Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?

Samet Oymak [1] **Mahdi Soltanolkotabi** [2]

Abstract

Many modern learning tasks involve fitting nonlinear models which are trained in an overparameterized regime where the parameters of the model exceed the size of the training dataset. Due to this overparameterization, the training loss may have infinitely many global minima and it is critical to understand the properties of the solutions found by first-order optimization schemes such as (stochastic) gradient descent starting from different initializations. In this paper we demonstrate that when the loss has certain properties over a minimally small neighborhood of the initial point, first order methods such as (stochastic) gradient descent have a few intriguing properties: (1) the iterates converge at a geometric rate to a global optima even when the loss is nonconvex, (2) among all global optima of the loss the iterates converge to one with a near minimal distance to the initial point, (3) the iterates take a near direct route from the initial point to this global optimum. As part of our proof technique, we introduce a new potential function which captures the tradeoff between the loss function and the distance to the initial point as the iterations progress. The utility of our general theory is demonstrated for a variety of problem domains spanning low-rank matrix recovery to shallow neural network training.

1. Introduction

1.1. Motivation

In a typical statistical estimation or supervised learning problem, we are interested in fitting a function $f(\cdot;\boldsymbol{\theta}) : \mathbb{R}^d \mapsto \mathbb{R}$ parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$ to a training data set of n input-output pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$. The training problem then consists of finding a parameter $\boldsymbol{\theta}$ that minimizes the empirical risk $\frac{1}{n}\sum_{i=1}^n \ell(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)$. The loss $\ell(\tilde{y}, y)$ measures the discrepancy between the output(or label) y and the model prediction $\tilde{y} = f(\boldsymbol{x}_i;\boldsymbol{\theta})$. For regression tasks one typically uses a least-squares loss $\ell(\tilde{y}, y) = \frac{1}{2}(\tilde{y} - y)^2$ so that the training problem reduces to a nonlinear least-squares problem of the form

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\sum_{i=1}^n \left(f(\boldsymbol{x}_i;\boldsymbol{\theta}) - y_i\right)^2. \tag{1.1}$$

In this paper we mostly focus on nonlinear least-squares problems however Section 5 of the supplementary material extends our results to a broader class of loss functions $\mathcal{L}(\boldsymbol{\theta})$.

Classical statistical estimation/learning theory postulates that to find a reliable model that avoids overfitting, the size of the training data must exceed the intrinsic dimension[1] of the model class $f(\cdot;\boldsymbol{\theta})$ used for empirical risk minimization (1.1). For many models such notions of intrinsic dimension are at least as large as the number of parameters in the model p, so that this literature requires the size of the training data to exceed the number of parameters in the model i.e. $n > p$. Contrary to this classical literature, modern machine learning models such as deep neural networks are often trained via first-order methods in an over-parameterized regime where the number of parameters in the model exceed the size of the training data (i.e. $n < p$). Statistical learning in this over-parameterized regime poses new challenges: Given the nonconvex nature of the training loss (1.1) can first-order methods converge to a globally optimal model that perfectly interpolate the training data? If so, which of the global optima do they converge to? What are the statistical properties of this model and how does this model vary as a function of the initial parameter used to start the iterative updates? What is the trajectory that iterative methods such as (stochastic) gradient descent take to reach this point? Why does a model trained using this approach *generalize* to

[1]Department of Electrical and Computer Engineering, University of California, Riverside [2]Department of Electrical and Computer Engineering, University of Southern California. Correspondence to: Samet Oymak <sametoymak@gmail.com>, Mahdi Soltanolkotabi <soltanol@usc.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]Some common notions of intrinsic dimension include Vapnik–Chervonenkis (VC) Dimension (Vapnik & Chervonenkis, 2015), Rademacher/Gaussian complexity (Bartlett & Mendelson, 2002; Mohri et al., 2018; Talagrand, 2006), as well as naive parameter counting.

new data and avoid overfitting to the training data?

In this paper we take a step towards addressing such challenges. We demonstrate that in many cases first-order methods do indeed converge to a globally optimal model that perfectly fits the training data. Furthermore, we show that among all globally optimal parameters of the training loss these algorithms tend to converge to one which has a near minimal distance to the parameter used for initialization. Additionally, the path that these algorithms take to reach such a global optima is rather short, with these algorithms following a near direct trajectory from initialization to this global optimum. We believe these key features may help demystify why models trained using first-order methods can achieve reliable learning in modern over-parametrized regimes without over-fitting to the training data.

1.2. Contributions

Our main contributions can be summarized as follows:

- We provide a general convergence result for overparameterized learning via gradient descent, that comes with matching upper and lower bounds, showing that under appropriate assumptions over a small neighborhood of the initialization, gradient descent (1) finds a globally optimal model, (2) among all possible globally optimal parameters it finds one which is approximately the closest to initialization and (3) it follows a nearly direct trajectory to find this global optima.
- We show that SGD exhibits the same behavior sand converges linearly without ever leaving a small neighborhood of the initialization even with rather large learning rates.
- We demonstrate the utility of our general results in the context of three overparameterized learning problems: generalized linear models, low-rank matrix regression, and shallow neural network training.

2. Convergence Analysis for Gradient Descent

The nonlinear least-squares problem in (1.1) can be written in the more compact form

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\left\|f(\boldsymbol{\theta})-\boldsymbol{y}\right\|_{\ell_2}^2, \tag{2.1}$$

where $\boldsymbol{y} := \begin{bmatrix}\boldsymbol{y}_1 & \boldsymbol{y}_2 & \dots & \boldsymbol{y}_n\end{bmatrix}^T \in \mathbb{R}^n$ and $f(\boldsymbol{\theta}) := \begin{bmatrix}f(\boldsymbol{x}_1;\boldsymbol{\theta}) & f(\boldsymbol{x}_2;\boldsymbol{\theta}) & \dots & f(\boldsymbol{x}_n;\boldsymbol{\theta})\end{bmatrix}^T \in \mathbb{R}^n$. A natural approach to optimizing (2.1) is to use gradient descent updates of the form

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

starting from some initial parameter $\boldsymbol{\theta}_0$. For the formulation (2.1) above the gradient takes the form

$$\nabla\mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})^T\left(f(\boldsymbol{\theta})-\boldsymbol{y}\right). \tag{2.2}$$

Here, $\mathcal{J}(\boldsymbol{\theta}) \in \mathbb{R}^{n\times p}$ is the Jacobian matrix associated with the mapping $f(\boldsymbol{\theta})$ with entries given by $\mathcal{J}_{ij} = \frac{\partial f(\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}$. We note that in the over-parameterized regime ($n < p$), the Jacobian has more columns than rows. Throughout, $\sigma_{\min}(\cdot)/\|\cdot\|$ denote the minimum/maximum singular value.

Our first assumption ensures that the Jacobian matrix smoothly changes as a function of the parameter $\boldsymbol{\theta}$.

Assumption 1 (Jacobian smoothness) *Consider a set $\mathcal{D} \subset \mathbb{R}^p$ containing the initial point $\boldsymbol{\theta}_0$ (i.e. $\boldsymbol{\theta}_0 \in \mathcal{D}$). We assume that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{D}$,*

$$\left\|\mathcal{J}(\boldsymbol{\theta}_2)-\mathcal{J}(\boldsymbol{\theta}_1)\right\| \le L\left\|\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\right\|_{\ell_2}.$$[2]

We will also assume that the spectrum of the Jacobian is bounded in a local neighborhood of the initialization.

Assumption 2 (Jacobian Spectrum) *Consider a set $\mathcal{D} \subset \mathbb{R}^p$ containing the initial point $\boldsymbol{\theta}_0$ (i.e. $\boldsymbol{\theta}_0 \in \mathcal{D}$). We assume that for all $\boldsymbol{\theta} \in \mathcal{D}$ the following inequality holds*

$$\alpha \le \sigma_{\min}\left(\mathcal{J}(\boldsymbol{\theta})\right) \le \left\|\mathcal{J}(\boldsymbol{\theta})\right\| \le \beta,$$

with α, β scalars obeying $\beta \ge \alpha > 0$.

With these assumptions in place we are now ready to state our main result.

Theorem 2.1 *Consider a nonlinear least-squares optimization problem of the form (2.1). Suppose the Jacobian mapping associated with f obeys Assumption 2 over a ball $\mathcal{D}$ of radius $R := \frac{4\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ around a point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$.*[3] *Furthermore, suppose Assumption 1 holds over $\mathcal{D}$ and set $\eta \le \frac{1}{2\beta^2}\cdot\min\left(1, \frac{\alpha^2}{L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}\right)$. Then, running gradient descent updates of the form $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ starting from $\boldsymbol{\theta}_0$, all iterates obey.*

$$\left\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\right\|_{\ell_2}^2 \le \left(1-\frac{\eta\alpha^2}{2}\right)^\tau \left\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\right\|_{\ell_2}^2, \tag{2.3}$$

$$\frac{1}{4}\alpha\left\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\right\|_{\ell_2} + \left\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\right\|_{\ell_2} \le \left\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\right\|_{\ell_2}. \tag{2.4}$$

Furthermore, the total gradient path is bounded. That is,

$$\sum_{\tau=0}^{\infty}\left\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\right\|_{\ell_2} \le \frac{4\left\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\right\|_{\ell_2}}{\alpha}. \tag{2.5}$$

To apply our main result, one can simply verify that Jacobian is nice at the initial point. The following corollary highlights the key relations between smoothness, residual, and initial Jacobian for global convergence.

[2] Note that, if $\frac{\partial\mathcal{J}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ is continuous, Lipschitzness condition holds over any compact domain (for possibly large L).

[3] That is, $\mathcal{D} = \mathcal{B}\left(\boldsymbol{\theta}_0, \frac{4\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}\right)$ with $\mathcal{B}(\boldsymbol{c},r) = \left\{\boldsymbol{\theta}\in\mathbb{R}^p : \|\boldsymbol{\theta}-\boldsymbol{c}\|_{\ell_2} \le r\right\}$

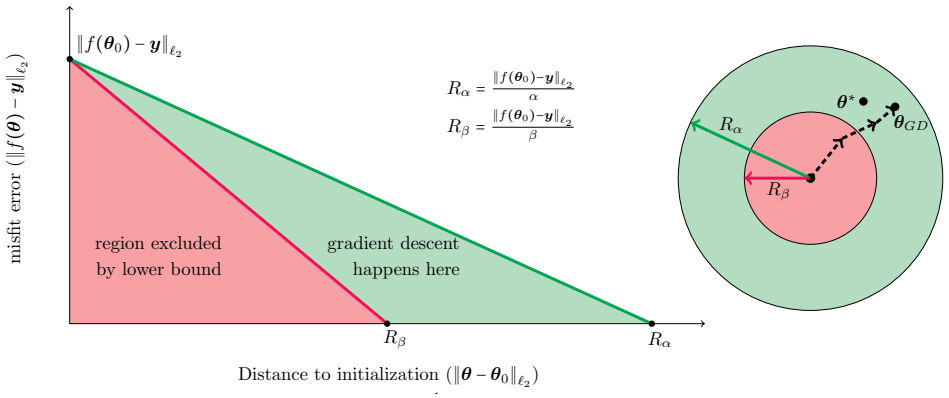


Figure 1. In the left figure we show that the gradient descent iterates in over-parameterized learning exhibit a sharp tradeoff between distance to the initial point ($\|\theta-\theta_0\|_{\ell_2}$) and the misfit error ($\|f(\theta)-y\|_{\ell_2}$). Our upper (equation (2.10)) and lower bounds (Theorem 2.4) guarantee that the gradient descent iterates must lie in the green region. Additionally this is the tightest region as we provide examples in Theorem 2.4 where gradient descent occurs only on the upper bound (green) line or on the lower bound (red line). Right figure shows the same behavior in the parameter space. Our theorems predict that the gradient descent trajectory ends at a globally optimal point θ_{GD} in the green region and this point will have approximately the same distance to the initialization parameter as the closest global optima to the initialization ($\theta^\star$). Furthermore, the GD iterates follow a near direct route from the initialization to this global optima.

Theorem 2.2 *Suppose the Jacobian at θ_0 obeys*

$$2\alpha \le \sigma_{\min}\left(\mathcal{J}(\theta_0)\right) \le \|\mathcal{J}(\theta_0)\| \le \beta/2.$$

Additionally, suppose Assumption 1 holds over a ball of radius $R = \frac{4\|f(\theta_0)-y\|_{\ell_2}}{\alpha}$ around θ_0 and

$$\alpha^2 \ge 4L\|y-f(\theta_0)\|_{\ell_2}. \tag{2.6}$$

Then, the conclusions of Theorem 2.1 hold with $\eta \le \frac{1}{2\beta^2}$.

Another trivial consequence of our theorem is the following.

Corollary 2.3 *Consider the setting and assumptions of Theorem 2.1 above. Let θ^* denote the global optima of the loss $\mathcal{L}(\theta)$ with smallest Euclidean distance to the initial parameter θ_0. Then, the gradient descent iterates θ_τ obey*

$$\|\theta_\tau-\theta_0\|_{\ell_2} \le 4\frac{\beta}{\alpha}\|\theta^*-\theta_0\|_{\ell_2}, \tag{2.7}$$

$$\sum_{\tau=0}^{\infty}\|\theta_{\tau+1}-\theta_\tau\|_{\ell_2} \le 4\frac{\beta}{\alpha}\|\theta^*-\theta_0\|_{\ell_2}. \tag{2.8}$$

The theorems and corollary above show that if the Jacobian of the nonlinear mapping has bounded/smooth deviations (Assumptions 1) and is well-conditioned (Assumption 2) in a ball of radius R around the initial point, then gradient descent enjoys three intriguing properties.

Zero traning error: The first property demonstrated by Theorem 2.1 above is that the iterates converge to a global optima θ_{GD}. This hold despite the fact that the fitting problem may be highly nonconvex in general. Indeed, based on (2.3) the fitting/training error $\|f(\theta_\tau)-y\|_{\ell_2}$ achieved by Gradient Descent (GD) iterates converges to zero. Therefore, GD can perfectly interpolate the data and achieve zero training error. Furthermore, this convergence is rather fast and the algorithm enjoys a geometric (a.k.a. linear) rate of convergence to this global optima.

Gradient descent iterates remain close to the initialization: The second interesting aspect of these results is that they guarantee the GD iterates never leave a neighborhood of radius $\frac{4}{\alpha}\|f(\theta_0)-y\|_{\ell_2}$ around the initial point. That is the GD iterates remain rather close to the initialization. In fact, based on (2.7) we can conclude that

$$\|\theta_{GD}-\theta_0\|_{\ell_2} = \lim_{\tau\to\infty}\|\theta_\tau-\theta_0\|_{\ell_2} \le 4\frac{\beta}{\alpha}\|\theta^*-\theta_0\|_{\ell_2}.$$

Thus the distance between the global optima GD converges to and the initial parameter θ_0 is within a factor $4\frac{\beta}{\alpha}$ of the distance between the closest global optima to θ_0 and the initialization. This shows that among all global optima of the loss, the GD iterates converge to one with a near minimal distance to the initialization. In particular, (2.4) shows that for all iterates the weighted sum of the distance to the initialization and the misfit error remains bounded so that as the loss decreases the distance to the initialization only moderately increases.

Gradient descent follows a short path: Another interesting aspect of the above results is that the total length of the path taken by gradient descent remains bounded. Indeed, based on (2.8) the length of the path taken by GD is within a factor of the distance between the closest global optima and the initialization. This implies that GD follows a near direct route from the initialization to a global optima!

We would like to note that Theorem 2.1 and Corollary 2.3 are special instances of a more general result stated in Theorem 9.3 of the supplementary material. This more

general result requires Assumptions 1 and 2 to hold in a smaller neighborhood and improves the approximation ratios. Specifically, the radius R can be chosen as small as

$$\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}, \tag{2.9}$$

and (2.4) can be improved to

$$\alpha\,\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}+\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2}\leq\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2} \tag{2.10}$$

This improves th approximation ratios in Corollary 2.3 to

$$\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}\leq\frac{\beta}{\alpha}\,\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}\,, \tag{2.11}$$

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2}\leq\frac{\beta}{\alpha}\,\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}\,. \tag{2.12}$$

The role of the sample size: Theorem 2.1 provides a good intuition towards the role of sample size in the overparameterized optimization landscape. First, observe that adding more samples can only increase the condition number of the Jacobian matrix (larger β and smaller α). Second, assuming samples are i.i.d, the initial misfit $\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}$ is proportional to $\sqrt{n}$. Together these imply that more samples lead to a more challenging optimization problem as: (1) More samples leads to a slower convergence rate by degrading the condition number of the Jacobian. (2) The required convergence radius R increases proportional to $\sqrt{n}$ and we need Jacobian to be well-behaved over a larger neighborhood.

A natural question about the results discussed so far is whether the size of the local neighborhood for which we require our assumptions to hold is optimal. In particular, one may hope to be able to show that a significantly smaller neighborhood is sufficient. We now state a lower bound showing that this is not possible.

Theorem 2.4 *Consider a nonlinear least-squares optimization problem of the form (2.1) and assume Assumption 1 holds over a set $\mathcal{D}$ around a point $\boldsymbol{\theta}_0\in\mathbb{R}^p$. Then,*

$$\|\boldsymbol{y}-f(\boldsymbol{\theta})\|_{\ell_2}+\beta\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\geq\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}, \tag{2.13}$$

holds for all $\boldsymbol{\theta}\in\mathcal{D}$. Hence, any $\boldsymbol{\theta}$ that sets the loss to zero satisfies $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\geq\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}/\beta$. Furthermore, for any α and β obeying $\alpha,\beta\geq 0$ and $\beta\geq\alpha$, there exists a linear regression problem such that

$$\|\boldsymbol{y}-f(\boldsymbol{\theta})\|_{\ell_2}+\alpha\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\geq\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}, \tag{2.14}$$

holds for all $\boldsymbol{\theta}$. Also, for any α and β obeying $\alpha,\beta\geq 0$ and $\beta\geq\alpha$, there also exists a linear regression problem where running gradient descent updates of the form $\boldsymbol{\theta}_{\tau+1}=\boldsymbol{\theta}_\tau-\eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ starting from $\boldsymbol{\theta}_0=0$ with a sufficiently small learning rate η, all iterates $\boldsymbol{\theta}_\tau$ obey

$$\|\boldsymbol{y}-f(\boldsymbol{\theta}_\tau)\|_{\ell_2}+\beta\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}=\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}. \tag{2.15}$$

The result above shows that any global optima is at least a distance $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\geq\frac{\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}}{\beta}$ away from the initialization so that the minimum ball around the initial point needs to have radius at least $R\geq\frac{\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}}{\beta}$ for convergence to a global optima to occur. Comparing this lower-bound with that of Theorem 2.1 and in particular the improvement discussed in (2.9) suggests that the size of the local neighborhood is optimal up to a factor β/α which is the condition number of the Jacobian in the local neighborhood. More generally, this result shows that the weighted sum of the residual/misfit to the model ($\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}$) and distance to initialization ($\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}$) has nearly matching lower/upper bounds (compare (2.10) and (2.13)). Theorem 2.4 also provides two specific examples in the context of linear regression which shows that both of these upper and lower bounds are possible under our assumptions.

Collectively our theorems (Theorem 2.1, Corollary 2.3, improvements in equations (2.9) and (2.10), and Theorem 2.4) demonstrate that the path taken by gradient descent is by no means arbitrary. Indeed as depicted in the left picture of Figure 1, gradient descent iterates in over-parameterized learning exhibit a sharp tradeoff between distance to the initial point ($\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}$) and the misfit error ($\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}$). Our upper (equation (2.10)) and lower bounds (Theorem 2.4) guarantee that the gradient descent iterates must lie in the green region in this figure. Additionally this is the tightest region as we provide examples in Theorem 2.4 where gradient descent occurs only on the upper bound (green) line or on the lower bound (red line). In the right picture of Figure 1 we also depict the gradient descent trajectory in the parameter space. As shown, the GD iterates end at a globally optimal point $\boldsymbol{\theta}_{GD}$ in the green region and this point will have approximately the same distance to the initialization parameter as the closest global optima to the initialization ($\boldsymbol{\theta}^*$). Furthermore, the GD iterates follow a near direct route from the initialization to this global optima.

3. Convergence Analysis for Stochastic Gradient Descent

Arguably the most widely used algorithm in modern learning is Stochastic Gradient Descent (SGD). For optimizing nonlinear least-squares problems (2.1) a natural implementation of SGD is to sample a data point at random and use that data point for the gradient updates. Specifically, let $\{\gamma_\tau\}_{\tau=0}^{\infty}$ be an i.i.d. sequence of integers chosen uniformly from $\{1,2,\ldots,n\}$. The SGD iterates take the form

$$\boldsymbol{\theta}_{\tau+1}=\boldsymbol{\theta}_\tau-\eta(f(\boldsymbol{x}_{\gamma_\tau};\boldsymbol{\theta}_\tau)-y_{\gamma_\tau})\nabla f(\boldsymbol{x}_{\gamma_\tau};\boldsymbol{\theta}_\tau). \tag{3.1}$$

We are interested in understanding the trajectory of SGD for over-parameterized learning. In particular, whether the intriguing properties of GD continue to hold for SGD. Our next theorem addresses this challenge.

Theorem 3.1 *Consider a nonlinear least-squares optimization problem of the form* $\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2}\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}^2$*, with* $f:\mathbb{R}^p\mapsto\mathbb{R}^n$ *and* $\boldsymbol{y}\in\mathbb{R}^n$*. Suppose the Jacobian mapping associated with* f *obeys Assumption 2 over a ball* $\mathcal{D}$ *of radius* $R := \nu\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ *around a point* $\boldsymbol{\theta}_0\in\mathbb{R}^p$ *with* ν *a scalar obeying* $\nu\geq 3$*. Also assume the rows of the Jacobian have bounded Euclidean norm over this ball, that is*

$$\max_i \|\mathcal{J}_i(\boldsymbol{\theta})\|_{\ell_2}\leq B \quad \textit{for all} \quad \boldsymbol{\theta}\in\mathcal{D}.$$

Furthermore, suppose Assumption 1 holds over $\mathcal{D}$ *and set* $\eta\leq\frac{\alpha^2}{\nu\beta^2B^2+\nu\beta BL\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}$*. Then, there exists an event* E *with* $\mathbb{P}(E)\geq 1-\frac{4}{\nu}\left(\frac{\beta}{\alpha}\right)^{\frac{1}{p}}$ *such that running SGD updates of the form (3.1) starting from* $\boldsymbol{\theta}_0$*, all iterates obey*

$$\mathbb{E}\left[\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2}^2\,\mathbb{1}_E\right]\leq\left(1-\frac{\eta\alpha^2}{2n}\right)^\tau\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}^2.$$

Furthermore, on this event the SGD iterates never leave the local neighborhood $\mathcal{D}$*.*

This result shows that SGD converges to a global optima that is close to the initialization. Furthermore, SGD always remains in close proximity to the initialization with high probability. Specifically, the neighborhood is on the order of $\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ which is consistent with the results on gradient descent and the lower bounds. However, unlike for gradient descent our approach to proving such a result is not based on the potential function (2.4). Rather we introduce a new potential function that keeps track of the average distances to multiple points around the initialization $\boldsymbol{\theta}_0$.

One interesting aspect of the result above is that the learning rate used is rather large. Indeed, ignoring an β/α ratio our convergence rate is on the order of $1-c/n$ so that n iterations of SGD correspond to a constant decrease in the misfit error on par with a full gradient iteration. This is made possible by a novel martingale-based technique which is in part inspired by (Tan & Vershynin, 2017) which studies SGD for nonconvex phase retrieval. Our novelty is analyzing SGD *without knowing where it eventually converges* by utilizing our potential function and ensuring that SGD iterations never exit the local neighborhood.

We note that it is possible to also used Azuma's inequality applied to the sequence $\log\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2}$ to show that the SGD iterates stay in a local neighborhood with very high probability. This idea has been utilized by recent related works (Allen-Zhu et al., 2018b; Li & Liang, 2018). However, such an argument requires a very small learning rate to ensure that one can take many steps without leaving the neighborhood at which point the concentration effect of Azuma becomes applicable. In contrast, our proof allows for using aggressive learning rates (on par with gradient descent) without ever leaving the local neighborhood.

4. Case studies

In this section we specialize and further develop our general convergence analysis in the context of three fundamental problems: fitting a generalized linear model, low-rank regression, and shallow neural network training.

4.1. Learning generalized linear models

In this section we focus on learning Generalized Linear Models (GLM) from data which involves fitting functions of the form $f(\cdot;\boldsymbol{\theta}):\mathbb{R}^d\to\mathbb{R}$

$$f(\boldsymbol{x};\boldsymbol{\theta})=\phi(\langle\boldsymbol{x},\boldsymbol{\theta}\rangle).$$

A natural approach for fitting such GLMs is via minimizing the nonlinear least-squares misfit of the form

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\mathcal{L}(\boldsymbol{\theta}):=\frac{1}{2}\sum_{i=1}^n\left(\phi(\langle\boldsymbol{x}_i,\boldsymbol{\theta}\rangle)-y_i\right)^2. \tag{4.1}$$

Define the data matrix $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ with rows given by $\boldsymbol{x}_i$ for $i=1,2,\ldots,n$. We thus recognize the above fitting problem as a special instance of (2.1) with $f(\boldsymbol{\theta})=\phi(\boldsymbol{X}\boldsymbol{\theta})$. Here, ϕ when applied to a vector means applying the nonlinearity entry by entry. We wish to understand the behavior of GD in the over-parameterized regime where $n\leq p$. This is the subject of the next two theorems.

Theorem 4.1 (Overparameterized GLM) *Consider a GLM fitting problem of the form (4.1) with* $\phi:\mathbb{R}\to\mathbb{R}$ *a strictly increasing nonlinearity with continuous derivatives obeying* $0<\gamma\leq\phi'(z)\leq\Gamma$ *for all* z*. Starting from arbitrary* $\boldsymbol{\theta}_0$*, we run gradient descent on the loss (4.1) with* $\eta\leq\frac{1}{\|\boldsymbol{X}\|^2\Gamma^2}$*. Furthermore, let* $\boldsymbol{\theta}^\star$ *denote the closest global optimum to* $\boldsymbol{\theta}_0$*. Then, all GD iterates obey*

$$\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}^\star\|_{\ell_2}\leq\left(1-\eta\gamma^2\lambda_{\min}\left(\boldsymbol{X}\boldsymbol{X}^T\right)\right)^\tau\|\boldsymbol{\theta}_0-\boldsymbol{\theta}^\star\|_{\ell_2}. \tag{4.2}$$

This theorem demonstrates that when fitting GLMs in the over-parameterized regime, gradient descent converges at a linear to a globally optimal model. Furthermore, this convergence is to the closest global optimum to the initialization.

4.2. Low-rank regression

A variety of modern learning problems spanning recommender engines to controls involve fitting low-rank models to data. In this problem given a data set of size n consisting of input/features $\boldsymbol{X}_i\in\mathbb{R}^{d\times d}$ and labels $y_i\in\mathbb{R}$ for $i=1,2,\ldots,n$, we aim to fit nonlinear models of the form

$$\boldsymbol{X}\mapsto f(\boldsymbol{X};\boldsymbol{\Theta})=\langle\boldsymbol{X},\boldsymbol{\Theta}\boldsymbol{\Theta}^T\rangle=\text{trace}\left(\boldsymbol{\Theta}^T\boldsymbol{X}\boldsymbol{\Theta}\right),$$

with $\boldsymbol{\Theta}\in\mathbb{R}^{d\times r}$ the parameter of the model. Fitting such models require optimizing losses of the form

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{d\times r}}\mathcal{L}(\boldsymbol{\Theta})=\frac{1}{2}\sum_{i=1}^n\left(y_i-\left\langle\boldsymbol{X}_i,\boldsymbol{\Theta}\boldsymbol{\Theta}^T\right\rangle\right)^2. \tag{4.3}$$

This approach, originally proposed by Burer and Monteiro (Burer & Monteiro, 2003), shifts the search space from a large low-rank positive semidefinite matrix $\boldsymbol{\Theta}\boldsymbol{\Theta}^T$ to its factor $\boldsymbol{\Theta}$. In this section we study the behavior of GD on this problem in the over-parameterized regime where $n < dr$.

Theorem 4.2 *Consider the problem of fitting a low-rank model of the form (4.3). Assume the input features $\boldsymbol{X}_i$ are i.i.d. matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Furthermore, assume the labels y_i are arbitrary and denote the vector of all labels by $\boldsymbol{y} \in \mathbb{R}^n$. Set the initial parameter $\boldsymbol{\Theta}_0 \in \mathbb{R}^{d\times r}$ to a matrix with singular values lying in the interval $\left[\frac{\sqrt{\|\boldsymbol{y}\|_{\ell_2}}}{\sqrt[4]{rn}}, 2\frac{\sqrt{\|\boldsymbol{y}\|_{\ell_2}}}{\sqrt[4]{rn}}\right]$ Furthermore, let $c, c_1, c_2 > 0$ be numerical constants and assume*

$$n \le cdr.$$

We set $\eta = \frac{c_1\sqrt{n}}{r^2 d\|\boldsymbol{y}\|_{\ell_2}}$ and run GD starting from $\boldsymbol{\Theta}_0$. Then, with probability at least $1 - 4e^{-\frac{n}{2}}$ all iterates obey

$$\sum_{i=1}^{n}\left(y_i - \left\langle \boldsymbol{X}_i, \boldsymbol{\Theta}_\tau\boldsymbol{\Theta}_\tau^T\right\rangle\right)^2 \le 100\left(1 - \frac{c_2}{r^{3/2}}\right)^\tau \|\boldsymbol{y}\|_{\ell_2}^2,$$

This theorem shows that with modest over-parametrization $dr \gtrsim n$, GD linearly converges to a globally optimal model and achieves zero loss. Note that the degrees of freedom of a $d\times r$ matrices is on the order of dr hence as soon as $n > dr$, gradient descent can no longer perfectly fit arbitrary labels highlighting a phase transition from zero loss to non-zero as sample size increases. Furthermore, our result holds despite the nonconvex nature of the Burer-Monteiro approach.

4.3. Training shallow neural networks

In this section we specialize our general approach in the context of training simple shallow neural networks. We shall focus on neural networks with only one hidden layer with d inputs, k hidden neurons and a single output. The overall input-output relationship of the neural network in this case is a function $f(\cdot;\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ that maps the input vector $\boldsymbol{x} \in \mathbb{R}^d$ into a scalar output via the following equation

$$\boldsymbol{x} \mapsto f(\boldsymbol{x};\boldsymbol{W}) = \sum_{\ell=1}^{k} \boldsymbol{v}_\ell \phi\left(\langle \boldsymbol{w}_\ell, \boldsymbol{x}\rangle\right).$$

In the above the vectors $\boldsymbol{w}_\ell \in \mathbb{R}^d$ contains the weights of the edges connecting the input to the ℓth hidden node and $\boldsymbol{v}_\ell \in \mathbb{R}$ is the weight of the edge connecting the ℓth hidden node to the output. Finally, $\phi : \mathbb{R} \to \mathbb{R}$ denotes the activation function applied to each hidden node. For more compact notation we gather the weights $\boldsymbol{w}_\ell/\boldsymbol{v}_\ell$ into larger matrices $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ of the form $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{w}_1 & \boldsymbol{w}_2 & \dots & \boldsymbol{w}_k\end{bmatrix}^T$ and $\boldsymbol{v} = \begin{bmatrix}v_1 & v_2 & \dots & v_k\end{bmatrix}^T$. We can now rewrite our input-output model in the more succinct form

$$\boldsymbol{x} \mapsto f(\boldsymbol{x};\boldsymbol{W}) := \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x}). \tag{4.4}$$

Here, we have used the convention that when ϕ is applied to a vector it corresponds to applying ϕ to each entry of that vector. In this paper we assume $\boldsymbol{v} \in \mathbb{R}^k$ is fixed and we train for the input-to-hidden weights $\boldsymbol{W}$. Without loss of generality we assume $\boldsymbol{v} \in \mathbb{R}^k$ has unit Euclidean norm i.e. $\|\boldsymbol{v}\|_{\ell_2} = 1$. The training problem then takes the form

$$\min_{\boldsymbol{W}\in\mathbb{R}^{k\times d}} \mathcal{L}(\boldsymbol{W}) := \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)^2. \tag{4.5}$$

The theorem below provides global geometric convergence guarantees for one-hidden layer neural networks in a simple over-parametrized regime.

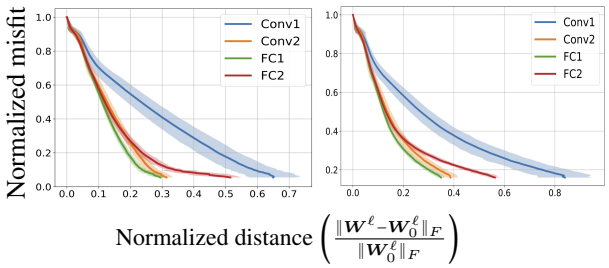
Theorem 4.3 *Consider a data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \dots, n$ aggregated as rows/entries of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a vector $\boldsymbol{y} \in \mathbb{R}^n$ with $n \le d$. Also consider a one-hidden layer neural network with k hidden units and one output as in (4.4). We assume the activation ϕ is strictly increasing with bounded derivatives i.e. $0 < \gamma \le \phi'(z) \le \Gamma$ and $\phi''(z) \le M$ for all z, $\boldsymbol{v}$ is fixed with unit Euclidean norm ($\|\boldsymbol{v}\|_{\ell_2} = 1$) and train only over $\boldsymbol{W}$. Starting from arbitrary $\boldsymbol{W}_0$, run gradient descent on the loss (4.5) with $\eta \le \frac{1}{2\Gamma^2\|\boldsymbol{X}\|^2}\min\left(1, \frac{\gamma^2}{\Gamma M}\frac{\sigma_{\min}(\boldsymbol{X})^2}{\|\boldsymbol{X}\|_{2,\infty}\|\boldsymbol{X}\|}\frac{1}{\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}}\right)$.[4] Then, all GD iterates obey*

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \left(1 - \eta\gamma^2\sigma_{\min}^2(\boldsymbol{X})\right)^\tau \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}.$$

This theorem demonstrates that the nice properties discussed in this paper also holds for one-hidden-layer networks in the regime where $n \le d$ from arbitrary initialization and the result is independent of number of hidden nodes k. This result holds for strictly increasing activations where ϕ' is bounded away from zero. While this might seem restrictive, we can obtain such a function by adding a small linear component to any non-decreasing function i.e. $\tilde{\phi}(x) = (1-\gamma)\phi(x) + \gamma x$. For instance, the commonly used leaky ReLU is obtained from ReLU in this way. We focus on such activations so as to ensure the result holds from arbitrary initialization. As we discuss below it is possible to relax this assumption when the algorithms are initialized at random.

We would like to emphasize that neural networks seem to work with much less over-parameterization e.g. for one hidden networks like the above $kd \gtrsim n$ seems to be sufficient. As such there is a huge gap between the $n \le d$ result above and practical use. That said, our main theoretical guarantees from Theorems 2.1 and 3.1 when combined with more intricate techniques from random matrix theory and stochastic processes continue to apply in this setting. In particular, in a companion paper (Oymak & Soltanolkotabi, 2019) we demonstrate that starting from a random initialization the result above continues to hold without the need for strictly increasing activations (including ReLU and softplus) and with much more modest amounts of over-parameterization.

[4] $\|\boldsymbol{X}\|_{2,\inf}$ denotes the maximum ℓ_2 norm of the rows of $\boldsymbol{X}$.



(a) $n = 500$ (b) $n = 5000$

Figure 2. The normalized misfit-distance trajectory for MNIST training for different layers of the network and different sample sizes. The layers from input to output are Conv1, Conv2, FC1, and FC2. Each curve represents the average normalized distance (for each layer of the network) corresponding to a fixed normalized misfit value over 20 independent realizations. The two standard deviation around the mean is highlighted via the shaded region.

5. Numerical Experiments

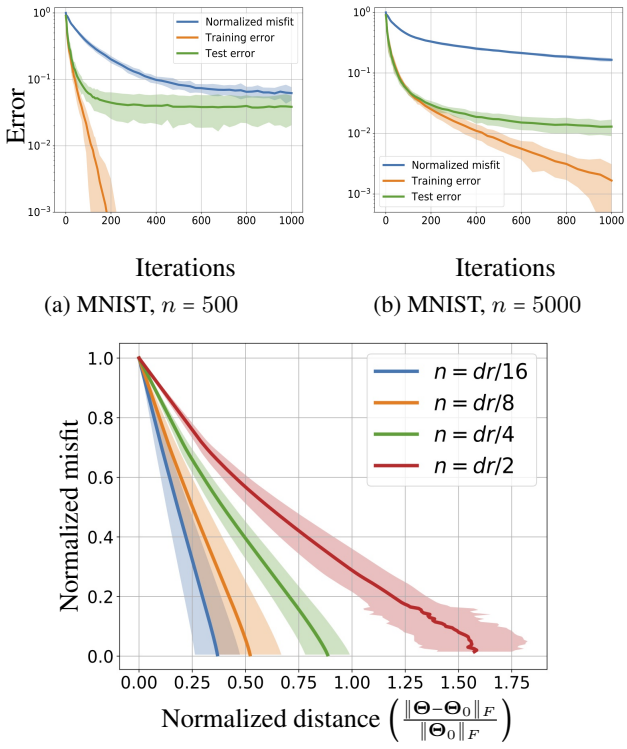
To verify our theoretical claims, we conducted experiments on MNIST classification and low-rank matrix regression. To illustrate the tradeoffs between the loss function and the distance to the initial point, we define *normalized misfit* and *normalized distance* as follows.

$$\text{misfit} = \frac{\|\boldsymbol{y} - f(\boldsymbol{\theta})\|_{\ell_2}}{\|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}} \quad , \quad \text{distance} = \frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}}{\|\boldsymbol{\theta}_0\|_{\ell_2}}.$$

5.1. MNIST Experiments

We consider MNIST digit classification task and use a standard LeNet model (LeCun et al., 1998) from Tensorflow (Abadi et al., 2016). This model has two convolutional layers followed by two fully-connected layers. Instead of cross-entropy loss, we use least-squares loss, without softmax layer, which falls within our nonlinear least-squares framework. We conducted two set of experiments with $n = 500$ and $n = 5000$. Both experiments use Adam with learning rate 0.001 and batch size 100 for 1000 iterations. At each iteration, we record the normalized misfit and distance to obtain a misfit-distance trajectory similar to Figure 1. We repeat the training 20 times (with independent initialization and dataset selection) to obtain the typical behavior.

Since layers have distinct goals (feature extraction vs classification), we kept track of the behavior of individual layers. Specifically, denote the weights of the ℓth layer of the neural network by $\boldsymbol{W}^\ell$, we consider the per-layer normalized distances $\frac{\|\boldsymbol{W}^\ell - \boldsymbol{W}_0^\ell\|_F}{\|\boldsymbol{W}_0^\ell\|_F}$ where layer ℓ is either convolutional (Conv1, Conv2) or fully-connected (FC1, FC2). In Figure 2, we depict the normalized misfit-distance tradeoff for different layers and sample sizes. Figure 2a illustrates the heavily overparameterized regime which has fewer samples. During the initial phase of the training (i.e. misfit ≤ 0.2) all layers follow a straight loss-distance line which is consistent with our theory (e.g. Figure 1). Towards the end of the training, the lines slightly level off which is most visible for the output layer FC2. This is likely due to the degradation of the Jacobian condition number as the model overfits to the data. Figure 3a plots the training and test errors together with normalized misfit to illustrate this. While misfit is around 0.05 at iteration 1000, the in-sample (classification) error hits 0 very quickly at iteration 200.



(a) MNIST, $n = 500$ (b) MNIST, $n = 5000$

(c) Low-rank regression

Figure 3. Figures 3a and 3b represent the test, training errors and normalized misfit corresponding to Figure 2. The x-axis is the number of iterations. Figure 3c highlights the loss-distance trajectory for low-rank matrix regression with $d = 100$ and $r = 4$.

In Figure 2b and 3b we increase the sample size to $n = 5000$. Similar to the first case, during the initial phase (misfit ≤ 0.4) the loss-distance curve is a straight line and levels off later on. Compared to $n = 500$, leveling off occurs earlier and is more visible. For instance, at misfit $= 0.2$, output layer FC2 has distance of 0.5 for $n = 5000$ and 0.25 for $n = 500$. This is consistent with Theorem 2.1 which predicts (i) more samples imply a Jacobian with worse condition number and (ii) the global minimizer lies further away from the initialization and it is less-likely that the Jacobian will be well-behaved over this larger neighborhood.

5.2. Low-rank regression

We consider a synthetic low-rank regression setup to test the predictions of Theorem 4.2. We generate input ma-

trices with i.i.d. standard normal entries and labels with i.i.d. Rademacher entries. We set $r = 4$ and $d = 100$ and initialize $\mathbf{\Theta}_0$ according to Theorem 4.2. We vary the sample size $n \in \{25, 50, 100, 200\} = \{dr/16, dr/8, dr/4, dr/2\}$ and run gradient descent for 200 iterations with a constant learning rate per Theorem 4.2. We observe a linear tradeoff in terms of misfit-distance to initialization with a narrow confidence interval consistent with our theoretical predictions in Figure 1. In the large sample size ($n = dr/2$), the problem is less over-parameterized and the confidence intervals become notably wider especially when the misfit is close to zero (i.e. by the time we reach a global minima). As predicted by our main theorem, the distance to initialization $\mathbf{\Theta}_0$ increases gracefully as the number of labels n increases.

6. Prior Art

Here, we briefly discuss some closely related literature. See the supplementary material for a more in depth discussion.

Implicit regularization: There is a growing interest in understanding properties of overparameterized problems. An interesting body of work investigate the implicit regularization capabilities of (stochastic) gradient descent for separable classification problems including (Azizan & Hassibi, 2018; Gunasekar et al., 2017; Nacson et al., 2018; Neyshabur et al., 2014; 2017; Soudry et al., 2017; Wilson et al., 2017). 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 (Bhojanapalli et al., 2016; Boumal et al., 2016; Burer & Monteiro, 2003; Li et al., 2018)). However, for regression problems based on a least-squares formulation the implicit bias/minimal norm property is proven under the assumption that gradient descent converges to a globally optimal solution which is not rigorously proven in most of these papers.

Overparameterized neural networks: A few recent papers (Arora et al., 2018a; Brutzkus & Globerson, 2018; Brutzkus et al., 2017b; Chizat & Bach, 2018; Ji & Telgarsky, 2018; Soltanolkotabi et al., 2018; Soudry & Carmon, 2016; Venturi et al., 2018; Zhang et al., 2016; Zhu et al., 2018) study the benefits of overparameterization for training neural networks and related optimization problems. Very recent works (Allen-Zhu et al., 2018a;b; Du et al., 2018a;b; Li & Liang, 2018; Zou et al., 2018) 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 results are not directly comparable to each other. We assume $n \leq d$ and use an arbitrary initialization where as these papers assume $poly(n) \lesssim k$ and start from random initialization. The results further defer in terms of other assumptions and conclusions. In contrast to these papers on neural nets, we focus on general nonlinearities and also on the gradient descent trajectory showing that among all the global optima, gradient descent converges to one with near minimal distance to the initialization. We would also like to note that the importance of the Jacobian for overparameterized neural network analysis has also been noted by other papers including (Du et al., 2018b; Soltanolkotabi et al., 2018) and also (Chaudhari et al., 2016; Keskar et al., 2016; Sagun et al., 2017) which investigate the optimization landscape and properties of SGD for training neural networks. An equally important question to understanding the convergence behavior of optimization algorithms for overparameterized models is understanding their generalization capabilities this is the subject of a few interesting recent papers (Arora et al., 2018b; Bartlett et al., 2017; Belkin et al., 2018a;b; Brutzkus et al., 2017a; Golowich et al., 2017; Li et al., 2019; Liang & Rakhlin, 2018; Oymak, 2018; Song et al., 2018). While our results do not directly address generalization, by characterizing the properties of the global optima that (stochastic) gradient descent converges to it may help demystify the generalization capabilities of overparametrized models.

7. Discussion and future directions

This work provides new insights and theory for overparameterized learning with nonlinear models. We first provided a general convergence result for gradient descent and matching upper and lower bounds showing that if the Jacobian of the nonlinear mapping is well-behaved in a minimally small neighborhood, gradient descent finds a global minimizer which has a nearly minimal distance to the initialization. Second, we extend the results to SGD to show that SGD exhibits the same behavior and converges linearly without ever leaving a minimally small neighborhood of initializtion. Finally, we specialize our general theory to provide new results for overparameterized learning with generalized linear models, low-rank regression and shallow neural network training. A key tool in our results is that we introduce a potential function that captures the tradeoff between the model misfit and the distance to the initial point: the decrease in loss is proportional to the distance from the initialization. Our numerical experiments on real and synthetic data further corroborate this intuition on the loss-distance tradeoff.

In this work we address important challenges surrounding the optimization of nonlinear over-parametrized learning via GD and SGD and some of its key features. The fact that gradient descent finds a nearby solution is a desirable property that hints as to why *generalization* to new data instances may be possible. However, we emphasize that this is only suggestive of the generalization capabilities of such algorithms to new data. Indeed, developing a clear understanding of the generalization capabilities of first order methods when solving over-parameterized nonlinear problems is an important future direction.

Acknowledgements

M. Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award #FA9550-18-1-0078, an NSF-CIF award #1813877, and a Google faculty research award.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pp. 265–283, 2016.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*, 2018a.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. 02 2018b. URL https://arxiv.org/pdf/1802.05296.

Azizan, N. and Hassibi, B. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. *arXiv preprint arXiv:1806.00952*, 2018.

Bartlett, P., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. 06 2017. URL https://arxiv.org/pdf/1706.08498.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Belkin, M., Hsu, D., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. 06 2018a. URL https://arxiv.org/pdf/1806.05161.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? 06 2018b. URL https://arxiv.org/pdf/1806.09471.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.

Boumal, N., Voroninski, V., and Bandeira, A. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pp. 2757–2765, 2016.

Brutzkus, A. and Globerson, A. Over-parameterization improves generalization in the xor detection problem. 10 2018. URL https://arxiv.org/pdf/1810.03037.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. Sgd learns over-parameterized networks that provably generalize on linearly separable data. 10 2017a. URL https://arxiv.org/pdf/1710.10174.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017b.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. 12 2017. URL https://arxiv.org/pdf/1712.06541.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. 10 2018. URL https://arxiv.org/pdf/1810.02032.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, M., Soltanolkotabi, M., and Oymak, S. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680*, 2019.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *NeurIPS*, 2018.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47, 2018.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. 08 2018. URL https://arxiv.org/pdf/1808.00387.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nacson, M. S., Srebro, N., and Soudry, D. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796*, 2018.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Neyshabur, B., Tomioka, R., Salakhutdinov, R., and Srebro, N. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Oymak, S. Learning compact neural networks with regularization. *International Conference on Machine Learning*, 2018.

Oymak, S. and Soltanolkotabi, M. Towards moderate overparameterization: global convergence guarantees for training neural networks. 2019.

Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

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

Soudry, D. and Carmon, Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. 05 2016. URL https://arxiv.org/pdf/1605.08361.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Talagrand, M. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

Tan, Y. S. and Vershynin, R. Phase retrieval via randomized kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, 2017.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, 2015.

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

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pp. 4148–4158, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhu, Z., Soudry, D., Eldar, Y. C., and Wakin, M. B. The global optimization geometry of shallow linear neural networks. 05 2018. URL https://arxiv.org/pdf/1805.04938.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.