Overparameterized Nonlinear Optimization with Applications to Neural Nets

Samet OYMAK
Department of Electrical and Computer Engineering
University of California, Riverside
Riverside, CA, USA
oymak@ece.ucr.edu

***Abstract*—Occam's razor is a fundamental problem-solving principal and states that one should seek the simplest possible explanation. Indeed, classical machine learning models such as (sparse) linear regression aims to find simple explanations to data by using with as few parameters as possible. On the other hand, modern techniques such as deep networks are often trained in the overparameterized regime where the model size exceeds the size of the training dataset. While this increases the risk of overfitting and the complexity of the explanation, deep networks are known to have good generalization properties. In this talk, we take a step towards resolving this paradox: We show that solution found by first order methods, such as gradient descent, has the property that it has near shortest distance to the initialization of the algorithm among all other solutions. We also advocate that shortest distance property can be a good proxy for the simplest explanation. We discuss the implications of these results on neural net training and also highlight some outstanding challenges.**

I. INTRODUCTION

Suppose we are given a dataset of n input-output pairs $(\boldsymbol{x}_i, y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$. In order to explain the relation between inputs and outputs, we shall pick a function class f parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$, a (non-negative) loss function $\mathcal{L}$, and solve the empirical risk minimization

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(\boldsymbol{x}_i, \boldsymbol{\theta})\right). \tag{1}$$

Our technical discussion will mostly focus on the quadratic loss $\mathcal{L}(a,b) = (a-b)^2/2$.

Classical statistical learning theory postulates that to find a model that generalizes well and avoids overfitting, the size of the training data n should be more than the model size p. For instance, consider the case of linear regression where $f(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T\boldsymbol{\theta}$. In this case, there is no unique solution in the overparameterized regime $n < p$ as we wish to solve the system of n equations $(y_i = \boldsymbol{x}_i^T\boldsymbol{\theta})_{i=1}^n$ and there are p unknowns. We do remark that one can guarantee unique solution by incorporating priors on $\boldsymbol{\theta}$ such as sparsity, low-rank, and subspace constraints. Indeed, much of the compressed sensing and low-rank approximation literature aligns with this direction.

Contrary to the classical literature, popular machine learning models such as deep networks are often trained via first-order methods in the over-parameterized regime $n < p$. This regime, in theory, allows the model to perfectly (over)fit to the training data. Despite this, deep networks have surprisingly good generalization abilities i.e. they perform well on unseen test datasets. It is also not at all clear when and how they overfit to the training data as the associated learning problem is highly nonconvex. These questions bring new challenges to understand the fundamental aspects of state-of-the-art machine learning models. In this work, we will aim to shed light on some of these challenges with a focus on optimization and first order methods. Gradient descent algorithm aims to solve (1) by starting from some initial point $\boldsymbol{\theta}_0$ and running the iterative updates

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_\tau), \tag{2}$$

where η is the step size. Since deep networks are trained using variants of gradient descent, it is important to understand the properties of the solution found by gradient updates (2). In this talk, we will argue that, under certain deterministic assumptions, solution found by gradient descent achieves zero training loss (i.e. $\mathcal{L}(\boldsymbol{\theta}_\infty) = 0$) and has the property that it has near shortest distance to the initialization $\boldsymbol{\theta}_0$ among all other solutions. Mathematically speaking, the latter means

$$\|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_0\|_{\ell_2} \approx \inf_{\mathcal{L}(\boldsymbol{\theta})=0} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}.$$

We will also discuss implications of these findings on the optimization of neural networks. In particular, what are the fundamental limitations on the overfitting ability of the neural networks i.e. under what conditions a neural network can achieve zero-training loss in (1)? We will focus on one-hidden layer neural networks with k hidden units characterized by an activation ϕ, an input weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ and output weight vector $\boldsymbol{v} \in \mathbb{R}^k$ via

$$f(\boldsymbol{x}, (\boldsymbol{v}, \boldsymbol{W})) = \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}). \tag{3}$$

We will provide a discussion of recent optimization results on this topic most of which focus on the problem of training input weights $\boldsymbol{W}$. We will demonstrate the sub-optimality of these bounds by showing that one can train only the output weights $\boldsymbol{v}$ to achieve strictly better bounds. This will lead to our discussion on outstanding challenges in optimization and generalization of neural networks.

II. Related works

Implicit regularization: An interesting body of related works investigate the implicit regularization capabilities of (stochastic) gradient descent for separable classification problems including [13], [16]–[18], [23], [26]. These results show that gradient descent does not converge to an arbitrary solution, for instance, it has a tendency to converge to the solution with the max margin or minimal norm. Some of this literature apply to regression problems as well (such as low-rank regression).

Overparameterized neural networks: Several recent papers [3], [8], [9], [21], [25], [27], [28] study the benefits of overparameterization for training neural networks and related optimization problems. Very recent works [1], [2], [10], [11], [14], [29] show that overparameterized neural networks can fit the data with random initialization if the number of hidden nodes are polynomially large in the size of the dataset. Our discussion is inherently connected to these works. In particular, we will illustrate that existing results require *extreme* overparameterization i.e. they need network to be much larger than the dataset. We will also highlight the fundamental optimization principles behind these works by studying overparameterized learning in a generic setup. An equally important question is understanding the generalization capabilities of overparameterized models. This is the subject of a few interesting recent papers [4]–[7], [12], [15], [22]. We will provide a discussion of generalization in terms of the *shortest distance* property of gradient descent.

III. Overparameterized Optimization

As a prelude to understanding the key properties of gradient descent in over-parameterized nonlinear learning we begin by focusing on the simple case of linear regression. In this case the mapping in (1) takes the form $f(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{x}_i^T \boldsymbol{\theta}$. Gathering the input data $\boldsymbol{x}_i$ and labels y_i as rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a vector $\boldsymbol{y} \in \mathbb{R}^n$, the fitting problem amounts to minimizing the loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_{\ell_2}^2$. Denote the projection operator to null space of $\boldsymbol{X}$ by Π_{null} and the pseudo-inverse solution, when $n \le p$, by

$$\boldsymbol{\theta}^\dagger = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}.$$

It can be shown that, gradient descent with a reasonable step size converges to a unique solution

$$\boldsymbol{\theta}_\infty = \boldsymbol{\theta}^\dagger + \Pi_{\text{null}}(\boldsymbol{\theta}_0).$$

This solution, perhaps not surprisingly, is the closest one to $\boldsymbol{\theta}_0$. This follows from the fact that gradient updates always lie on the row space of $\boldsymbol{X}$ and never touches the null space. The distance to $\boldsymbol{\theta}_0$ is equal to the length of the pseudo-inverse $\boldsymbol{\theta}^\dagger$.

The natural question is how to move from least-squares to nonlinear problems such as neural net training. A key idea in the recent works [1], [2], [10], [11], [14], [19], [29] is based on replacing the data matrix $\boldsymbol{X}$ with a $n \times p$ *nonlinear feature matrix* obtained by the *Jacobian* of the problem. Jacobian is a function of data and model weights and given by the matrix of partial derivatives

$$\mathcal{J}(\boldsymbol{X}, \boldsymbol{\theta}) = \left[\frac{\partial f(\boldsymbol{x}_1,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \dots \quad \frac{\partial f(\boldsymbol{x}_n,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T$$

In contrast to neural net specific results which utilize properties of randomized initialization of $\boldsymbol{\theta}_0$ (e.g. random Gaussian weights), in [19], we provide three deterministic assumptions that govern when gradient descent attains zero loss for nonlinear least squares. Here, we provide a variation of the result of [19] which doesn't require smoothness of Jacobian.

Our first assumption is that, at the initial point $\boldsymbol{\theta}_0$ the minimum singular value of the Jacobian is lower bounded.

Assumption 1. *Fix a point $\boldsymbol{\theta}_0$. We have that $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \ge 2\alpha$ where $\sigma_{\min}()$ returns the minimum singular value.*

The second assumption is that, spectral norm (denoted by $\|\cdot\|$) of Jacobian is upper bounded by some quantity β.

Assumption 2. *For all $\boldsymbol{\theta} \in \mathbb{R}^p$, we have that $\|\mathcal{J}(\boldsymbol{\theta})\| \le \beta$.*

Our final assumption is there is some neighborhood of $\boldsymbol{\theta}_0$ around which Jacobian doesn't deviate much.

Assumption 3. *Fix a point $\boldsymbol{\theta}_0$ and a number $R > 0$. For any $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \le R$, we have that $\|\mathcal{J}(\boldsymbol{\theta}_0) - \mathcal{J}(\boldsymbol{\theta})\| \le \alpha/3$.*

These three properties essentially ensure that, even if the problem is nonlinear and nonconvex, it has a linear-regression-like behavior. The following theorem formalizes this.

Theorem 1. *Set $\mathcal{L}(x, y) = (x-y)^2/2$. Given $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, suppose Assumptions 1, 2, and 3 hold with $\beta \ge 2\alpha$ and*

$$R = \frac{5\|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}}{\alpha}.$$

Then, picking constant learning rate $\eta \le \frac{1}{\beta^2}$, all gradient iterations (2) *obey the followings*

$$\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le (1 - \frac{\eta\alpha^2}{4})^\tau \|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2} \tag{4}$$

$$\frac{\alpha}{5}\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le \|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}. \tag{5}$$

The line (4) shows that gradient descent converges linearly fast to achieve zero loss and setting $\eta = 1/\beta^2$, it finds an ε-approximate global minima in $\frac{\beta^2}{\alpha^2}\log(\frac{1}{\varepsilon})$ steps. The second line (5) is of particular importance. Ignoring the residual $\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2}$ term, (5) guarantees that, the model parameter $\boldsymbol{\theta}$ will never step out of the radius

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \le \frac{5}{\alpha}\|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}.$$

Hence gradient descent finds a solution within this radius. In [19], we complement this by showing that there is no global minimizer within the region $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \le \frac{1}{\beta}\|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}$. These upper and lower bounds can be shown to be sharp highlighting the fact that gradient descent finds a solution with near shortest distance to $\boldsymbol{\theta}_0$ among all global minima.

Implications for generalization: Ensuring that model performs well on unseen data is of critical importance. The generalization ability can be characterized in terms of the Rademacher complexity of the model space with respect to dataset $(\boldsymbol{x}_i, y_i)_{i=1}^n$. If the distance to initialization $\|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_0\|_{\ell_2}$ is guaranteed to be upper bounded by some Γ, then, we can

provide generalization guarantees for $\boldsymbol{\theta}_\infty$ as a function of the search space $\{\boldsymbol{\theta} \in \mathbb{R}^p \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \leq \Gamma\}$. In particular, as the sample size n increases, the generalization error (on fresh samples) will decay as $\mathcal{O}(\frac{\Gamma}{\sqrt{n}})$. There are interesting initial results in this direction for neural net training [1].

IV. Some Outstanding Challenges on Neural Nets

In this section, we focus on recent developments in overparameterized neural network training from a critical point of view. As discussed in the introduction, it is of interest to understand under what conditions neural nets achieve zero training error, in particular, networks with one-hidden layer as described by (3). Here, the challenge arise from the nonlinear activation ϕ and simultaneous optimization over input/output layers. Recent works [2], [10], [11], [20], [29] govern the overfitting ability in terms of number of hidden units k and focus on optimization over $\boldsymbol{W}$ (some of these works also apply to deep networks). Intuitively, $\boldsymbol{W}$ is the right parameter to focus as compared to $\boldsymbol{v}$ since it contains many more parameters. Input layer also learns and extracts the useful features from data and is critical for good generalization. Existing results require that (i) the number of hidden nodes k should be polynomially large in dataset size n and (ii) dataset satisfies certain separability conditions such as Assumption 1. Our recent work [20] improves over the results of [10], [11] and establishes the best known optimization bounds to find that

- $k \gtrsim \mathcal{O}(n^2/d)$ is sufficient for smooth activations ϕ,
- $k \gtrsim \mathcal{O}(n^4/d^3)$ is sufficient for ReLU activation $\phi(x) = \max(x, 0)$,

to ensure gradient descent on input layer achieves zero training error from a randomly initialized weight matrix $\boldsymbol{W}_0$. Unfortunately, even these improved results are fairly suboptimal compared to best possible bounds. Input layer has kd degrees of freedom and intuitively it should be able to fit the dataset as long as $n < \mathcal{O}(kd)$. This intuition is supported by empirical observations as well [27]. In contrast, bounds provided above require $n < \mathcal{O}(\sqrt{kd})$ and do not allow k to scale linearly in n.

Here, we shall further formalize this gap between theory and practice by considering the optimization over output layer (for fixed $\boldsymbol{W}$) and showing that, one can achieve zero training error by only optimizing $\boldsymbol{v}$ in the regime $k > \mathcal{O}(n)$. Note that output layer optimization is fairly straightforward. Letting ϕ apply entry-wise, set $\boldsymbol{\Phi} = \phi(\boldsymbol{W}\boldsymbol{X}^T) \in \mathbb{R}^{k \times n}$. $\boldsymbol{\Phi}$ represents the features generated by the network and optimal $\boldsymbol{v}$ is given by the pseudo-inverse

$$\boldsymbol{v}^\dagger = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{y}.$$

Pseudo-inverse will clearly achieve zero training error if $\boldsymbol{\Phi}$ is full-rank and gradient descent will converge to pseudo-inverse solution with sufficiently small step size. Hence the key question we wish to address is: under what conditions $\boldsymbol{\Phi}$ is full-rank? Towards this goal, we introduce a variation of Assumption 1 to characterize the optimization landscape around $\boldsymbol{v}$ when $\boldsymbol{W}$ is randomly initialized and fixed.

Assumption 4. *Let $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. For some constant $\alpha > 0$ (as a function of input dataset $\boldsymbol{X}$), suppose the covariance matrix obeys*

$$\mathbb{E}[\phi(\boldsymbol{X}\boldsymbol{g})\phi(\boldsymbol{X}\boldsymbol{g})^T] \geq \alpha^2.$$

Here, $\boldsymbol{g}$ is intended to correspond to a single row of the randomly initialized matrix $\boldsymbol{W}$. Under Assumption 4, for $\boldsymbol{W} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, we have that

$$\frac{1}{k}\mathbb{E}[\boldsymbol{\Phi}^T\boldsymbol{\Phi}] \geq \alpha^2.$$

Building on this observation following theorem provides high probability lower bound on $\sigma_{\min}(\boldsymbol{\Phi})$.

Theorem 2. *Suppose input samples $(\boldsymbol{x}_i)_{i=1}^n$ have unit Euclidian norm and input weight matrix obeys $\boldsymbol{W} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Additionally, suppose Assumption 4 holds, and ϕ is the ReLU activation ($\phi(x) = \max(x, 0)$). If*

$$k \geq \mathcal{O}\left(\frac{n\log(n)\log(k)}{\alpha^2}\right),$$

with probability $1 - n^{-100} - k^{-100}$, $\boldsymbol{\Phi} = \phi(\boldsymbol{W}\boldsymbol{X}^T)$ is full-rank and obeys

$$\frac{\sigma_{\min}(\boldsymbol{\Phi})}{\sqrt{k}} \geq \frac{\alpha}{2}.$$

Proof. The proof is based on standard concentration arguments involving subgaussian distributions and Matrix Chernoff bound [24]. First, denoting ith row of $\boldsymbol{W}$ by $\boldsymbol{w}_i$ and setting $\boldsymbol{z}_i = \boldsymbol{X}\boldsymbol{w}_i$, observe that $\boldsymbol{z}_i$ and $\phi(\boldsymbol{z}_i)$ are $\|\boldsymbol{X}\|$-Lipschitz function of $\boldsymbol{w}_i$. Following this, using a union bound and Gaussian concentration and noticing $\mathbb{E}[\|\boldsymbol{z}_i\|_{\ell_2}] \leq \|\boldsymbol{X}\|_F = \sqrt{n}$, with probability $1 - k\exp(-\mathcal{O}(\delta^2\frac{n}{\|\boldsymbol{X}\|^2}))$, we have that

$$\|\phi(\boldsymbol{z}_i)\|_{\ell_2} \leq \|\boldsymbol{z}_i\|_{\ell_2} \leq \delta\sqrt{n} \quad \text{for all} \quad 1 \leq i \leq k, \tag{6}$$

On this (truncation) event (6), $\boldsymbol{z}_i$'s are still statistically i.i.d. and it can be shown that covariance does not change much i.e. $\mathbb{E}[\phi(\boldsymbol{z}_i)\phi(\boldsymbol{z}_i)^T] \geq \alpha^2/2$ (by using Assumption 4, subgaussian tail, and picking large δ to ensure high probability). To proceed, applying Matrix Chernoff with ℓ_2 norm bounds of (6) and covariance lower bounds of $\alpha^2/2$ yields that,

$$\mathbb{P}(\frac{1}{k}\sum_{i=1}^k \phi(\boldsymbol{z}_i)\phi(\boldsymbol{z}_i)^T \geq \frac{\alpha^2}{4}) \geq 1 - n\exp(-\mathcal{O}(\frac{k\alpha^2}{\delta^2 n})).$$

We conclude by union bounding the above probability with that of (6) after setting $\delta = \mathcal{O}(\sqrt{\log k})$ and noticing that k is assumed to be large enough to ensure high probability. □

Discussion: Ignoring log factors and assuming constant α[1], this result yields the aforementioned result $k \gtrsim \mathcal{O}(n)$ to guarantee zero training error by optimizing output layer alone. We obviously acknowledge that optimizing over $\boldsymbol{v}$ alone is much easier than optimizing over $\boldsymbol{W}$ or all layers of a deep network. Indeed, earlier works involve fairly complicated arguments and utilize deeper ideas (e.g. Theorem 1). However, it is somewhat

[1] Similar assumptions are made by related works [2], [10], [11], [20].

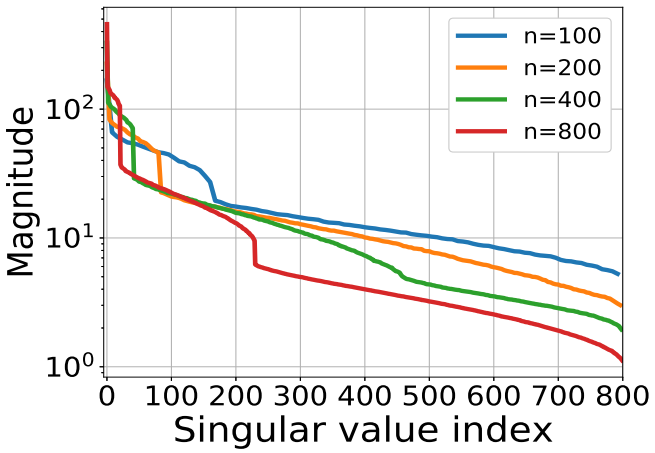

Fig. 1: Singular value spectrum of $\mathbf{\Phi} = \phi(\boldsymbol{W}\boldsymbol{X}^T)$ for $k = 1600$ and $n = 100$ to 800. Smaller n plots are stretched to fit $[0, 800]$ interval.

surprising that such a simple strategy can easily achieve better bounds than what we have for input layer.

To demonstrate that Theorem 2 is indeed sensible, in Figure 1, we plotted the singular value distribution of the $\mathbf{\Phi}$ matrix for varying sample sizes $n = 100, 200, 400, 800$ and fixed input dimension $d = 20$ and hidden units $k = 1600$. In this plot, $n = 100, 200, 400$ is properly stretched to be aligned with $n = 800$. Here, we generated $\boldsymbol{W}$ and $\boldsymbol{X}$ with i.i.d. $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 1/d)$ entries respectively. While $\mathbf{\Phi}$ has a bimodal spectrum with some large and some small singular values, the minimum singular value is persistently nonzero despite n and k being much larger than d. However, we still lack a good understanding of the evolution of this spectrum in terms of (n, k, d): for instance, larger n appears to amplify the larger singular values and dampens the smaller ones (compare red and blue curves).

While discussion so far focused on networks with one-hidden layers, another outstanding question is on the role of depth. Interestingly, depth appears to hurt the performance in very recent results [2], [10] (i.e. they require larger network to fit the data if there are multiple layers). We believe, this is more than likely due to the sub-optimality of the analysis. An improved understanding of when depth helps/hurts optimization and generalization can further shed light on the success of deep learning models.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[3] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*, 2018.

[4] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. 02 2018.

[5] Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. 06 2017.

[6] Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. 06 2018.

[7] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? 06 2018.

[8] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[9] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

[10] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[11] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[12] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. 12 2017.

[13] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[14] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *NeurIPS*, 2018.

[15] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. 08 2018.

[16] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *arXiv preprint arXiv:1712.06559*, 2017.

[17] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796*, 2018.

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

[19] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? *arXiv preprint arXiv:1812.10004*, 2018.

[20] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training neural networks. 2019.

[21] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

[22] Mei Song, A Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. In *Proceedings of the National Academy of Sciences*, volume 115, pages E7665–E7671, 2018.

[23] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

[24] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[25] L. Venturi, A. Bandeira, and J. Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*, 2018.

[26] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

[27] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[28] Zhihui Zhu, Daniel Soudry, Yonina C. Eldar, and Michael B. Wakin. The global optimization geometry of shallow linear neural networks. 05 2018.

[29] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.