C_F$, the width M is sufficiently over-parameterized as $M \geq C_M \exp(\alpha^{-2})/\alpha$, and the initial solution satisfies

$$\max\{W_\infty(\tau, \nu_0^+),\ W_\infty(\tau, \nu_0^-)\} \leq (J_0 - J^*)/C,$$

then we have the following convergence properties:
(1) Global exploration: There exists $k_0 \geq C'(J_0 - J^*)^{-(2+\epsilon)}$ such that for any $k \geq k_0$, it holds that

$$J(\nu_k) - J^* \leq J_0 - J^*.$$

(2) Local convergence: For any $k \geq k_0$, it holds that

$$J(\nu_k) - J^* \leq (J(\nu_0) - J^*)(1 - \kappa_0)^{k-k_0}.$$

Therefore, combining these results, we see that $J(\nu_k)$ converges to $J(\nu^*)$.

The proof can be found in Appendix B. This theorem implies that the norm-dependent gradient descent can converge to the global optimal solution in terms of both the measure on parameters and the function value. Its dynamics consists of two phases: (1) the global exploration regime, and (2) the local linear convergence regime. In the first phase, the gradient descent explores the parameter space to roughly capture the location of the optimal parameters. In the second phase, the dynamics enters a local region around the optimal parameters where the objective is locally strongly convex. After entering this phase, the parameters converge to the optimal solution linearly. In that sense, J_0 represents a threshold that separates the global region and local near strongly convex region. During the optimization, the sparse regularization works for eliminating the amplitudes of nodes that are far away from the optimal parameters. This kind of "two phase" dynamics has been pointed out by several authors (e.g., Li & Yuan (2017); Chizat (2019)), but it has not been shown for the ReLU fully connected neural networks.

The condition $\max\{W_\infty(\tau, \nu_0^+),\ W_\infty(\tau, \nu_0^-)\} \leq (J_0 - J^*)/C$ requires that M is sufficiently over-parameterized. It is known that $W_\infty(\tau, \nu_0^\pm) = \mathrm{O}_p((\log M)^{1/(d-1)} M^{-1/(d-1)})$ for $d > 3$ (Trillos & Slepčev, 2015). Therefore, it is implicitly assumed that $M \geq \Omega((J_0 - J^*)^{-(d-1)} \log_+(1/(J_0 - J^*))^{(d-1)})$[2]. The condition $M \geq C_M \exp(\alpha^{-2})/\alpha$ also requires the over-parameterization and the right side may be quite large. This condition is only required for the global exploration ((1) in Theorem 4.8). The over-parameterization and the norm-dependency of stepsize ensure that $(\theta_{j,k})_{j=1}^M$ do not move far away from initialization until the function value decrease enough. By this property, the gradient descent can "identify" an informative subset of parameters $(\theta_{j,k})_{j=1}^M$, which are close to the optimal parameters $(\theta_j^*)_{j=1}^{m*}$. It may be possible to ensure that under the less number of parameters M, the gradient descent "automatically" reaches around each of the optimal parameters and can accomplish the global exploration. We leave this issue for future work. Finally, we mention a remark on a condition on the constant ρ and the regularization parameter λ for Theorem 4.8. Roughly speaking, ρ represents a diameter of a local smooth region around each optimal parameter θ_j^*. Under Assumptions 3.1 and 3.2, it suffices to take $\rho = \mathrm{O}_p(1/nm)$ if θ_j^* and θ_j° are sufficiently close for any $j \in [m]$ (see Lemma B.18). It can be shown that this closeness condition between θ_j^* and θ_j° holds with high probability by setting $\lambda = \mathrm{O}(1/nm^{3/2})$ by Theorem 3.5. These estimates are derived from conservative evaluations and could be larger for each concrete realization of $(x_i)_{i=1}^n$.

In addition to this convergence property in terms of the objective function, we can show convergence in terms of the L_∞-norm.

Theorem 4.9. *Under Assumptions 4.1, 4.3–4.7, there exists $C'' > 0$ such that for all $k \geq k_0$, it holds that*

$$\|f(x; \nu_k) - f(x; \nu^*)\|_\infty \leq C''(J(\nu_0) - J^*)(1 - \kappa_0)^{k-k_0},$$

where k_0 and κ_0 are those introduced in Theorem 4.8.

To show this, we prove that the measure representation ν_k converges to the optimal representation ν^* in terms of a modified 2-Wasserstein distance. The details can be found in Section B.6.

Near Exact Recovery by Gradient Descent Finally, combining Theorem 3.5 and Theorem 4.9, we obtain the following corollary that asserts that the student network converges near the teacher network by the gradient descent method. To show this, we need to prove that Assumptions 3.1 and 3.2 implies Assumptions 4.1, 4.3–4.7. The proof can be found in Section B.6.

[2] $\log_+(x)$ denotes $\max\{\log(x), 1\}$.

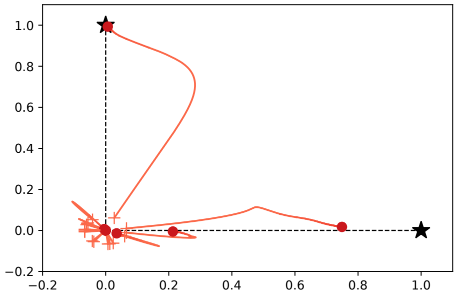

Figure 1. Illustration of the optimization dynamics with $d = 2$ and $m = 2$. The true parameters are indicated by $\star$, the initial solution of each node is indicated by an orange $+$, and its final state is indicated by the red $\circ$.

Corollary 4.10. *Under Assumptions 3.1 and 3.2, suppose that $n > \mathrm{poly}(m, d, \log 1/\delta)$ for $\delta > 0$, then, with probability at least $1 - \delta$, it holds that the $L_2(P_\mathcal{X})$-norm between $f(\cdot; \nu_k)$ and f° can be bounded as*

$$\begin{aligned}&\|f(\cdot; \nu_k) - f^\circ\|_{L_2(P_\mathcal{X})}^2 \\ &\leq 2C''^2 (J(\nu_0) - J^*)^2 (1 - \kappa_0)^{2(k-k_0)} + O(m\lambda^2),\end{aligned}$$

dependent on the observation D_n. for all $k \geq k_0$ where k_0 and κ_0 are constants introduced in Theorem 4.8 that could depend on the observation D_n.

5. Numerical Experiments

In this section, we conduct numerical experiments to justify our theoretical results.

Illustration in two dimensional space. First, we give an illustrative example in which the dynamics of the student network is depicted in a two dimensional setting $d = 2$. In this experiment, we employ $m = 2$ with $r_1^\circ = r_2^\circ = 1$ and $\theta_1^\circ = (1, 0)^\top$, $\theta_2^\circ = (0, 1)^\top$, $M = 15$, and $n = 100$. Figure 1 shows the optimization trajectory of $(a_{j,k}, w_{j,k})_{j=1}^M$. We can see that the nodes with initialization near to a teacher parameter approaches one of the nodes in the teacher network and, on the other hand, the nodes with initialization far away from any teacher node finally vanish. This behavior is induced by the sparse regularization, that is, the sparse regularization "selects" informative nodes and discard non-informative nodes. We also see that the selected nodes explore a wide area in the early stage and after that they finally head to the direction of one of the teacher nodes. This well justifies our theoretical analysis.

Effect of over-parameterization for convergence. Next, we investigate how the over-parameterization affects the dynamics. In this experiment, we employ $m = 5$ for the

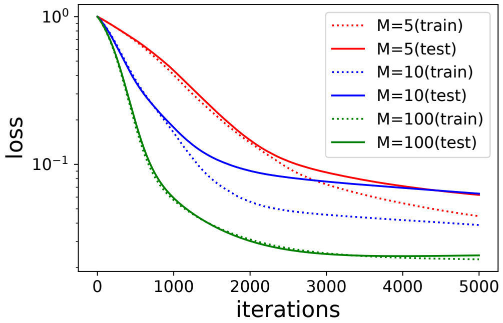

Figure 2. Convergence of the training/test loss for different student width $M = 5, 10, 100$.

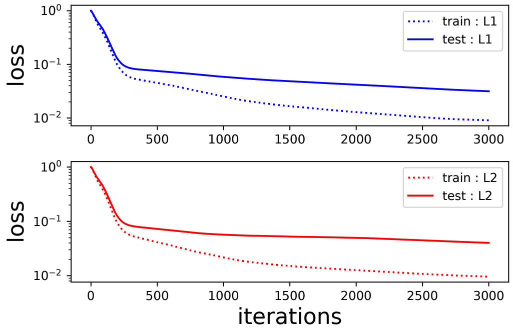

Figure 3. Comparison of L_1 and L_2- regularizations.

teacher width, $d = 5$ for the dimensionality and $n = 100$ for the sample size. As for the student network, we compare the dynamics between $M = 5, 10, 100$. Figure 2 depicts the training loss and test loss against the number of iterations. Each line corresponds to different setting of M. We can see that a sufficiently over-parameterized network ($M = 100$) appropriately estimates the true function while a narrow network ($M = 5$) does not reach the global optimal solution. We also note that the test loss is almost same as the training loss in the over-parameterized setting while we observe over-fitting for $M = 5$ and $M = 10$. This means that the solution in the over-parameterized setting ($M = 100$) finally converges to the optimal "sparse" solution that avoids the over-fitting. This is consistent to the findings by the existing studies (Safran & Shamir, 2018; Safran et al., 2020).

Comparison of L_1 and L_2 Regularization Inspired by Eq. (5), we also conduct norm-dependent gradient descent for the L_2-regularized problem:

$$\min_{\Theta \in (\mathbb{R} \times \mathbb{R}^d)^M} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i; \Theta))^2 + \frac{\lambda}{2} \sum_{j=1}^{M} (a_j^2 + \|w_j\|^2). \tag{13}$$

We give a comparison of the loss evolution between the L_1-regularization and L_2-regularization in Figure 3. In this experiment, we employ $m = 5$ for the teacher width, $d = 5$ for the dimensionality, $n = 100$ for the sample size and $M = 10$ for the student width. We can see that both regularizations show the almost same trajectory of the loss functions. This indicates the usefulness of the practical use of the L_2-regularization.

6. Conclusion

In this paper, we have investigated identifiability of the true target function via the gradient descent method for two-layer ReLU neural networks in teacher-student settings. We have shown that with the sparse regularization, the global minima can be arbitrarily close to the teacher network. Furthermore, we have proposed a gradient method with norm-dependent step size which is guaranteed to converge to the global minima, and shown that this framework can be applied to the teacher-student setting. The key ingredient in this analysis is the measure representation of the ReLU network. With this perspective, the gradient method can be associated with gradient descent in the measure space. We believe that this analysis gives a new insight into learnability in the teacher-student setting.

Acknowledgement

TS was partially supported by JSPS KAKENHI (18H03201, and 20H00576), Japan Digital Design and JST CREST.

References

Absil, P. A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *Proceedings of International Conference on Machine Learning*, pp. 242–252, 2019.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

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

Borwein, J. M. and Zhu, Q. J. *Techique of Variational Analysis*. Springer, 2005.

Bredies, K. and Pikkarainen, H. K. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

Cai, T. T., Fan, J., and Jiang, T. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 14:1837–1864, 2013.

Candès, E. J. and Fernandez-Granda, C. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

Chizat, L. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. volume 31, pp. 3036–3046, 2018.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

De Castro, Y. and Gamboa, F. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

de Dios, J. and Bruna, J. On sparsity in over-parametrised shallow ReLU networks. *arXiv preprint arXiv:2006.10225*, 2020.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations 7*, 2019.

Duval, V. and Peyré, G. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

E, W., Ma, C., and Wu, L. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

Erdogdu, M. A., Mackey, L., and Shamir, O. Global nonconvex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems 31*, pp. 9671–9680. 2018.

Flinth, A., de Gournay, F., and Weiss, P. On the linear convergence rates of exchange and continuous methods for total variation minimization. *Mathematical Programming*, pp. 1–37, 2020.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 315–323, 2011.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, pp. 6981–6991, 2019.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9482–9491, 2018.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 8580–8589, 2018.

Klusowski, J. M. and Barron, A. R. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pp. 6389–6399, 2018.

Li, Y. and Yuan, Y. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 597–607. Curran Associates, Inc., 2017.

Li, Y., Ma, T., and Zhang, H. R. Learning over-parametrized two-layer neural networks beyond NTK. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2613–2682. PMLR, 2020.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464, 2019.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pp. 1376–1401, 2015.

Nitanda, A. and Suzuki, T. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

Poon, C., Keriven, N., and Peyré, G. The geometry of off-the-grid compressed sensing. *arXiv preprint arXiv:1802.08464*, 2018.

Poon, C., Keriven, N., and Peyré, G. Support localization and the fisher metric for off-the-grid sparse regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1341–1350. PMLR, 2019.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Nonconvex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *arXiv e-prints*, pp. arXiv:1702.03849, 2017.

Rockafeller, R. T. Duality and stability in extremum problems involving convex functions. *Pacific Journal of Mathematics*, (1):167–188, 1967.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, pp. 4433–4441. PMLR, 2018.

Safran, I., Yehudai, G., and Shamir, O. The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. *arXiv preprint arXiv:2006.01005*, 2020.

Suzuki, T. and Akiyama, S. Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. In *International Conference on Learning Representations*, 2021.

Tian, Y. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

Tian, Y. Student specialization in deep rectified networks with finite width and input dimension. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9470–9480. PMLR, 2020.

Trillos, N. G. and Slepčev, D. On the rate of convergence of empirical measures in-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

Tropp, J. A. *An Introduction to Matrix Concentration Inequalities*, volume 8 of *Foundations and Trends in Machine Learning*. Now Publishers Inc. Hanover, MA, USA, 2015.

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

Weinan, E., Ma, C., and Wu, L. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, pp. 1–24, 2020.

Welling, M. and Teh, Y.-W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673. PMLR, 09–12 Jul 2020.

Yehudai, G. and Shamir, O. Learning a single neuron with gradient methods. *arXiv preprint arXiv:2001.05205*, 2020.

Zhang, X., Yu, Y., Wang, L., and Gu, Q. Learning one-hidden-layer relu networks via gradient descent. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1524–1534. PMLR, 2019.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

Zhou, M., Ge, R., and Jin, C. A local convergence theory for mildly over-parameterized two-layer neural network. *arXiv preprint arXiv:2102.02410*, 2021.