[

Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?

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

Abstract

Many modern learning tasks involve fitting nonlinear models to data 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 optima. As part of our proof technique, we introduce a new potential function which captures the precise tradeoff between the loss function and the distance to the initial point as the iterations progress. For Stochastic Gradient Descent (SGD), we develop novel martingale techniques that guarantee SGD never leaves a small neighborhood of the initialization, even with rather large learning rates. We demonstrate the utility of our general theory 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. In Section 5 we discuss results that apply 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

[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. Copyright 2019 by the author(s).

[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.

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 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 the global optima. We believe these key features of first-order methods may help demystify why models trained using these simple algorithms can achieve reliable learning in modern over-parametrized regimes without over-fitting to the training data.

1.2. Insights from Linear Regression

As a prelude to understanding the key properties of (stochastic) gradient descent in over-parameterized nonlinear learning we begin by focusing on the simple case of linear regression. In this case the mapping in (1.1) 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 $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y}\right\|_{\ell_2}^2$. Therefore, starting from an initialization $\boldsymbol{\theta}_0$, gradient descent iterations with a step size η take the form

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau) = \boldsymbol{\theta}_\tau - \eta\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{\theta}_\tau-\boldsymbol{y}\right).$$

As long as the matrix $\boldsymbol{X}$ has full row rank the set $\mathcal{G} := \{\boldsymbol{\theta}\in\mathbb{R}^p : \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}\}$ is nonempty and the global minimum of the loss is 0. Using simple algebraic manipulations the residual vector $\boldsymbol{r}_\tau = \boldsymbol{X}\boldsymbol{\theta}_{\tau+1} - \boldsymbol{y}$ obeys

$$\boldsymbol{r}_{\tau+1} = \left(\boldsymbol{I}-\eta\boldsymbol{X}\boldsymbol{X}^T\right)\boldsymbol{r}_\tau \quad\Rightarrow\quad \left\|\boldsymbol{r}_{\tau+1}\right\|_{\ell_2} \le \left\|\boldsymbol{I}-\eta\boldsymbol{X}\boldsymbol{X}^T\right\|\left\|\boldsymbol{r}_\tau\right\|_{\ell_2}.$$

Therefore, using a step size of $\eta \le \frac{1}{\|\boldsymbol{X}\|^2}$ the residual iterates converge at a geometric rate to zero. This yields the first key property of gradient methods for over-parametrized learning:

> ***Key property I:*** *Gradient descent iterates converge at a geometric rate to a global optima.*

Let $\boldsymbol{\theta}^*$ denote the global minima we converge to and $\Pi_{\mathcal{R}}$ and $\Pi_{\mathcal{N}}$ denote the projections onto the row space and null space of $\boldsymbol{X}$, respectively. Since the gradients lie on the row space of $\boldsymbol{X}$ and $\boldsymbol{X}$ is full row rank, denoting the unique pseudo-inverse solution by $\boldsymbol{\theta}^\dagger$, we have

$$\Pi_{\mathcal{N}}(\boldsymbol{\theta}^*) = \Pi_{\mathcal{N}}(\boldsymbol{\theta}_0) \qquad\text{and}\qquad \Pi_{\mathcal{R}}(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^\dagger.$$

The equalities above imply that $\boldsymbol{\theta}^*$ is the closest global minima to $\boldsymbol{\theta}_0$; which highlights the second property:

> ***Key property II:*** *Gradient descent converges to the closest global optima to initialization.*

Finally, it can also be shown that the total path length $\sum_{\tau=0}^{\infty}\left\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\right\|_{\ell_2}$ can be upper bounded by the distance $\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\right\|_{\ell_2}$ (up to multiplicative factors depending on condition number of $\boldsymbol{X}$). This leads us to:

> ***Key property III:*** *Gradient descent takes a near direct trajectory to reach the closest global optima.*

In this paper we show that similar properties continue to hold for a broad class of *nonlinear* over-parameterized learning problems.

1.3. Contributions

Our main technical 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 as gradient descent and 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

$$\boldsymbol{y} := \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^n \quad \text{and} \quad f(\boldsymbol{\theta}) := \begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\theta}) \\ f(\boldsymbol{x}_2;\boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_n;\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^n.$$

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 nonlinear least-squares formulation (2.1) above the gradient takes the form

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

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.

The particular form of the gradient in (2.2) suggests that the singular values of the Jacobian matrix may significantly impact the convergence of gradient descent. Our main technical assumption in this paper is that the spectrum of the Jacobian matrix is bounded from below and above in a local neighborhood of the initialization.

Assumption 1 (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 β and α scalars obeying $\beta\ge\alpha>0$. Here, $\sigma_{\min}(\cdot)$ and $\|\cdot\|$ denote the minimum singular value and the spectral norm respectively.

Our second technical assumption ensures that the Jacobian matrix is not too sensitive to changes in the parameters of the nonlinear mapping. Specifically we require the Jacobian to have either bounded or smooth variations as detailed next.

Assumption 2 (Jacobian Deviations) *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 one of the following two conditions holds:*
(a) Bounded deviation: *For all $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathcal{D}$*

$$\left\|\mathcal{J}(\boldsymbol{\theta}_2)-\mathcal{J}(\boldsymbol{\theta}_1)\right\| \le \frac{(1-\lambda)\alpha^2}{\beta},$$

holds for some $0\le\lambda\le 1$. Here, α and β are the bounds on the Jacobian spectrum over $\mathcal{D}$ per Assumption 1.
(b) Smooth deviation: *For all $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathcal{D}$*

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

[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).

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*

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

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 1 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 one of the following statements is valid.*

- *Assumption 2 (a) holds over $\mathcal{D}$ with $\lambda = 1/2$ and set $\eta \le \frac{1}{2\beta^2}$.*
- *Assumption 2 (b) 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 along. The following corollary highlights the key relations between smoothness, residual, and initial Jacobian for global convergence.

Theorem 2.2 *Let $\beta,\alpha>0$ and suppose the Jacobian at $\boldsymbol{\theta}_0$ obeys*

$$2\alpha \le \sigma_{\min}\left(\mathcal{J}(\boldsymbol{\theta}_0)\right) \le \|\mathcal{J}(\boldsymbol{\theta}_0)\| \le \beta/2.$$

Additionally, suppose Assumption 2 holds over a ball of radius $R = \frac{4\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ around $\boldsymbol{\theta}_0$ and

$$\alpha^2 \ge 4L\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}. \tag{2.6}$$

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

Proof We simply need to verify the conditions of Theorem 2.1 over $R = \frac{4\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ neighborhood of $\boldsymbol{\theta}_0$. Since $\mathcal{J}(\boldsymbol{\theta})$ has Lipschitz spectral norm and $L \le \frac{\alpha^2}{4\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}}$, we have $LR\le\alpha\le\beta/2$. Hence, for any $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\le R$, we find

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta})) \ge \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) - LR \ge 2\alpha-\alpha=\alpha, \tag{2.7}$$

$$\|\mathcal{J}(\boldsymbol{\theta})\| \ge \|\mathcal{J}(\boldsymbol{\theta}_0)\| + LR \le \frac{\beta}{2}+\alpha\le\beta. \tag{2.8}$$

Hence, Assumption 1 holds and conclusions of Theorem 2.1 follows. What remains is determining learning rate, in particular ensuring $\eta\le\frac{1}{2\beta^2}\cdot\min\left(1,\frac{\alpha^2}{L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}\right)$. This follows from (2.6). ■

Another trivial consequence of our main theorem is the following corollary.

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

Corollary 2.3 *Consider the setting and assumptions of Theorem 2.1 above. Let $\boldsymbol{\theta}^*$ denote the global optima of the loss $\mathcal{L}(\boldsymbol{\theta})$ with smallest Euclidean distance to the initial parameter $\boldsymbol{\theta}_0$. Then, the gradient descent iterates $\boldsymbol{\theta}_\tau$ obey*

$$\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}\le 4\frac{\beta}{\alpha}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}, \tag{2.9}$$

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2}\le 4\frac{\beta}{\alpha}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}. \tag{2.10}$$

The theorem and corollary above show that if the Jacobian of the nonlinear mapping is well-conditioned (Assumption 1) and has bounded/smooth deviations (Assumptions 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 $\boldsymbol{\theta}_{GD}$. This hold despite the fact that the fitting problem may be highly nonconvex in general. Indeed, based on (2.3) the fitting/training error $\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2}$ 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 $\frac{4}{\alpha}\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}$ around the initial point. That is the GD iterates remain rather close to the initialization. In fact, based on (2.9) we can conclude that

$$\|\boldsymbol{\theta}_{GD}-\boldsymbol{\theta}_0\|_{\ell_2}=\left\|\lim_{\tau\to\infty}\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\right\|_{\ell_2}=\lim_{\tau\to\infty}\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}\le 4\frac{\beta}{\alpha}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}.$$

Thus the distance between the global optima GD converges to and the initial parameter $\boldsymbol{\theta}_0$ is within a factor $4\frac{\beta}{\alpha}$ of the distance between the closest global optima to $\boldsymbol{\theta}_0$ 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 (2.10) 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 the proofs (Theorem 9.3 stated in Section 9.2).[4] This more general result requires Assumptions 1 and 2 to hold in a smaller neighborhood and improves the approximation ratios. Specifically, this more general result allows the radius R to be chosen as small as

$$\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}, \tag{2.11}$$

and (2.4) to be improved to

$$\alpha\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}+\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2}\le\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2} \tag{2.12}$$

Also the approximation ratios in Corollary 2.3 can be improved to

$$\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}\le\frac{\beta}{\alpha}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}, \tag{2.13}$$

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2}\le\frac{\beta}{\alpha}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}. \tag{2.14}$$

However, this requires a smaller learning rate and hence leads to a slower converge guarantee.

[4] Theorem 2.1 and Corollary 2.3 above are a special case of this theorem with $\lambda=1/2$ and $\rho=1$.

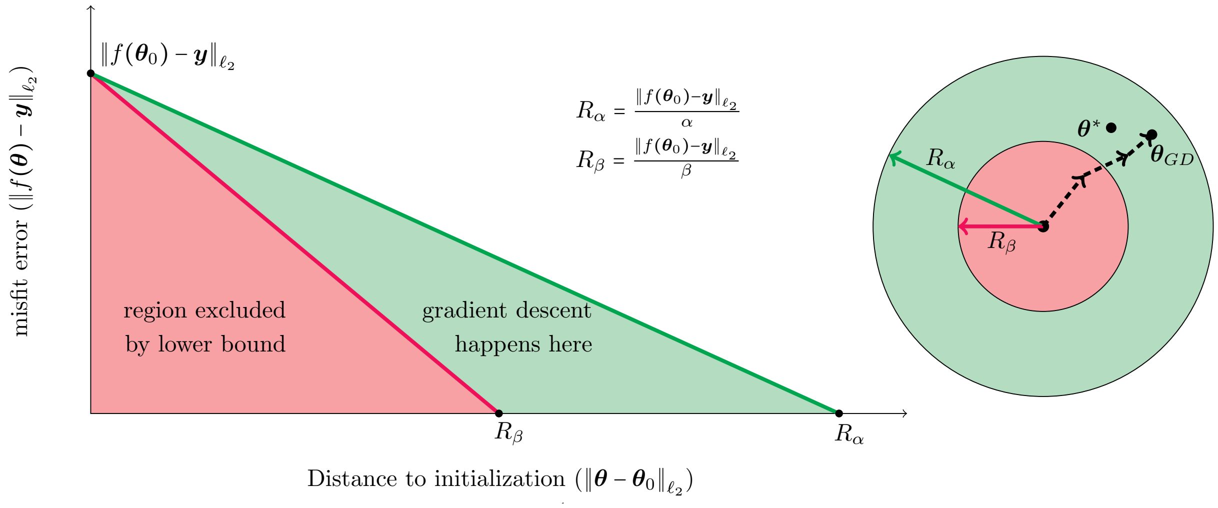

$\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}$
misfit error ($\|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2}$)
$R_\alpha = \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$
$R_\beta = \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\beta}$
region excluded by lower bound
gradient descent happens here
R_β
R_α
Distance to initialization ($\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}$)
$\boldsymbol{\theta}^*$
$\boldsymbol{\theta}_{GD}$
R_α
R_β

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 ($\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}$) and the misfit error ($\|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2}$). Our upper (equation (2.12)) 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 $\boldsymbol{\theta}_{GD}$ 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.

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). Secondly, 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 follows.

- More samples leads to a slower convergence rate by degrading the condition number of the Jacobian,
- The required convergence radius R increases proportional to $\sqrt{n}$ and we need Jacobian to be well-behaved over a larger neighborhood for fast convergence.

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*

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

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 1 over a set $\mathcal{D}$ around a point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. Then,

$$\|\boldsymbol{y} - f(\boldsymbol{\theta})\|_{\ell_2} + \beta\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \geq \|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}, \tag{2.15}$$

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

$$\|\boldsymbol{y} - f(\boldsymbol{\theta})\|_{\ell_2} + \alpha\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \geq \|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}, \tag{2.16}$$

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

$$\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2} + \beta\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} = \|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}. \tag{2.17}$$

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 (2.11) 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 (2.12) and (2.15)). 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 (2.11) and (2.12), 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 (2.12)) 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 learning nonlinear least-squares problems of the form (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

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

Here, $G(\boldsymbol{\theta}_\tau;\gamma_\tau)$ is the gradient on the γ_τth training sample. We are interested in understanding the trajectory of SGD for over-parameterized learning. In particular, whether the three intriguing properties discussed in the previous section for 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 1 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 one of the following statements is valid.

- *Assumption 2 (a) holds over $\mathcal{D}$ and set $\eta\leq\frac{\alpha^2}{\nu\beta^2B^2}$.*
- *Assumption 2 (b) 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 which holds with probability at least $\mathbb{P}(E)\geq 1-\frac{4}{\nu}\left(\frac{\beta}{\alpha}\right)^{\frac{1}{p}}$ and running stochastic gradient descent updates of the form (3.1) starting from $\boldsymbol{\theta}_0$, all iterates obey

$$\mathbb{E}\left[\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2}^2\mathbb{1}_E\right]\leq\left(1-\frac{\eta\alpha^2}{2n}\right)^\tau\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}^2, \tag{3.2}$$

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 showing that the weighted sum of the misfit and distance to initialization remains bounded per (2.4). Rather we show a more intricate function (discussed in detail in Lemma 9.11 and illustrated in Figure 4 in the proofs) remains bounded. This function 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 that keeps track of the average distances to a set of points close to the initialization and ensures 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. 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 guarantees that SGD can use 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 neural network training.

4.1. Learning generalized linear models

Nonlinear data-fitting problems are fundamental to many supervised learning tasks in machine learning. Given training data consisting of n pairs of input features $\boldsymbol{x}_i\in\mathbb{R}^p$ and desired outputs $y_i\in\mathbb{R}$ we wish to infer a function that best explains the training data. In this section we focus on learning Generalized Linear Models (GLM) from data which involves fitting functions of the form $f(\cdot;\boldsymbol{\theta}):\mathbb{R}^d\rightarrow\mathbb{R}$

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

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 $n\le p$. This is the subject of the next two theorems.

Theorem 4.1 (Overparameterized GLM) *Consider a data set of input/label pairs $\boldsymbol{x}_i\in\mathbb{R}^p$ and y_i for $i=1,2,\ldots,n$ aggregated as rows/entries of a matrix $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ and a vector $\boldsymbol{y}\in\mathbb{R}^n$ with $n\le p$. Also consider a Generalized Linear Model (GLM) of the form $\boldsymbol{x}\mapsto\phi\left(\langle\boldsymbol{x},\boldsymbol{\theta}\rangle\right)$ with $\phi:\mathbb{R}\rightarrow\mathbb{R}$ a strictly increasing nonlinearity with continuous derivatives (i.e. obeying $0<\gamma\le\phi'(z)\le\Gamma$ for all z). Starting from an initial parameter $\boldsymbol{\theta}_0$ we run gradient descent updates of the form $\boldsymbol{\theta}_{\tau+1}=\boldsymbol{\theta}_\tau-\eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ on the loss (4.1) with $\eta\le\frac{1}{\|\boldsymbol{X}\|^2\Gamma^2}$. Furthermore, let $\boldsymbol{\theta}^\star$ denote the closest global optima to $\boldsymbol{\theta}_0$. Then, all GD iterates obey*

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

The above 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 optima to the initialization parameter. Also, we can deduce from (4.2) that the total gradient path length when using a step size on the order of $\frac{1}{\|\boldsymbol{X}\|^2\Gamma^2}$

is bounded by

$$\sum_{\tau=0}^{\infty} \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_{\tau}\|_{\ell_2} \leq \frac{\Gamma^2}{\gamma^2} \frac{\lambda_{\max}\left(\boldsymbol{X}\boldsymbol{X}^T\right)}{\lambda_{\min}\left(\boldsymbol{X}\boldsymbol{X}^T\right)} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_{\ell_2}, \tag{4.3}$$

so that the total path length is a constant multiple of the distance between initialization and the closest global optima. Furthermore, applying Theorem 2.1 with a smaller learning rate, the right hand side can be improved to $\frac{\Gamma}{\gamma} \frac{\|\boldsymbol{X}\|}{\sigma_{\min}(\boldsymbol{X})} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_{\ell_2}$. Thus, gradient descent takes a near direct route.

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 $i = 1, 2, \dots, n$, 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

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{d\times r}} \mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{2}\sum_{i=1}^{n} \left(y_i - \left\langle \boldsymbol{X}_i, \boldsymbol{\Theta}\boldsymbol{\Theta}^T \right\rangle\right)^2. \tag{4.4}$$

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 and SGD 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 $\boldsymbol{X} \mapsto f(\boldsymbol{X}; \boldsymbol{\Theta}) = trace\left(\boldsymbol{\Theta}^T \boldsymbol{X} \boldsymbol{\Theta}\right)$ with $\boldsymbol{\Theta} \in \mathbb{R}^{d\times r}$ with $r \leq d$ to a data set $(y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^{d\times d}$ for $i = 1, 2, \dots, n$ via the loss (4.4). Assume the input features $\boldsymbol{X}_i$ are random and distributed i.i.d. with entries i.i.d. $\mathcal{N}(0, 1)$. 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*

$$n \leq cdr.$$

We run gradient descent iterations of the form $\boldsymbol{\Theta}_{\tau+1} = \boldsymbol{\Theta}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{\Theta}_\tau)$ starting from $\boldsymbol{\Theta}_0$ with $\eta = \frac{c_1\sqrt{n}}{r^2 d\|\boldsymbol{y}\|_{\ell_2}}$. Then, with probability at least $1 - 4e^{-\frac{n}{2}}$ all GD iterates obey

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

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 degrees of freedom of $d \times r$ matrices is 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

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_k^T \end{bmatrix} \quad \text{and} \quad \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}.$$

We can now rewrite our input-output model in the more succinct form

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}; \boldsymbol{W}) \coloneqq \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}). \tag{4.5}$$

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. When training a neural network, one typically has access to a data set consisting of n feature/label pairs $(\boldsymbol{x}_i, y_i)$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ representing the feature and y_i the associated label. We wish to infer the best weights $\boldsymbol{v}, \boldsymbol{W}$ such that the mapping f best fits the training data. 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 optimization problem then takes the form

$$\min_{\boldsymbol{W}\in\mathbb{R}^{k\times d}} \mathcal{L}(\boldsymbol{W}) \coloneqq \frac{1}{2}\sum_{i=1}^n \left(\boldsymbol{v}^T \phi\left(\boldsymbol{W}\boldsymbol{x}_i\right) - y_i\right)^2. \tag{4.6}$$

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

Theorem 4.3 (Overparameterized Neural Nets) *Consider a data set of input/label pairs $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$ 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 of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$ with $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and $\boldsymbol{v} \in \mathbb{R}^k$ the input-to-hidden and hidden-to-output weights. 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. We assume $\boldsymbol{v}$ is fixed with unit Euclidean norm ($\|\boldsymbol{v}\|_{\ell_2} = 1$) and train only over $\boldsymbol{W}$. Starting from an initial weight matrix $\boldsymbol{W}_0$ we run gradient descent updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ on the loss (4.6) with $\eta \le \frac{1}{2\Gamma^2\|\boldsymbol{X}\|^2}\min\left(1, \frac{\gamma^2}{\Gamma M}\frac{\sigma_{\min}(\boldsymbol{X})^2}{\|\boldsymbol{X}\|_{2,\infty}\|\boldsymbol{X}\|}\frac{1}{\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}}\right)$.[5] Then, all GD iterates obey*

$$\|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \left(1 - \eta\gamma^2\sigma_{\min}^2(\boldsymbol{X})\right)^\tau \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}, \tag{4.7}$$

$$\frac{\gamma\sigma_{\min}(\boldsymbol{X})}{4}\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F + \|f(\boldsymbol{W}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{W}_0) - \boldsymbol{y}\|_{\ell_2}. \tag{4.8}$$

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.

[5] Here, $\|\boldsymbol{X}\|_{2,\text{inf}}$ denotes the maximum Euclidean norm of the rows of $\boldsymbol{X}$.

5. Beyond nonlinear least-squares

In this section we explore generalizations of our results beyond nonlinear least-squares problems. In particular we focus on optimizing a general loss $\mathcal{L}(\boldsymbol{\theta})$ over $\boldsymbol{\theta}\in\mathbb{R}^p$. For exposition purposes throughout this section we assume that $\mathcal{L}$ is differentiable and the global minimum is zero, i.e. $\min\limits_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})=0$[6]. This generalization will be based on a local variant of Polyak-Lojasiewicz (PL) inequality. We begin by discussing this local PL condition formally.

Definition 5.1 (Local PL condition) *We say that the Local PL inequality holds over a set $\mathcal{D}\subseteq\mathbb{R}^p$ with $\mu>0$ if for all $\boldsymbol{\theta}\in\mathcal{D}$ we have*

$$\|\nabla\mathcal{L}(\boldsymbol{\theta})\|_{\ell_2}^2\ge 2\mu\mathcal{L}(\boldsymbol{\theta}).$$

Our first result shows that when the PL inequality holds around a minimally small neighborhood of the initialization, the intriguing properties of gradient descent discussed in Theorem 2.1 and Corollary 2.3 continue to hold beyond nonlinear least-squares problems.

Theorem 5.2 *Let $\mathcal{L}:\mathbb{R}^p\rightarrow\mathbb{R}$ be a loss function. Let $\boldsymbol{\theta}_0\in\mathbb{R}^p$ be an initialization parameter and define the set $\mathcal{D}$ to a local neighborhood around this point as follows*

$$\mathcal{D}=\mathcal{B}(\boldsymbol{\theta}_0,R)\quad\text{with}\quad R=\sqrt{\frac{8\mathcal{L}(\boldsymbol{\theta}_0)}{\mu}}\quad\text{and}\quad\mu>0.$$

Assume the loss $\mathcal{L}$ obeys the local PL condition per Definition 5.1 and is L-smooth over $\mathcal{D}$ ($\|\nabla\mathcal{L}(\boldsymbol{\theta}_2)-\nabla\mathcal{L}(\boldsymbol{\theta}_1)\|_{\ell_2}\le L\|\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\|_{\ell_2}$ for all $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathcal{D}$). Then, starting from $\boldsymbol{\theta}_0$ running gradient descent updates of the form

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

with $\eta\le 1/L$, all iterates $\boldsymbol{\theta}_\tau$ obey the following inequalities

$$\mathcal{L}(\boldsymbol{\theta}_\tau)\le(1-\eta\mu)^\tau\mathcal{L}(\boldsymbol{\theta}_0),\tag{5.1}$$

$$\sqrt{\frac{\mu}{8}}\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}+\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)}\le\sqrt{\mathcal{L}(\boldsymbol{\theta}_0)}.\tag{5.2}$$

Furthermore, the total path length of gradient descent is bounded via

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2}\le\sqrt{\frac{8\mathcal{L}(\boldsymbol{\theta}_0)}{\mu}}.\tag{5.3}$$

Similar to Corollary 2.3 a trivial consequence of the above theorem is the following corollary.

Corollary 5.3 *Consider the setting and assumptions of Theorem 5.2 above. Let $\boldsymbol{\theta}^*$ denote the global optima of the loss $\mathcal{L}(\boldsymbol{\theta})$ with smallest Euclidean distance to the initial parameter $\boldsymbol{\theta}_0$. Then, the gradient descent iterates $\boldsymbol{\theta}_\tau$ obey*

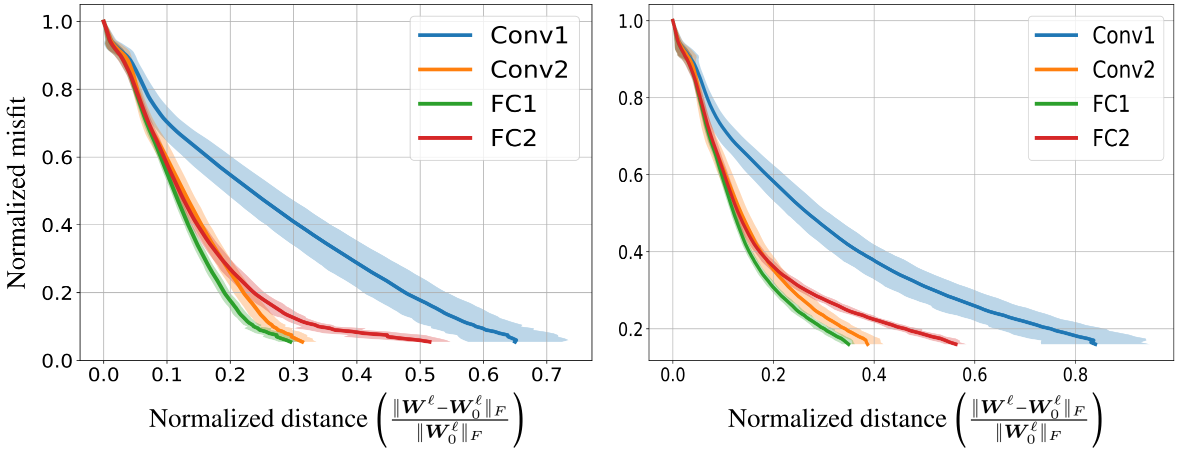
$$\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}\le 2\frac{L}{\mu}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2},\tag{5.4}$$

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2}\le 2\frac{L}{\mu}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}.\tag{5.5}$$

Similar to their nonlinear least-squares counter parts the theorem and corollary above show that if the loss function obeys the local PL condition and is smooth in a ball of radius R around the initial point then gradient descent enjoys three intriguing properties: (i) the iterates converge at a linear rate to a global optima, (ii) Gradient descent iterates remain close to the initialization and never leave a neighborhood of radius $2\frac{L}{\mu}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}$, and (iii) gradient descent iterates follow a near direct route to the global optima with the length of the path taken by GD iterates within a factor of the distance between the closest global optima and the initialization.

We end this section by discussing a simple lower bound which demonstrates that the required radius over which the Local PL result must hold per Theorem 5.2 is optimal up to a factor of two.

[6] Note that this is without loss of generality as for any loss we can apply the results to the shifted loss $\tilde{\mathcal{L}}(\boldsymbol{\theta})=\mathcal{L}(\boldsymbol{\theta})-\min\limits_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$.



(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 average distance is highlighted via the shaded region.

Theorem 5.4 *Let $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$ be an L-smooth loss function over a ball of radius R centered around a point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ ($\mathcal{B}(\boldsymbol{\theta}_0, R)$). Then there is no global minima over $\mathcal{B}(\boldsymbol{\theta}_0, R)$ when $R < \sqrt{2\mathcal{L}(\boldsymbol{\theta}_0)/L}$. Furthermore, for any μ and L obeying $L \geq \mu \geq 0$, there exists a loss $\mathcal{L}$ such that there is no global minima over the set $\mathcal{B}(\boldsymbol{\theta}_0, R)$ as long as $R < \sqrt{2\mathcal{L}(\boldsymbol{\theta}_0)/\mu}$.*

The result above shows that any global optima is at least a distance $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \geq \sqrt{2\mathcal{L}(\boldsymbol{\theta}_0)/L}$ away from the initialization so that the minimum ball around the initial point needs to have radius at least $R \geq \sqrt{2\mathcal{L}(\boldsymbol{\theta}_0)/L}$ for convergence to a global optima to occur. Comparing this lower-bound with that of Theorem 5.2 suggests that the size of the local neighborhood is optimal up to a factor 2. Collectively our theorems demonstrate that the path taken by gradient descent is by no means arbitrary. Indeed, under local PL and smoothness assumptions similar to Figure 1, gradient descent iterates exhibit a sharp tradeoff between distance to the initial point ($\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}$) and square root of loss value ($\sqrt{\mathcal{L}(\boldsymbol{\theta}_0)}$).

6. 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.

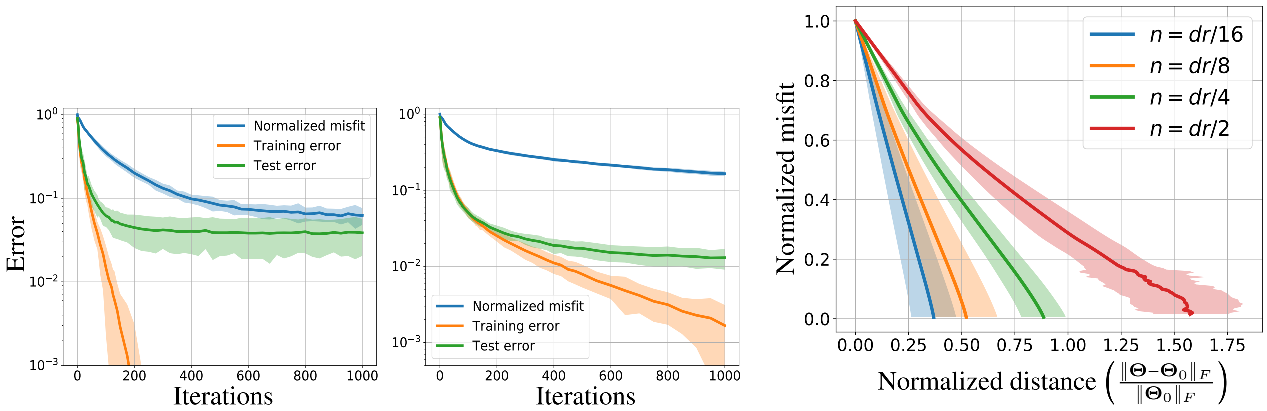
$$\text{Normalized misfit} = \frac{\|\boldsymbol{y} - f(\boldsymbol{\theta})\|_{\ell_2}}{\|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2}} \quad , \quad \text{Normalized distance} = \frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}}{\|\boldsymbol{\theta}_0\|_{\ell_2}}. \tag{6.1}$$

6.1. MNIST Experiments

We consider MNIST digit classification task and use a standard LeNet model (LeCun et al., 1998) from Tensorflow (Abadi et al., 2016)[7]. 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 9.13. 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.

[7] https://github.com/tensorflow/models/blob/master/research/slim/nets/lenet.py



(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 $r = 4$.

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 9.13). 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.

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.

6.2. Low-rank regression

We consider a synthetic low-rank regression setup to test the predictions of Theorem 4.2. We generate input matrices with i.i.d. standard normal entries and labels with i.i.d. Rademacher entries. We set $r = 4$ and $d = 100$ and initialize $\boldsymbol{\Theta}_0$ according to Theorem 4.2. We vary the sample size to be $n \in \{25, 50, 100, 200\} = \{dr/16, dr/8, dr/4, dr/2\}$ 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 9.13. 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 $\boldsymbol{\Theta}_0$ increases gracefully as the number of labels n increases.

7. Prior Art

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). 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 these papers. A recent paper (Li et al., 2017) does prove global convergence from a small initialization but requires fitting a $d \times d$ matrix ($\boldsymbol{U} \in \mathbb{R}^{d \times d}$) and operates with a suboptimal number of observations ($dr^2 \log^3 d$).

Overparametrized low-rank regression. As discussed in Section 4.2, there is a rich literature which studies global optimality of nonconvex low-rank factorization formulations such as the Burer-Monteiro factorization in the overparametrized regime(Bhojanapalli et al., 2016; Boumal et al., 2016; Burer & Monteiro, 2003; Li et al., 2018b). These results typically require the factorization rank to be at least $\sqrt{n}$ to guarantee convergence of gradient descent. In contrast, with random data but arbitrary features, our results guarantee global convergence as long as $r \gtrsim n/d$. Specifically, for the problem of nonconvex low-rank regression discussed in this paper if one assumes the labels are created according to a low-rank matrix of rank r^* (i.e. $y_i = \langle \boldsymbol{X}_i, \boldsymbol{\Theta}_* \boldsymbol{\Theta}_*^T \rangle$ with $\boldsymbol{\Theta}_* \in \mathbb{R}^{d \times r^*}$) and the number of labels is on the order of dr^* (i.e. $n = cdr^*$) then these classical results require the fitted rank to be $r \geq \sqrt{dr^*}$ where as our results work as soon as $r \gtrsim r^*$.

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 which show global convergence to a point somewhere around the initialization, 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 via a direct path. 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; Liang & Rakhlin, 2018; Oymak, 2018a; 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 trained via first order methods. Rigorous understanding of this relationship is an interesting and important subject for future research. See (Fabian et al., 2019) for some recent results in this direction.

Stochastic methods: SGD performance guarantees are typically in expectation rather than in probability. Martingale-based methods have been utilized to give probabilistic guarantees (De Sa et al., 2015; Rakhlin et al., 2012). The main challenge in nonconvex analysis of SGD, is to ensure SGD iterates stay within a region where nonconvex analysis can apply even when using rather large learning rates. While a few papers (Allen-Zhu et al., 2018b; Li & Liang, 2018) show that SGD stays in a specific region with high probability in specific instances, these results require using very small learning rates (which translates into very small variance) to ensure standard concentration arguments apply. In contrast, our approach allows for much larger learning rates by using martingale stopping time arguments. Our approach is in part inspired by (Tan & Vershynin, 2017) which studies SGD for nonconvex phase retrieval but involves different assumptions on the loss.

Nonconvex optimization: A key idea for solving nonconvex optimization problems is ensuring that optimization landscape has desirable properties. These properties include Polyak-Lojasiewicz (PL) condition (Lojasiewicz, 1963; Polyak, 1963) and the regularity condition (e.g. local strong convexity) (Candes et al., 2015; Hassani et al., 2017; Kalan et al., 2019; Li et al., 2018a; Oymak et al., 2015; Soltanolkotabi, 2017a; Xu et al., 2018). PL condition is particularly suited for analyzing overparameterized problems and has been utilized by several recent papers (Bassily et al., 2018; Karimi et al., 2016; Lei et al., 2017; Ma et al., 2017; Vaswani et al., 2018). Unlike these works, we show that overparameterized gradient descent trajectory stays in a small neighborhood and we only need properties such as PL to hold over this region. There is also a large body of work that study the applications discussed in this paper in the over determined regime $p \leq n$. For instance, Low-rank regression and generalized linear models have been considered by various works including (Bhojanapalli et al., 2016; Chen et al., 2018; Josz et al., 2018; Sun et al., 2018; Tu et al., 2015) in such an overdetermined setting. More recently, provable first order methods for learning neural networks have been investigated by multiple papers including (Alon Brutzkus &

Globerson, 2017; Ge et al., 2017; Oymak, 2018b; Soltanolkotabi, 2017b; Zhong et al., 2017) in the overdetermined setting.

8. 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 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. Making progress towards this generalization puzzle requires merging insights gained from optimization with more intricate tools from statistical learning and is an interesting topic for future research.

9. Proofs

9.1. Notations and definitions

Before we begin the proof we briefly discuss some notation and definitions that will be used throughout. The spectral norm and the minimum singular value of a matrix $\boldsymbol{A}$ is denoted by $\|\boldsymbol{A}\|/\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ respectively. $\|\boldsymbol{A}\|_{2,\infty}$ denotes the largest ℓ_2 norm among the rows of $\boldsymbol{A}$. $\mathcal{B}(\boldsymbol{\theta},R)$ denotes the ℓ_2 ball of radius R around a vector $\boldsymbol{\theta}$.

We introduce the following matrix and vector which play a crucial role in the convergence analysis of our algorithms

Definition 9.1 (Average Jacobian) *We define the average Jacobian along the path connecting two points $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^p$ as*

$$\mathcal{J}(\boldsymbol{y},\boldsymbol{x}) := \int_0^1 \mathcal{J}(\boldsymbol{x}+\alpha(\boldsymbol{y}-\boldsymbol{x}))d\alpha. \tag{9.1}$$

Definition 9.2 (Residual error) *We also define the residual error at iteration τ, denoted by $\boldsymbol{r}_\tau\in\mathbb{R}^n$, as the vector of misfits of the model to the labels. That is,*

$$\boldsymbol{r}_\tau = f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}.$$

9.2. Gradient descent convergence proofs (Theorem 2.1 and Corollary 2.3)

Theorem 2.1 and Corollary 2.3 are a special case of a more general result stated below. Theorem 2.1 and Corollary 2.3 then follows by setting $\lambda = 1/2$ and $\rho = 1$.

Theorem 9.3 *Consider a nonlinear least-squares optimization problem of the form*

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

with $f:\mathbb{R}^p\mapsto\mathbb{R}^n$ and $\boldsymbol{y}\in\mathbb{R}^n$. Let λ a scalar obeying $0<\lambda\le 1$. Suppose the Jacobian mapping associated with f obeys Assumption 1 over a ball of radius $R := \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{(\lambda-\eta\beta^2/2)\alpha}$ around a point $\boldsymbol{\theta}_0\in\mathbb{R}^p$, that is $\mathcal{D} = \mathcal{B}\left(\boldsymbol{\theta}_0,\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{(\lambda-\eta\beta^2/2)\alpha}\right)$. Furthermore, suppose one of the following statements is valid.

- *Assumption 2 (a) holds over $\mathcal{D}$ and set $\eta\le\frac{\lambda}{\beta^2}$.*

- *Assumption 2 (b) holds over $\mathcal{D}$ and set* $\eta \le \frac{1}{\beta^2} \cdot \min\left(\lambda, \frac{2(1-\lambda)\alpha^2}{L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}\right)$.

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.

$$\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2}^2 \le \left(1-\alpha^2\lambda\eta\right)^\tau \|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}^2, \tag{9.2}$$

$$(\lambda-\eta\beta^2/2)\alpha\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2} + \|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}. \tag{9.3}$$

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

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2} \le \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{(\lambda-\eta\beta^2/2)\alpha}. \tag{9.4}$$

Let $\boldsymbol{\theta}^*$ denote the global optimum of the loss $\mathcal{L}(\boldsymbol{\theta})$ with smallest Euclidean distance to the initial parameter $\boldsymbol{\theta}_0$. Then, the gradient descent iterates $\boldsymbol{\theta}_\tau$ also obey

$$\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2} \le \frac{\beta}{(\lambda-\eta\beta^2/2)\alpha}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}, \tag{9.5}$$

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2} \le \frac{\beta}{(\lambda-\eta\beta^2/2)\alpha}\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\|_{\ell_2}. \tag{9.6}$$

Proof Sketch. To prove the above theorem we begin by noting that the residual $\boldsymbol{r}_\tau$ satisfies the recursion

$$\begin{aligned}\boldsymbol{r}_{\tau+1} =& \boldsymbol{r}_\tau - f(\boldsymbol{\theta}_\tau) + f(\boldsymbol{\theta}_{\tau+1})\\ \overset{(a)}{=}& \boldsymbol{r}_\tau + \mathcal{J}(\boldsymbol{\theta}_{\tau+1},\boldsymbol{\theta}_\tau)(\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau)\\ \overset{(b)}{=}& \boldsymbol{r}_\tau - \eta\mathcal{J}(\boldsymbol{\theta}_{\tau+1},\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\\ =& (\boldsymbol{I}-\eta\boldsymbol{C}(\boldsymbol{\theta}_\tau))\boldsymbol{r}_\tau.\end{aligned} \tag{9.7}$$

where $\boldsymbol{C}(\boldsymbol{\theta}_\tau) := \mathcal{J}(\boldsymbol{\theta}_{\tau+1},\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T$. Here, (a) follows from fundamental rule of calculus and (b) from the gradient identity $\nabla\mathcal{L}(\boldsymbol{\theta}_\tau) = \mathcal{J}^T(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau$. If $\boldsymbol{I}-\eta\boldsymbol{C}(\boldsymbol{\theta}_\tau)$ has spectral norm less than 1, the the residual verctors will converge linearly. We build on this observation and show that one only needs this requirement over a minimally small neighborhood of $\boldsymbol{\theta}_0$. To this aim, we first introduce a potential set which contains the space of parameters that can be reached by gradient descent.

Definition 9.4 (Potential sub-level set) *Given a scalar $\zeta > 0$, define the radius $R_\zeta = \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}$. The potential sub-level set $\mathcal{P}(\boldsymbol{\theta}_0, R_\zeta)$ is defined as*

$$\mathcal{P}(\boldsymbol{\theta}_0, R_\zeta) = \left\{\boldsymbol{\theta}\in\mathbb{R}^p \;\middle|\; \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2} + \frac{\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}}{\zeta} \le R_\zeta\right\}. \tag{9.8}$$

Note that $\mathcal{P}(\boldsymbol{\theta}_0, R_\zeta) \subseteq \mathcal{B}(\boldsymbol{\theta}_0, R_\zeta)$. Our first lemma shows that, if an iterate $\boldsymbol{\theta}_\tau \in \mathcal{P} := \mathcal{P}(\boldsymbol{\theta}_0, R_\zeta)$, then the next iterate $\boldsymbol{\theta}_{\tau+1}$ stays in the set $\mathcal{D} := \mathcal{B}(\boldsymbol{\theta}_0, R_\zeta)$.

Lemma 9.5 *Suppose Assumption 1 holds over the domain $\mathcal{D} = \mathcal{B}\left(\boldsymbol{\theta}_0, \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$ for some ζ obeying $\zeta \le \alpha$. Also assume $\boldsymbol{\theta} \in \mathcal{P}(\boldsymbol{\theta}_0, R_\zeta)$, then gradient iterate $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta\nabla\mathcal{L}(\boldsymbol{\theta})$ with $\eta \le \frac{1}{\beta^2}$ satisfies $\boldsymbol{\theta}^+ \in \mathcal{D}$.*

Proof We begin by noting that

$$\|\boldsymbol{\theta}^+-\boldsymbol{\theta}\|_{\ell_2} = \eta\|\mathcal{J}^T(\boldsymbol{\theta})(f(\boldsymbol{\theta})-\boldsymbol{y})\|_{\ell_2} \overset{(a)}{\le} \eta\beta\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2} \overset{(b)}{\le} \frac{\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}}{\beta} \overset{(c)}{\le} \frac{\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}}{\alpha} \overset{(d)}{\le} \frac{\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}}{\zeta}. \tag{9.9}$$

In the above, (a) follows from the upper bound on the Jacobian over $\mathcal{D}$ per Assumption 1, (b) from the fact that $\eta \le \frac{1}{\beta^2}$, (c) from $\alpha \le \beta$, and (d) from $\zeta \le \alpha$. The latter combined with the triangular inequality yields

$$\|\boldsymbol{\theta}^+-\boldsymbol{\theta}_0\|_{\ell_2} \le \|\boldsymbol{\theta}^+-\boldsymbol{\theta}\|_{\ell_2} + \|\boldsymbol{\theta}_0-\boldsymbol{\theta}\|_{\ell_2} \le \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2} + \frac{\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}}{\zeta} \le R_\zeta,$$

concluding the proof of $\boldsymbol{\theta}^+ \in \mathcal{D}$. ■

The next lemma establishes the convergence to a global minima that lies in a minimally small local neighborhood under a Jacobian condition (9.10). The proof of this lemma is deferred to Section 9.2.1.

Lemma 9.6 *Suppose the Jacobian mapping associated with f obeys Assumption 1 over a ball of radius $R_\zeta := \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}$ around a point $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, that is $\mathcal{D} = \mathcal{B}\left(\boldsymbol{\theta}_0, \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$. Let λ be a scalar obeying $0 < \lambda \le 1$ and set $\zeta = (\lambda - \eta\beta^2/2)\alpha$. Also assume*

$$\boldsymbol{C}(\boldsymbol{\theta}) \ge \lambda \mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T \tag{9.10}$$

holds for all $\boldsymbol{\theta} \in \mathcal{P}\left(\boldsymbol{\theta}_0, \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$. Then, staring from $\boldsymbol{\theta}_0$ the GD iterates $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ with $\eta \le \frac{\lambda}{\beta^2}$ obey

$$\|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2}^2 \le \left(1 - \alpha^2\lambda\eta\right)^\tau \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}^2, \tag{9.11}$$

$$\zeta \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}. \tag{9.12}$$

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

$$\sum_{\tau=0}^{\infty} \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} \le \frac{\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\zeta}. \tag{9.13}$$

The next lemma shows that (9.10) indeed holds. We defer the proof of this lemma to Section 9.2.2.

Lemma 9.7 *Consider a point $\boldsymbol{\theta} \in \mathbb{R}^p$ and the result of a gradient update $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta\nabla\mathcal{L}(\boldsymbol{\theta})$ staring from $\boldsymbol{\theta}$. Suppose Assumption 1 and one of the following two statements hold over $\mathcal{D} = \mathcal{B}\left(\boldsymbol{\theta}_0, \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$ for a ζ obeying $0 \le \zeta \le \alpha$*

- *Assumption 2(a) holds over $\mathcal{D}$ and $\eta \le \frac{1}{\beta^2}$*
- *Assumption 2(b) holds over $\mathcal{D}$ and $\eta \le \frac{1}{\beta^2}\min\left(1, \frac{2(1-\lambda)\alpha^2}{L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}\right)$.*

Then for all $\boldsymbol{\theta} \in \mathcal{P}\left(\boldsymbol{\theta}_0, \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$,

$$\boldsymbol{C}(\boldsymbol{\theta}) := \mathcal{J}(\boldsymbol{\theta}^+, \boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T \ge \lambda \mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T.$$

With these lemmas in place we are now ready to prove Theorem 9.3.
Proof of Theorem 9.3: Set $\zeta = (\lambda - \eta\beta^2/2)\alpha$ and observe that

- Since assumptions of Theorem 9.3 subsume those of Lemma 9.7, for all $\boldsymbol{\theta} \in \mathcal{P}\left(\boldsymbol{\theta}_0, \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$, (9.10) holds i.e. we have that $\boldsymbol{C}(\boldsymbol{\theta}) \ge \lambda\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T$.
- Based on the above, the assumptions of Theorem 9.3 also subsume those of Lemma 9.6. Thus (9.11), (9.12), and (9.13) hold for all τ.

This completes the bounds (9.2), (9.3), and (9.4) of Theorem 9.3. The proofs of (9.5) and (9.6) follow immediately from (9.3) and (9.4) by noting that for any global optima (including the closest global optimum to $\boldsymbol{\theta}_0$ denoted by $\boldsymbol{\theta}^*$) we have

$$\begin{aligned}
\|\boldsymbol{y} - f(\boldsymbol{\theta}_0)\|_{\ell_2} &= \|f(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta}_0)\|_{\ell_2} \\
&= \left\|\int_0^1 \mathcal{J}^T\left(\boldsymbol{\theta}_0 + t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\right)(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)dt\right\|_{\ell_2} \\
&\le \sup_{0\le 1\le t} \|\mathcal{J}\left(\boldsymbol{\theta}_0 + t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)\right)\| \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_{\ell_2} \\
&\le \sup_{\boldsymbol{\theta}\in\mathcal{D}} \|\mathcal{J}(\boldsymbol{\theta})\| \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_{\ell_2} \\
&\le \beta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_{\ell_2}.
\end{aligned}$$

This concludes the proof of Theorem 9.3. All that remains is to prove Lemmas 9.6 and 9.7 which are the subject of the two sections below.

9.2.1. Proof of Lemma 9.6

We will prove this lemma by induction. Assume the claim holds until iteration τ. First, since (9.12) holds, applying Lemma 9.5 and using the facts that $\eta \le 1/\beta^2$ and $\zeta \le \alpha$, we can conclude that $\boldsymbol{\theta}_{\tau+1} \in \mathcal{D}$.

Next, we will simultaneously monitor how the distance to the initial parameter $\boldsymbol{\theta}_0$ ($\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2}$) and the Euclidean norm of the residual ($\|\boldsymbol{r}_\tau\|_{\ell_2}$) change from iteration τ to $\tau+1$. For the distance to initialization, using triangular inequality and the formula for the gradient we have

$$\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\|_{\ell_2} \le \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} = \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \eta\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2}. \tag{9.14}$$

For the norm of the residual using the fact that $\boldsymbol{C}(\boldsymbol{\theta}) \ge \lambda\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T$ (per assumption (9.10)) we have

$$\begin{aligned}
\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2 &\overset{(a)}{=} \|(\boldsymbol{I} - \eta\boldsymbol{C}(\boldsymbol{\theta}_\tau))\boldsymbol{r}_\tau\|_{\ell_2}^2,\\
&= \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - 2\eta\boldsymbol{r}_\tau^T\boldsymbol{C}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau + \eta^2\boldsymbol{r}_\tau^T\boldsymbol{C}(\boldsymbol{\theta}_\tau)^T\boldsymbol{C}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau,\\
&\overset{(b)}{\le} \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - 2\lambda\eta\boldsymbol{r}_\tau^T\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau + \eta^2\beta^2\boldsymbol{r}_\tau^T\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau,\\
&\overset{(c)}{\le} \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - (2\lambda - \eta\beta^2)\eta\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\|_{\ell_2}^2.
\end{aligned} \tag{9.15}$$

Here, (a) follows from (9.7), (b) from (9.10) and the upper bound on the spectral norm of the Jacobian, (c) and from merging the terms on the right hand side. Combining (9.15) with $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_\tau)) \ge \alpha$, and using $\eta \le \lambda/\beta^2$, we conclude that

$$\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2 \le \left(1 - \alpha^2(2\lambda - \eta\beta^2)\eta\right)\|\boldsymbol{r}_\tau\|_{\ell_2}^2 \le \left(1 - \lambda\alpha^2\eta\right)\|\boldsymbol{r}_\tau\|_{\ell_2}^2,$$

completing the proof of (9.11). For the remainder of discussion, denote $\gamma = (\lambda - \eta\beta^2/2)\eta$. γ is nonnegative due to upper bound on η and we have

$$\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2 \le \|\boldsymbol{r}_\tau\|_{\ell_2}^2 - 2\gamma\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\|_{\ell_2}^2.$$

We now turn our attention to proving (9.12). To this aim we start from (9.15) and complete the square to conclude that

$$\begin{aligned}
\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2 &= \left(\|\boldsymbol{r}_\tau\|_{\ell_2} - \gamma\frac{\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}}\right)^2 - \left(\gamma\frac{\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}}\right)^2,\\
&\le \left(\|\boldsymbol{r}_\tau\|_{\ell_2} - \gamma\frac{\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}}\right)^2.
\end{aligned} \tag{9.16}$$

Also note that using the upper bound on spectrum of $\mathcal{J}$ and $\gamma \le \lambda\eta \le \frac{1}{\beta^2}$ we have

$$\|\boldsymbol{r}_\tau\|_{\ell_2}^2 \ge \frac{1}{\beta^2}\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2 \ge \gamma\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2 \quad \Rightarrow \quad \|\boldsymbol{r}_\tau\|_{\ell_2} - \gamma\frac{\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}} \ge 0.$$

Thus, taking square root from both sides of (9.16) we reach the following identity for changes in the norm of residual

$$\|\boldsymbol{r}_{\tau+1}\|_{\ell_2} \le \|\boldsymbol{r}_\tau\|_{\ell_2} - \gamma\frac{\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}}. \tag{9.17}$$

To combine the identities (9.14) and (9.17) in such a way to yield our theorem we proceed by defining the potential/Lyapunov function below with $\zeta = \alpha\gamma/\eta$.

$$\mathcal{V}_\tau := \|\boldsymbol{r}_\tau\|_{\ell_2} + \zeta\sum_{t=0}^{\tau-1}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_{\ell_2}. \tag{9.18}$$

A unique feature of the $\mathcal{V}_\tau$ potential is that it is non-increasing. To see this note that using (9.17) we have

$$\begin{aligned}\frac{1}{\eta}\left(\mathcal{V}_{\tau+1}-\mathcal{V}_\tau\right) =&\frac{1}{\eta}\left(\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}-\|\boldsymbol{r}_\tau\|_{\ell_2}\right)+\frac{\zeta}{\eta}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2},\\ \overset{(a)}{=}&\frac{1}{\eta}\left(\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}-\|\boldsymbol{r}_\tau\|_{\ell_2}\right)+\zeta\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2},\\ \overset{(b)}{\le}&-\frac{\gamma}{\eta}\frac{\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}}+\zeta\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2},\\ =&\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}\left(\zeta-\frac{\gamma}{\eta}\frac{\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}}{\|\boldsymbol{r}_\tau\|_{\ell_2}}\right),\\ \overset{(c)}{\le}&\left\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\boldsymbol{r}_\tau\right\|_{\ell_2}\left(\zeta-\alpha\frac{\gamma}{\eta}\right),\\ =&\,0.\end{aligned}\tag{9.19}$$

Here, (a) follows from the gradient formula, (b) from (9.17), (c) from $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_\tau))\ge\alpha$, and (d) from $\zeta=\alpha\gamma/\eta$. Using this non-increasing property and triangle inequality over $(\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2})_{\tau\ge0}$ we can conclude that

$$\|\boldsymbol{r}_\tau\|_{\ell_2}+\zeta\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}\le\mathcal{V}_\tau\le\mathcal{V}_0=\|\boldsymbol{r}_0\|_{\ell_2},$$

proving (9.12).

Finally using the definition of $\mathcal{V}_\tau$ and its non-increasing property (9.19) we have

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2}\le\frac{\mathcal{V}_\infty}{\zeta}\le\frac{\mathcal{V}_0}{\zeta}=\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{\zeta},$$

concluding the proof of (9.13) and Lemma 9.6 when we substitute $\zeta=(\lambda-\eta\beta^2/2)\alpha$.

9.2.2. Proof of Lemma 9.7

First note that since $\boldsymbol{\theta}\in\mathcal{P}\left(\boldsymbol{\theta}_0,\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$, we have

$$\|\boldsymbol{y}-f(\boldsymbol{\theta})\|_{\ell_2}\le\|\boldsymbol{y}-f(\boldsymbol{\theta}_0)\|_{\ell_2}.\tag{9.20}$$

Second, applying Lemma 9.5, we also have $\boldsymbol{\theta}^+=\boldsymbol{\theta}-\eta\nabla\mathcal{L}(\boldsymbol{\theta})\in\mathcal{D}:=\mathcal{B}\left(\boldsymbol{\theta}_0,\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\zeta}\right)$. To prove

$$\boldsymbol{C}(\boldsymbol{\theta})\ge\lambda\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T,\tag{9.21}$$

we consider the two cases related to Assumption 2 separately.

If Assumption 2(a) holds then for any $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathcal{D}$ we have

$$\begin{aligned}\|\mathcal{J}(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1)-\mathcal{J}(\boldsymbol{\theta}_1)\|=&\left\|\int_0^1\left(\mathcal{J}\left(\boldsymbol{\theta}_1+t\left(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\right)\right)-\mathcal{J}(\boldsymbol{\theta}_1)\right)dt\right\|,\\ \le&\int_0^1\|\mathcal{J}\left(\boldsymbol{\theta}_1+t\left(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\right)\right)-\mathcal{J}(\boldsymbol{\theta}_1)\|\,dt,\\ \le&\int_0^1\frac{(1-\lambda)\alpha^2}{\beta}dt,\\ \le&\frac{(1-\lambda)\alpha^2}{\beta}.\end{aligned}$$

Thus for $\boldsymbol{\theta}, \boldsymbol{\theta}^+ \in \mathcal{D}$ we have

$$\begin{aligned}\left\|\left(\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})-\mathcal{J}(\boldsymbol{\theta})\right)\mathcal{J}(\boldsymbol{\theta})^T\right\| &\le \|\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})-\mathcal{J}(\boldsymbol{\theta})\|\,\|\mathcal{J}(\boldsymbol{\theta})\|\\ &\le \frac{(1-\lambda)\alpha^2}{\beta}\beta\\ &= (1-\lambda)\alpha^2\\ &\le (1-\lambda)\sigma_{\min}^2\left(\mathcal{J}(\boldsymbol{\theta})\right).\end{aligned}$$

Thus we have

$$\begin{aligned}\mathcal{C}(\boldsymbol{\theta}) =&\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T,\\ =&\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T-\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T+\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T,\\ \ge&\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T-\boldsymbol{I}_n\left\|\left(\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})-\mathcal{J}(\boldsymbol{\theta})\right)\mathcal{J}(\boldsymbol{\theta})^T\right\|,\\ \ge&\lambda\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T.\end{aligned}$$

This implies the desired bound (9.21).

Next, suppose Assumption 2(b) holds. Then, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{D}$ we have

$$\begin{aligned}\|\mathcal{J}(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1)-\mathcal{J}(\boldsymbol{\theta}_1)\| =&\left\|\int_0^1\left(\mathcal{J}\left(\boldsymbol{\theta}_1+t\left(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\right)\right)-\mathcal{J}(\boldsymbol{\theta}_1)\right)dt\right\|,\\ \le&\int_0^1\left\|\mathcal{J}\left(\boldsymbol{\theta}_1+t\left(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\right)\right)-\mathcal{J}(\boldsymbol{\theta}_1)\right\|dt,\\ \le&\int_0^1 tL\left\|\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\right\|_{\ell_2}dt,\\ \le&\frac{L}{2}\left\|\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\right\|_{\ell_2}.\end{aligned}\tag{9.22}$$

Thus, for $\eta \le \frac{2(1-\lambda)\alpha^2}{L\beta^2\|\boldsymbol{r}_0\|_{\ell_2}}$,

$$\|\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})-\mathcal{J}(\boldsymbol{\theta})\| \le \frac{L}{2}\|\boldsymbol{\theta}^+-\boldsymbol{\theta}\|_{\ell_2} = \frac{\eta L}{2}\left\|\mathcal{J}^T(\boldsymbol{\theta})\left(f(\boldsymbol{\theta})-\boldsymbol{y}\right)\right\|_{\ell_2} \le \frac{\eta\beta L}{2}\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2} \overset{(9.20)}{\le} \frac{\eta\beta L}{2}\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2} \le \frac{(1-\lambda)\alpha^2}{\beta},$$

Repeating the previous argument (with Assumption 2(a)), we again conclude with (9.21).

9.3. Lower bounds proofs (Theorem 2.4)

We begin by proving (2.15). To show this we first use the upper bound on the Jacobian matrix to prove that the nonlinear mapping is Lipschitz. To this aim note that

$$f(\boldsymbol{\theta})-f(\boldsymbol{\theta}_0) = \int_0^1\mathcal{J}\left(\boldsymbol{\theta}_0+t(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)dt = \mathcal{J}(\boldsymbol{\theta},\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0).$$

Hence,

$$\|f(\boldsymbol{\theta})-f(\boldsymbol{\theta}_0)\|_{\ell_2} \le \|\mathcal{J}(\boldsymbol{\theta},\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_{\ell_2} \le \beta\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2},$$

completing the proof of the Lipschitz property. This Lipschitz property combined with the triangular inequality allows us to conclude

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

completing the proof of (2.15).

Next we turn out attention to providing the counter examples. Consider a least squares problem where the loss is equal to $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2$ and the data matrix $\boldsymbol{X}$ has orthogonal rows. Suppose the first row $\boldsymbol{x}_1$ has the smallest ℓ_2 norm which is α and the last row $\boldsymbol{x}_n$ has the largest ℓ_2 norm equal to β. We also set the labels to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^\star$ where $\boldsymbol{\theta}^\star = \gamma \boldsymbol{x}_1 / \|\boldsymbol{x}_1\|_{\ell_2}$ with $\gamma = \beta/\alpha$. For this linear regression problem, the Jacobian is equal to $\boldsymbol{X}$ and since the matrix is orthogonal α, β are the minimum/maximum singular values of the Jacobian.

For any $\alpha, \beta \geq 0$ obeying $\alpha \leq \beta$ and any $\boldsymbol{\theta}$, we have

$$\|\boldsymbol{y} - f(\boldsymbol{\theta})\|_{\ell_2} = \|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|_{\ell_2} \geq \|\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_{\ell_2} \geq \|\boldsymbol{x}_1^T(\boldsymbol{\theta} - \boldsymbol{\theta}_\star)\|_{\ell_2} \geq \|\boldsymbol{x}_1^T\boldsymbol{\theta}_\star\|_{\ell_2} - \|\boldsymbol{x}_1^T\boldsymbol{\theta}\|_{\ell_2} \geq \alpha(\gamma - \|\boldsymbol{\theta}\|_{\ell_2}).$$

This yields $\|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2} + \alpha\|\boldsymbol{\theta}\|_{\ell_2} \geq \|\boldsymbol{y}\|_{\ell_2} = \gamma\alpha$ which in turns implies (2.16) with $\boldsymbol{\theta}_0 = \boldsymbol{0}$.

To show (2.17), we set the labels to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^\star$ where $\boldsymbol{\theta}^\star = \gamma \frac{\boldsymbol{x}_n}{\|\boldsymbol{x}_n\|_{\ell_2}}$. In this case, gradient iteration starting from $\boldsymbol{\theta}_0 = \boldsymbol{0}$ is simply

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau + \eta \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\tau).$$

If $\boldsymbol{\theta}_\tau \subset \text{span}(\boldsymbol{x}_n)$, it is clear that $\boldsymbol{\theta}_{\tau+1} \subset \text{span}(\boldsymbol{x}_n)$ as well as $\boldsymbol{X}^T\boldsymbol{y} \subset \text{span}(\boldsymbol{x}_n)$. Since $\boldsymbol{\theta}_0 = 0$, this implies that gradient descent recursion is one dimensional over $\boldsymbol{x}_n$ i.e. $\boldsymbol{\theta}_\tau = \frac{\boldsymbol{x}_n}{\|\boldsymbol{x}_n\|_{\ell_2}}\theta_\tau$ with θ_τ a scalar obeying the recursion,

$$\theta_{\tau+1} = \theta_\tau + \eta\beta^2(\theta^\star - \theta_\tau).$$

If $\eta \leq 1/\beta^2$, all iterations satisfy $0 \leq \theta_\tau \leq \theta^\star = \gamma$. On the other hand, the misfit in each iteration obeys

$$\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2} = \|\boldsymbol{X}(\boldsymbol{\theta}^\star - \boldsymbol{\theta}_\tau)\|_{\ell_2} = \beta\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_\tau\|_{\ell_2} = \beta|\theta^\star - \theta_\tau| = \beta(\theta^\star - \theta_\tau).$$

The last two identities imply $\|\boldsymbol{y} - f(\boldsymbol{\theta}_\tau)\|_{\ell_2} + \beta\|\boldsymbol{\theta}_\tau\|_{\ell_2} = \beta\gamma = \|\boldsymbol{y}\|_{\ell_2}$ completing the proof of (2.17).

9.4. SGD proofs (Proof of Theorem 3.1)

9.4.1. Roadmap of SGD proof

We begin our SGD analysis by writing the SGD iterates in terms of the Jacobian matrix. To this aim define the matrix $\mathcal{J}(\boldsymbol{\theta}_\tau; \gamma_\tau)$ which keeps the γ_τ-th row of $\mathcal{J}(\boldsymbol{\theta}_\tau)$ and sets the remaining rows to zero. We note that

$$G(\boldsymbol{\theta}_\tau; \gamma_\tau) = \mathcal{J}(\boldsymbol{\theta}_\tau; \gamma_\tau)^T\left(f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\right) \quad \text{and} \quad \mathbb{E}[\mathcal{J}(\boldsymbol{\theta}_\tau; \gamma_\tau)] = \frac{1}{n}\mathcal{J}(\boldsymbol{\theta}_\tau). \tag{9.23}$$

Also define the matrix $\boldsymbol{C}(\boldsymbol{\theta}_\tau; \gamma_\tau) = \mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau; \gamma_\tau)^T \in \mathbb{R}^{n\times n}$ which can be thought of as a stochastic version of $\boldsymbol{C}(\boldsymbol{\theta}_\tau)$ obeying

$$\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta}_\tau; \gamma_\tau)] = \frac{1}{n}\boldsymbol{C}(\boldsymbol{\theta}_\tau).$$

Similar to the GD proof we begin by noting that the residual $\boldsymbol{r}_\tau$ satisfies the recursion

$$\begin{aligned}
\boldsymbol{r}_{\tau+1} =& \boldsymbol{r}_\tau - f(\boldsymbol{\theta}_\tau) + f(\boldsymbol{\theta}_{\tau+1}),\\
\overset{(a)}{=}& \boldsymbol{r}_\tau + \mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau)\left(\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\right),\\
\overset{(b)}{=}& \boldsymbol{r}_\tau - \eta\mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau)G(\boldsymbol{\theta}_\tau; \gamma_\tau),\\
\overset{(c)}{=}& \left(\boldsymbol{I}_n - \eta\boldsymbol{C}(\boldsymbol{\theta}_\tau; \gamma_\tau)\right)\boldsymbol{r}_\tau.
\end{aligned} \tag{9.24}$$

Here, (a) follows from the fundamental rule of calculus, (b) from the stochastic update rule, and (c) from combining the form of the stochastic gradient in (9.23) with the definition of $\boldsymbol{C}(\boldsymbol{\theta}_\tau; \gamma_\tau)$.

Given that $\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta}_\tau; \gamma_\tau)] = \boldsymbol{C}(\boldsymbol{\theta}_\tau)/n$, similar to the GD proof we can show that under the two assumptions $\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta}_\tau; \gamma_\tau)]$ is positive-definite and thus with a sufficiently small learning rate η this implies linear convergence of the expected residual via (9.24) as long as $\boldsymbol{\theta}_i \in \mathcal{D}$.

It is completely unclear if SGD stays inside a neighborhood around the initial model to ensure the on average convergence argument discussed above is useful. We will develop a novel martingale-based argument to show that SGD does indeed stay in this local neighborhood. We briefly discuss the intuition behind this approach here. Since SGD is inherently random, ideally, we would like to show that, a variant of (2.4) holds. Specifically, define

$$\mathcal{V}_\tau = c\alpha\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2}, \tag{9.25}$$

we wish to show that $\mathcal{V}_\tau$ is bounded. One approach to do this is to show $\mathbb{E}[\mathcal{V}_\tau] \le \mathcal{V}_{\tau-1}$ where the expectation is over the τ'th SGD step given first $\tau - 1$ steps. If this holds, $\mathcal{V}_\tau$ is a *supermartingale* with respect to the filtration generated by random SGD steps. This allows us to utilize martingale maximal inequality (Revuz & Yor, 2013) which bounds the supremum of $\mathcal{V}_\tau$ via a Markov-like inequality

$$\mathbb{P}(\sup_{\tau\ge 0} \mathcal{V}_\tau \ge C\,\mathbb{E}[\mathcal{V}_0]) \le \frac{1}{C}.$$

This immediately establishes that $\mathcal{V}_\tau$ is uniformly bounded by $C\,\mathbb{E}[\mathcal{V}_0]$ and thus $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \le \frac{C\,\mathbb{E}[\mathcal{V}_0]}{c\alpha}$, hence $\boldsymbol{\theta}_\tau$ doesn't leave this neighborhood. However, unfortunately such a strategy does not work and a more nuanced argument is required. In particular, we need to overcome two challenges:

- The first challenge is that (9.4.1) is not a super martingale for reasonably large values of c. However, large values of c are desirable as they yield a small convergence radius (e.g. $c = 1/4$ in (2.4)). We overcome this challenge by proposing a new potential function which tracks distances to multiple anchor points around $\boldsymbol{\theta}_0$ rather than only $\boldsymbol{\theta}_0$. Denoting these anchor points by $\{\boldsymbol{p}_\ell\}_{\ell=1}^K$, we utilize the potential

  $$\mathcal{V}_\tau := \mathcal{V}(\boldsymbol{\theta}_\tau) := 12\|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2} + \frac{\alpha}{K}\sum_{\ell=1}^K \|\boldsymbol{\theta}_\tau - \boldsymbol{p}_\ell\|_{\ell_2}.$$

  Figure 4 provides a pictorial illustration of this potential function.

- The second challenge is that the optimization landscape is assumed to have nice properties only over a small neighborhood $\mathcal{D}$ around the initial point. Hence, the super martingale inequality $\mathbb{E}[\mathcal{V}_\tau] \le \mathcal{V}_{\tau-1}$ applies only if the current and next iterate is over $\mathcal{D}$ and optimization essentially fails if we step outside. We overcome this by showing that the chance that SGD iterates exit this neighborhood is small using martingale stopping time arguments. The latter argument is inspired by/adapted from the work of Tan and Vershynin (Tan & Vershynin, 2017) in the context of phase retrieval.

The outline of this Section is as follows. We show in Section 9.4.2 that from one SGD iterate to the next the misfit decreases in expectation. Then in Section 9.4.3 show that from one SGD iterate to the next the average distance to the chosen points $\{\boldsymbol{p}_\ell\}_{\ell=1}^K$ do not increase by a significant amount. We then combine the latter two results in Section 9.4.4 to formally show that the potential $\mathcal{V}(\boldsymbol{\theta}_\tau)$ is indeed a supermartingale. Next, in section we deploy a martingale stopping time argument to show that with high probability the SGD iterates stay inside a neighborhood around the initial model. Finally, we put together all of these different arguments to complete the proof of Theorem 3.1 in Section 9.4.6.

9.4.2. Decrease of the expected misfit

In this section we will show that under the assumption that SGD iterates always remain close to the initialization, the expected value of the norm of the residual will decrease in each iteration. Concretely, in this section we prove the following lemma.

Lemma 9.8 *Consider a point $\boldsymbol{\theta} \in \mathbb{R}^p$ and the result of a stochastic gradient update $\boldsymbol{\theta}^+ := \boldsymbol{\theta} - \eta G(\boldsymbol{\theta};\gamma) = \boldsymbol{\theta} - \eta\left(f(\boldsymbol{x}_\gamma;\boldsymbol{\theta}) - y_\gamma\right)\nabla f(\boldsymbol{x}_\gamma;\boldsymbol{\theta})$ staring from $\boldsymbol{\theta}$ with the index γ chosen uniformly at random from $\{1,2,\ldots,n\}$. Also consider the set*

$$\mathcal{B}(\nu) = \mathcal{B}\left(\boldsymbol{\theta}_0, \nu\frac{\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\alpha}\right) \bigcap \left\{\boldsymbol{\theta} \in \mathbb{R}^p \middle| \|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2} \le \frac{2\nu}{3}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}\right\}, \tag{9.26}$$

Assume $\boldsymbol{\theta} \in \mathcal{D}' := \mathcal{B}(\nu/2)$ with ν a scalar obeying $\nu \ge 3$. Also assume the Jacobian associated with f obeys Assumption 1 over the set $\mathcal{D} := \mathcal{B}(\nu)$ and the rows of the Jacobian have bounded Euclidean norm over this set, that is

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

Also assume

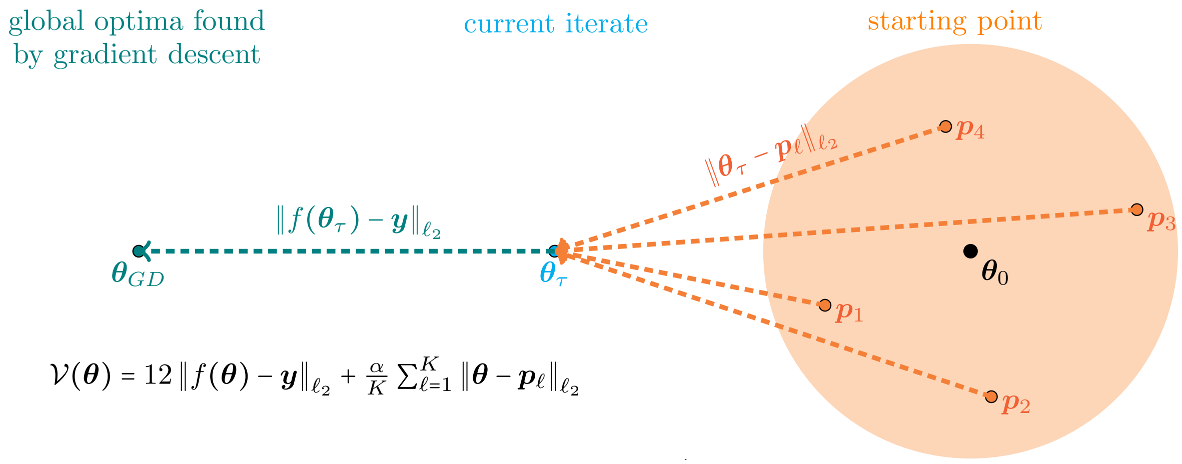

global optima found
by gradient descent
current iterate
starting point
$\|\theta_\tau - p_\ell\|_{\ell_2}$
p_4
p_3
$\|f(\theta_\tau) - y\|_{\ell_2}$
θ_{GD}
θ_τ
θ_0
p_1
p_2
$\mathcal{V}(\theta) = 12\,\|f(\theta) - y\|_{\ell_2} + \frac{\alpha}{K}\sum_{\ell=1}^{K}\|\theta - p_\ell\|_{\ell_2}$

Figure 4. SGD potential function is similar to the gradient descent potential (2.4). It provides a balance between misfit error and distance to the initial point. However, to show that this potential is non-increasing, unlike gradient descent, we keep track of distances to multiple points around the initial point $\boldsymbol{\theta}_0$. This smooths out the potential function and guarantees the desired non-increasing property. Intuitively, the misfit ($\|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2}$) can be viewed as a proxy for distance to the global minima ($\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_{GD}\|_{\ell_2}$) as illustrated.

- *Assumption 2(a) holds over $\mathcal{D}$ and $\eta \le \frac{\alpha^2}{2\beta^2 B^2}$.*
- *Assumption 2(b) holds over $\mathcal{D}$ and $\eta \le \frac{1}{2\beta B} \cdot \min\left(\frac{\alpha^2}{B\beta}, \frac{3\alpha^2}{\nu L \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}\right)$.*

Then,

$$\mathbb{E}[\|f(\boldsymbol{\theta}^+) - \boldsymbol{y}\|_{\ell_2}] \le \|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2} - \frac{\eta}{4n}\frac{\|\mathcal{J}^T(\boldsymbol{\theta})(f(\boldsymbol{\theta}) - \boldsymbol{y})\|_{\ell_2}^2}{\|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2}}, \tag{9.27}$$

$$\mathbb{E}\left[\|f(\boldsymbol{\theta}^+) - \boldsymbol{y}\|_{\ell_2}^2\right] \le \left(1 - \frac{\eta\alpha^2}{2n}\right)^\tau \|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2}^2. \tag{9.28}$$

For simplicity of exposition of the proof of this lemma we define $r(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) - \boldsymbol{y}$ and $r(\boldsymbol{\theta}^+) = f(\boldsymbol{\theta}^+) - \boldsymbol{y}$. We prove the lemma in three steps.

- **Step I:** We show that as long as $\eta \le \frac{1}{\beta B}$, then $\boldsymbol{\theta}^+ \in \mathcal{D}$.
- **Step II:** We prove that the matrix $\boldsymbol{C}(\boldsymbol{\theta}) := \mathcal{J}(\boldsymbol{\theta}^+, \boldsymbol{\theta})\mathcal{J}^T(\boldsymbol{\theta})$ obeys

$$\boldsymbol{C}(\boldsymbol{\theta}) \succeq \frac{1}{2}\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T. \tag{9.29}$$

- **Step III:** We use Step I and II to show the inequalities (9.27) and (9.28) which are equivalent to

$$\mathbb{E}[\|\boldsymbol{r}(\boldsymbol{\theta}^+)\|_{\ell_2}] \le \|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} - \frac{\eta}{4n}\frac{\|\mathcal{J}^T(\boldsymbol{\theta})(f(\boldsymbol{\theta}) - \boldsymbol{y})\|_{\ell_2}^2}{\|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2}}, \tag{9.30}$$

$$\mathbb{E}[\|\boldsymbol{r}(\boldsymbol{\theta}^+)\|_{\ell_2}] \le \left(1 - \frac{\eta\alpha^2}{2n}\right)\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2. \tag{9.31}$$

Step I: We begin this step by noting that

$$\|G(\boldsymbol{\theta};\gamma)\|_{\ell_2} \le \max_{1\le i\le n}\|\nabla f(\boldsymbol{x}_i;\boldsymbol{\theta})\|_{\ell_2}\,|f(\boldsymbol{x}_i;\boldsymbol{\theta}) - \boldsymbol{y}_i| \le B\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}. \tag{9.32}$$

Using this inequality we can conclude that

$$\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_{\ell_2} \le \eta \|G(\boldsymbol{\theta};\gamma)\|_{\ell_2} \overset{(a)}{\le} \eta B \|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} \overset{(b)}{\le} \frac{\eta\nu B}{3} \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} \overset{(c)}{\le} \frac{1}{3}\frac{\nu}{\beta} \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}. \tag{9.33}$$

Here, (a) follows from (9.32), (b) from the fact that $\boldsymbol{\theta} \in \mathcal{D}' := \mathcal{B}(\nu/2)$, and (c) from $\eta \le \frac{1}{\beta B}$. Furthermore, the simple fact that $\alpha \le \beta$ implies that

$$\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_{\ell_2} \le \frac{1}{3}\frac{\nu}{\alpha} \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}. \tag{9.34}$$

Next note that

$$\begin{aligned}\|r(\boldsymbol{\theta}^+)\|_{\ell_2} &\le \|r(\boldsymbol{\theta})\|_{\ell_2} + \|r(\boldsymbol{\theta}^+) - r(\boldsymbol{\theta})\|_{\ell_2}, \\ &= \|r(\boldsymbol{\theta})\|_{\ell_2} + \|\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})(\boldsymbol{\theta}^+ - \boldsymbol{\theta})\|_{\ell_2}, \\ &\le \|r(\boldsymbol{\theta})\|_{\ell_2} + \|\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})\| \, \|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_{\ell_2}, \\ &\overset{(a)}{\le} \|r(\boldsymbol{\theta})\|_{\ell_2} + \frac{\nu}{3} \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}, \\ &\overset{(b)}{\le} \frac{2}{3}\nu \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}.\end{aligned} \tag{9.35}$$

Here, (a) follows from (9.33) and the fact that $\|\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})\| \le \beta$ and (b) follows from the fact that $\boldsymbol{\theta} \in \mathcal{D}' := \mathcal{B}(\nu/2)$. Combining (9.34) and (9.35) we conclude that $\boldsymbol{\theta}^+ \in \mathcal{D} := \mathcal{B}(\nu)$.

Step II: The proof of (9.29) is very similar to the proof of Lemma 9.7 with $\lambda = 1/2$. In particular, under Assumption 2(a) the exact same argument yields (9.29). To show the result under Assumption 2(b) we combine (9.22) from the proof of Lemma 9.7, (9.32), and $\boldsymbol{\theta} \in \mathcal{D}' = \mathcal{B}(\nu/2)$ to conclude that

$$\|\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta})\| \le \frac{L}{2}\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_{\ell_2} \le \frac{\eta BL}{2}\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} \le \frac{\eta\nu BL}{3}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} \le \frac{\alpha^2}{2\beta},$$

where in the last inequality we use the fact that $\eta \le \frac{3}{2}\frac{\alpha^2}{\nu\beta BL\|\boldsymbol{r}_0\|_{\ell_2}}$. The remainder of the proof of (9.29) is exactly the same as the proof of Lemma 9.7.

Step III: From the arguments of Steps I and II we know that

(i) $\boldsymbol{\theta}^+ \in \mathcal{D}$,

(ii) $\|\boldsymbol{C}(\boldsymbol{\theta})\| \le \beta^2$ and $\|\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})\| \le \beta$,

(iii) $\boldsymbol{C}(\boldsymbol{\theta}) \ge \frac{1}{2}\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T$.

Using (ii) $\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta}^+,\boldsymbol{\theta})^T \le \beta^2 \boldsymbol{I}_n$, so that

$$\boldsymbol{C}(\boldsymbol{\theta};\gamma)^T\boldsymbol{C}(\boldsymbol{\theta};\gamma) \le \beta^2\mathcal{J}(\boldsymbol{\theta};\gamma)\mathcal{J}(\boldsymbol{\theta};\gamma)^T.$$

Furthermore, $\mathcal{J}(\boldsymbol{\theta};\gamma)\mathcal{J}(\boldsymbol{\theta};\gamma)^T$ is a diagonal matrix with a single nonzero entry which is bounded by B^2. Thus,

$$\mathbb{E}\left[\boldsymbol{C}(\boldsymbol{\theta};\gamma)^T\boldsymbol{C}(\boldsymbol{\theta};\gamma)\right] \le \frac{\beta^2 B^2}{n}. \tag{9.36}$$

Also the fact that $\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta};\gamma)] = \frac{1}{n}\boldsymbol{C}(\boldsymbol{\theta})$ (also noted in Section 9.4.1) together with (iii) allows us to conclude that

$$\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta};\gamma)] = \frac{1}{n}\boldsymbol{C}(\boldsymbol{\theta}) \ge \frac{1}{2n}\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T. \tag{9.37}$$

Using the latter two inequalities allows us to conclude

$$\begin{aligned}\eta \boldsymbol{r}(\boldsymbol{\theta})^T\,\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta};\gamma)^T\boldsymbol{C}(\boldsymbol{\theta};\gamma)]\boldsymbol{r}(\boldsymbol{\theta}) &\overset{(a)}{\leq} \frac{\eta\beta^2B^2}{n}\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2,\\ &\overset{(b)}{\leq} \frac{\alpha^2}{2n}\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2,\\ &\overset{(c)}{\leq} \frac{1}{2n}\boldsymbol{r}(\boldsymbol{\theta})^T\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r}(\boldsymbol{\theta}),\\ &\overset{(d)}{\leq} \boldsymbol{r}(\boldsymbol{\theta})^T\,\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta};\gamma)]\boldsymbol{r}(\boldsymbol{\theta}).\end{aligned} \tag{9.38}$$

Here, (a) follows from (9.36), (b) from the fact that the step size obeys $\eta \leq \frac{\alpha^2}{2B^2\beta^2}$, (c) from $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta})) \geq \alpha$, and (d) from (9.37). These inequalities allow us to conclude

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{r}(\boldsymbol{\theta}^+)\|_{\ell_2}^2] &\overset{(a)}{\leq} \boldsymbol{r}(\boldsymbol{\theta})^T\left(\boldsymbol{I}_n - 2\eta\,\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta};\gamma)] + \eta^2\,\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta};\gamma)^T\boldsymbol{C}(\boldsymbol{\theta};\gamma)]\right)\boldsymbol{r}(\boldsymbol{\theta}),\\ &\overset{(b)}{\leq} \boldsymbol{r}(\boldsymbol{\theta})^T\left(\boldsymbol{I}_n - \eta\,\mathbb{E}[\boldsymbol{C}(\boldsymbol{\theta};\gamma)]\right)\boldsymbol{r}(\boldsymbol{\theta}),\\ &\overset{(c)}{\leq} \boldsymbol{r}(\boldsymbol{\theta})^T\left(\boldsymbol{I}_n - \frac{\eta}{2n}\mathcal{J}(\boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T\right)\boldsymbol{r}(\boldsymbol{\theta}),\\ &= \|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2 - \frac{\eta}{2n}\|\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2,\\ &\overset{(d)}{\leq} \left(\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} - \frac{\eta}{4n}\frac{\left\|\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r}(\boldsymbol{\theta})\right\|_{\ell_2}^2}{\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}}\right)^2.\end{aligned} \tag{9.39}$$

Here, (a) follows from the calculation in (9.24) applied to $\boldsymbol{r}(\boldsymbol{\theta})$ and $\boldsymbol{r}(\boldsymbol{\theta}^+)$, (b) from (9.38), (c) from (9.37), and (d) from completing the square. Finally, note that using the upper bound on the spectrum of the Jacobian and the fact that $\eta \leq \frac{\alpha^2}{2\beta^2B^2} \leq \frac{1}{2\beta^2}$[8] we have

$$\frac{\eta}{4n}\left\|\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r}(\boldsymbol{\theta})\right\|_{\ell_2}^2 \leq \eta\frac{\beta^2}{4n}\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2 \leq \|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2,$$

so that the term inside the parentheses of right-hand sided of (9.39) is positive. Consequently, combining Jensen's inequality with the square root of both sides of (9.39) yields

$$\mathbb{E}[\|\boldsymbol{r}(\boldsymbol{\theta}^+)\|_{\ell_2}] \leq \sqrt{\mathbb{E}[\|\boldsymbol{r}(\boldsymbol{\theta}^+)\|_{\ell_2}^2]} \leq \|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} - \frac{\eta}{4n}\frac{\left\|\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r}(\boldsymbol{\theta})\right\|_{\ell_2}^2}{\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}},$$

concluding the proof (9.30). To prove (9.31) we use the penultimate inequality from (9.39) together with the fact that $\sigma_{\min}\left(\mathcal{J}(\boldsymbol{\theta})\right) \geq \alpha$ to conclude that

$$\mathbb{E}[\|\boldsymbol{r}(\boldsymbol{\theta}^+)\|_{\ell_2}^2] \leq \|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2 - \frac{\eta}{2n}\|\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2 \leq \left(1 - \frac{\eta\alpha^2}{2n}\right)\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2,$$

completing the proof of (9.31).

9.4.3. Bounding the increase of expected average distance to anchor points

In this section we will show that under the assumption that SGD iterates always remain close to the initialization, the expected value of the average distance to the anchor points will not significantly increase in each iteration. Specifically, the anchor points we pick are an ϵ cover of the neighborhood of the initialization denoted by $\mathcal{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_K\}$. and we monitor the following average distance

$$d_{\mathcal{P}}(\boldsymbol{\theta}) := \frac{1}{K}\sum_{\ell=1}^{K}\|\boldsymbol{\theta} - \boldsymbol{p}_\ell\|_{\ell_2}. \tag{9.40}$$

Concretely, in this section we prove the following lemma.

[8] Note that $\alpha \leq B$.

Lemma 9.9 *Consider the setting and assumptions of Lemma 9.8. Also assume $\eta \le \frac{3}{\nu B^2}$. Furthermore, fix $K \ge \sqrt{n}\frac{\beta}{\alpha}$ and let $\mathcal{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_K\}$ be an $\epsilon := \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ packing of a ball of radius $R_p := 1.25\left(\frac{\beta}{\alpha}\right)^{1/p}\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ around $\boldsymbol{\theta}_0$ so that pairwise distances in this set are at least ϵ apart.*[9] *Then, for $d_{\mathcal{P}}$ given by (9.40) we have*

$$\mathbb{E}[d_{\mathcal{P}}(\boldsymbol{\theta}^+)] \le d_{\mathcal{P}}(\boldsymbol{\theta}) + \frac{3\eta}{n}\|\mathcal{J}^T(\boldsymbol{\theta})\left(f(\boldsymbol{\theta}) - \boldsymbol{y}\right)\|_{\ell_2}. \tag{9.41}$$

For simplicity of exposition of the proof of this lemma we define $r(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) - \boldsymbol{y}$ and $r(\boldsymbol{\theta}^+) = f(\boldsymbol{\theta}^+) - \boldsymbol{y}$. We start the proof by monitoring the evolution of the parameter vector with respect to a particular reference point $\boldsymbol{p} \in \mathcal{P}$. In particular define $\boldsymbol{w} = \boldsymbol{\theta} - \boldsymbol{p}$ and note that $\boldsymbol{w}^+ = \boldsymbol{\theta} - \boldsymbol{p} = \boldsymbol{w} - \eta\mathcal{J}^T(\boldsymbol{\theta};\gamma)\boldsymbol{r}(\boldsymbol{\theta})$ and $\mathbb{E}[\mathcal{J}(\boldsymbol{\theta};\gamma)] = \mathcal{J}(\boldsymbol{\theta})/n$. Thus,

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{w}^+\|_{\ell_2}^2] &= \mathbb{E}[\|\boldsymbol{w}\|_{\ell_2}^2 - 2\eta\boldsymbol{w}^T\mathcal{J}^T(\boldsymbol{\theta};\gamma)\boldsymbol{r}(\boldsymbol{\theta}) + \eta^2\|\mathcal{J}^T(\boldsymbol{\theta};\gamma)\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2],\\
&= \|\boldsymbol{w}\|_{\ell_2}^2 - 2\frac{\eta}{n}\boldsymbol{w}^T\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta}) + \eta^2\,\mathbb{E}[\|\mathcal{J}^T(\boldsymbol{\theta};\gamma)\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2],\\
&\le \|\boldsymbol{w}\|_{\ell_2}^2 + \frac{2}{n}\eta\|\boldsymbol{w}\|_{\ell_2}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} + \frac{\eta^2}{n}B^2\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2, &(9.42)\\
&\le \left(\|\boldsymbol{w}\|_{\ell_2} + \frac{2\eta}{n}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}\right)^2 + \frac{\eta}{n}\left(\eta B^2\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2 - 2\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}\|\boldsymbol{w}\|_{\ell_2}\right). &(9.43)
\end{aligned}$$

Using (9.42) and $\|\mathcal{J}^T(\boldsymbol{\theta})\| \le \beta$, we also have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{w}^+\|_{\ell_2}] \le&\sqrt{\mathbb{E}[\|\boldsymbol{w}^+\|_{\ell_2}^2]},\\
\le&\left(\|\boldsymbol{w}\|_{\ell_2}^2 + \frac{2}{n}\eta\|\boldsymbol{w}\|_{\ell_2}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} + \frac{\eta^2}{n}B^2\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2\right)^{1/2}\\
\le&\left(\|\boldsymbol{w}\|_{\ell_2}^2 + \frac{2\beta}{n}\eta\|\boldsymbol{w}\|_{\ell_2}\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} + \frac{\eta^2}{n}\beta^2\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2\right)^{1/2}\\
\le&\|\boldsymbol{w}\|_{\ell_2} + \frac{\eta}{\sqrt{n}}\beta\,\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}. &(9.44)
\end{aligned}$$

We also prove the following simple lemma.

Lemma 9.10 *If $\|\boldsymbol{w}\|_{\ell_2} \ge \epsilon/2$, then $\eta B^2\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}^2 - 2\|\mathcal{J}(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}\|\boldsymbol{w}\|_{\ell_2} \le 0$.*

Proof Using the assumption $\boldsymbol{\theta} \in \mathcal{B}(\nu/2)$, we have

$$\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} \le \frac{\nu}{3}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}. \tag{9.45}$$

Consequently, using $\eta \le \frac{3}{B^2\nu}$ and $\sigma_{\min}\left(\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\right) \ge \alpha\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}$, we have

$$\eta B^2\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} \le \frac{\eta B^2\nu}{3}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} \le \|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} = 2\alpha\frac{\epsilon}{2} \le 2\alpha\|\boldsymbol{w}\|_{\ell_2}.$$

■

Hence, the lemma above combined with (9.43) implies that if $\|\boldsymbol{w}\|_{\ell_2} \ge \varepsilon/2$

$$\mathbb{E}[\|\boldsymbol{w}^+\|_{\ell_2}] \le \sqrt{\mathbb{E}[\|\boldsymbol{w}^+\|_{\ell_2}^2]} \le \|\boldsymbol{w}\|_{\ell_2} + \frac{2\eta}{n}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}. \tag{9.46}$$

Combining (9.44) and (9.46), we conclude that

$$\mathbb{E}[\|\boldsymbol{w}^+\|_{\ell_2}] \le \begin{cases}\|\boldsymbol{w}\|_{\ell_2} + \frac{2\eta}{n}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} & \text{if} \quad \|\boldsymbol{w}\|_{\ell_2} \ge \frac{\epsilon}{2}\\ \|\boldsymbol{w}\|_{\ell_2} + \frac{\eta}{\sqrt{n}}\beta\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} & \text{otherwise}\end{cases}. \tag{9.47}$$

[9] Classical results guarantee that, we can find a $(R_p/\varepsilon)^p$ ϵ-packing of an R_p-ball. In our case using the fact that $p \ge n$ this reduces to $\left(1.25\left(\frac{\beta}{\alpha}\right)^{1/p}\right)^p \ge \left(\frac{5}{4}\right)^p\frac{\beta}{\alpha} \ge \sqrt{n}\frac{\beta}{\alpha} \ge K$ so that such a packing is possible.

Now define $\boldsymbol{w}_\ell := \boldsymbol{\theta} - \boldsymbol{p}_\ell$ as the difference between the parameter and the ℓth anchor point. Now observe that out of all vectors $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K\}$, at most one of them can satisfy $\|\boldsymbol{w}_\ell\|_{\ell_2} \le \frac{\epsilon}{2}$ due to the packing property. Specifically, if $\|\boldsymbol{w}_\ell\|_{\ell_2} \le \frac{\epsilon}{2}$, then for any $\widetilde{\ell} \ne \ell$ we have

$$\|\boldsymbol{w}_{\widetilde{\ell}}\|_{\ell_2} = \|\boldsymbol{p}_\ell - \boldsymbol{p}_{\widetilde{\ell}}\|_{\ell_2} - \|\boldsymbol{w}_\ell\|_{\ell_2} \ge \frac{\epsilon}{2}.$$

Hence, at least $K-1$ of $\boldsymbol{w}_\ell$ satisfies first line and at most 1 satisfies the second line of (9.47). Next, note that

$$\sqrt{n}\beta\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} \le \sqrt{n}\frac{\beta}{\alpha}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} \le K\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}.$$

Using the latter two identities we conclude that

$$\begin{aligned}\mathbb{E}\Big[\sum_{\ell=1}^K \|\boldsymbol{w}_\ell^+\|_{\ell_2}\Big] \le& \sum_{\ell=1}^K \|\boldsymbol{w}_\ell\|_{\ell_2} + (K-1)\frac{2\eta}{n}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2} + \frac{\eta}{\sqrt{n}}\beta\|\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}\\ \le& \sum_{\ell=1}^K \|\boldsymbol{w}_\ell\|_{\ell_2} + \frac{3K\eta}{n}\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})\|_{\ell_2}.\end{aligned}$$

Dividing both sides by K completes the proof of (9.41).

9.4.4. Shortest path potential is a supermartingale

In this section we show that the shortest path potential

$$\mathcal{V}_\tau := \mathcal{V}(\boldsymbol{\theta}_\tau) := 12\|f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\|_{\ell_2} + \frac{\alpha}{K}\sum_{\ell=1}^K \|\boldsymbol{\theta}_\tau - \boldsymbol{p}_\ell\|_{\ell_2}. \tag{9.48}$$

is a supermartingale. Specifically we prove the following lemma.

Lemma 9.11 *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 1 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* $\nu \ge 3$. *Also consider the set*

$$\mathcal{B}(\nu) = \mathcal{B}\left(\boldsymbol{\theta}_0, \nu\frac{\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\alpha}\right) \bigcap \left\{\boldsymbol{\theta} \in \mathbb{R}^p \middle| \|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2} \le \frac{2\nu}{3}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}\right\}. \tag{9.49}$$

Also assume the rows of the Jacobian have bounded Euclidean norm over this ball, that is

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

Furthermore, suppose one of the following statements is valid.

- *Assumption 2 (a) holds over* $\mathcal{D}$ *and set* $\eta \le \frac{\alpha^2}{\nu\beta^2 B^2}$.
- *Assumption 2 (b) holds over* $\mathcal{D}$ *and set* $\eta \le \frac{\alpha^2}{\nu\beta^2 B^2 + \nu\beta B L\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}$.

Fix $K \ge \sqrt{n}\frac{\beta}{\alpha}$ *and let* $\{\boldsymbol{p}_\ell\}_{\ell=1}^K$ *be an* $\epsilon := \frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ *packing of a ball of radius* $R_p := 1.25\left(\frac{\beta}{\alpha}\right)^{1/p}\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$ *around* $\boldsymbol{\theta}_0$ *so that pairwise distances in this set are at least* ϵ *and define the potential* $\mathcal{V}(\boldsymbol{\theta})$ *associated with this packing per* (9.48). *Starting from* $\boldsymbol{\theta}_0$ *we run stochastic gradient updates of the form* (3.1). *Then,* $\mathcal{V}(\boldsymbol{\theta}_0) \le 14\left(\frac{\beta}{\alpha}\right)^{1/p}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}$. *Furthermore, if* $\boldsymbol{\theta}_\tau \in \mathcal{B}(\nu/2)$, *then* $\mathbb{E}[\mathcal{V}(\boldsymbol{\theta}_{\tau+1})] \le \mathcal{V}(\boldsymbol{\theta}_\tau)$.

To bound $\mathcal{V}(\boldsymbol{\theta}_0)$ not that each anchor point in the packing obeys $\|\boldsymbol{p}_\ell - \boldsymbol{\theta}_0\|_{\ell_2} \le 1.25\left(\frac{\beta}{\alpha}\right)^{1/p}\frac{\|f(\boldsymbol{\theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$, we have

$$\begin{aligned}\mathcal{V}(\boldsymbol{\theta}_0) \le& 12\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} + \frac{\alpha}{K}\sum_{i=1}^K \|\boldsymbol{\theta}_0 - \boldsymbol{p}_\ell\|_{\ell_2} \le 12\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} + 1.25\left(\frac{\beta}{\alpha}\right)^{1/p}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}\\ \le& 14\left(\frac{\beta}{\alpha}\right)^{1/p}\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}.\end{aligned}$$

Turning our attention to the supermartingale property, define $\boldsymbol{r}_\tau = f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}$ and note that when $\boldsymbol{\theta}_\tau \in \mathcal{B}(\nu/2)$, by Lemmas 9.9 and 9.10 we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}] \leq& \|\boldsymbol{r}_\tau\|_{\ell_2} - \frac{\eta}{4n}\frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}},\\ \mathbb{E}[d_{\mathcal{P}}(\boldsymbol{\theta}_{\tau+1})] \leq& d(\boldsymbol{\theta}_\tau) + \frac{3\eta}{n}\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2}.\end{aligned}$$

Summing these two identities with a scaling of the first inequality by 12 and the second one by α, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{V}(\boldsymbol{\theta}_{\tau+1})] - \mathcal{V}(\boldsymbol{\theta}_\tau) &\leq 12\left(\mathbb{E}[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}] - \|\boldsymbol{r}_\tau\|_{\ell_2}\right) + \alpha\left(\mathbb{E}[d_{\mathcal{P}}(\boldsymbol{\theta}_{\tau+1})] - d_{\mathcal{P}}(\boldsymbol{\theta}_\tau)\right)\\ &\leq \frac{-12\eta}{4n}\frac{\|\mathcal{J}(\boldsymbol{\theta}_i)\boldsymbol{r}_\tau\|_{\ell_2}^2}{\|\boldsymbol{r}_\tau\|_{\ell_2}} + \frac{3\eta\alpha}{n}\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2}\\ &\leq \frac{3\eta}{n}\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2}\left(\alpha - \frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)\boldsymbol{r}_\tau\|_{\ell_2}}{\|\boldsymbol{r}_\tau\|_{\ell_2}}\right)\\ &\leq 0.\end{aligned}$$

9.4.5. SGD REMAINS IN THE LOCAL NEIGHBORHOOD

In this section we show that SGD iterates remain close to the initialization. Specifically we prove the following lemma.

Lemma 9.12 *Consider the setup of Lemma 9.11 and the potential function $\mathcal{V}$ from (9.48). Also define the stopping time $T = \min\{\tau : \boldsymbol{\theta}_\tau \notin \mathcal{B}(\nu/2)\}$. Under the stated assumptions,*

$$\mathbb{P}\left\{T = \infty\right\} \geq 1 - \frac{4}{\nu}\left(\frac{\beta}{\alpha}\right)^{\frac{1}{p}}.$$

Proof Assume $\boldsymbol{\theta}_\tau \notin \mathcal{B}(\nu/2)$. This implies $\|\boldsymbol{r}_\tau\|_{\ell_2} \geq \frac{\nu}{3}\|\boldsymbol{r}_0\|_{\ell_2}$, hence the potential $\mathcal{V}(\boldsymbol{\theta}_\tau)$ can be lower bounded as

$$\mathcal{V}_\tau = \mathcal{V}(\boldsymbol{\theta}_\tau) \geq \frac{\alpha}{K}\sum_{\ell=1}^{K}\|\boldsymbol{\theta}_\tau - \boldsymbol{p}_\ell\|_{\ell_2} + 12\|\boldsymbol{r}_\tau\|_{\ell_2} \geq 12\|\boldsymbol{r}_\tau\|_{\ell_2} \geq 4\nu\|\boldsymbol{r}_0\|_{\ell_2}.$$

Define the stopping time $\widetilde{T}$ which is the first instance $\mathcal{V}_\tau \geq 4\nu\|\boldsymbol{r}_0\|_{\ell_2}$. Clearly $\widetilde{T} \leq T$ and $\mathbb{P}\{T = \infty\} \geq \mathbb{P}\{\widetilde{T} = \infty\}$. To show that $\widetilde{T} = \infty$ holds with high probability we utilize an argument similar to (Tan & Vershynin, 2017). Define $a \wedge b = \min(a, b)$ and the stopped process $\mathcal{U}_\tau = \mathcal{V}_{\tau\wedge\widetilde{T}}$. We will show that $\mathcal{U}_\tau$ is a supermartingale. Let $\mathcal{F}_\tau$ denote the σ-algebra generated by the first τ SGD random variables $\gamma_1, \gamma_2, \dots, \gamma_\tau$. By construction, $\boldsymbol{\theta}_\tau, \boldsymbol{r}_\tau, \mathcal{V}_\tau$ are measurable with respect to $\mathcal{F}_\tau$. We can decompose the expectation based on the event $\widetilde{T} > \tau$ as follows

$$\begin{aligned}\mathbb{E}[\mathcal{U}_{\tau+1} \mid \mathcal{F}_\tau] &= \mathbb{E}[\mathcal{V}_{(\tau+1)\wedge\widetilde{T}}\mathbb{1}_{\widetilde{T}\leq\tau} \mid \mathcal{F}_\tau] + \mathbb{E}[\mathcal{V}_{(\tau+1)\wedge\widetilde{T}}\mathbb{1}_{\widetilde{T}>\tau} \mid \mathcal{F}_\tau],\\ &= \mathbb{E}[\mathcal{V}_{\tau\wedge\widetilde{T}}\mathbb{1}_{\widetilde{T}\leq\tau} \mid \mathcal{F}_\tau] + \mathbb{E}[\mathcal{V}_{\tau+1}\mathbb{1}_{\widetilde{T}>\tau} \mid \mathcal{F}_\tau].\end{aligned}$$

The $\mathcal{V}_{\tau\wedge\widetilde{T}}$ term is measurable with respect to filteration $\mathcal{F}_\tau$, hence

$$\mathbb{E}[\mathcal{V}_{\tau\wedge\widetilde{T}}\mathbb{1}_{\widetilde{T}\leq\tau} \mid \mathcal{F}_\tau] = \mathcal{V}_{\tau\wedge\widetilde{T}}\mathbb{1}_{\widetilde{T}\leq\tau} = \mathcal{U}_\tau\mathbb{1}_{\widetilde{T}\leq\tau}.$$

Therefore we can focus on the $\mathcal{V}_{\tau+1}\mathbb{1}_{\widetilde{T}>\tau}$ term. As previously discussed, $\widetilde{T} > \tau$ implies $\boldsymbol{\theta}_\tau \in \mathcal{B}(\nu/2)$ and Lemma 9.11 is applicable. This yields

$$\mathbb{E}[\mathcal{V}_{\tau+1}\mathbb{1}_{\widetilde{T}>\tau} \mid \mathcal{F}_\tau] = \mathbb{E}[\mathcal{V}_{\tau+1} \mid \mathcal{F}_\tau]\mathbb{1}_{\widetilde{T}>\tau} \leq \mathcal{V}_\tau\mathbb{1}_{\widetilde{T}>\tau}.$$

Also note that $\mathcal{V}_\tau\mathbb{1}_{\widetilde{T}>\tau} = \mathcal{V}_{\tau\wedge\widetilde{T}}\mathbb{1}_{\widetilde{T}>\tau} = \mathcal{U}_\tau\mathbb{1}_{\widetilde{T}>\tau}$. Combining the latter two identities we have

$$\mathbb{E}[\mathcal{U}_{\tau+1} \mid \mathcal{F}_\tau] \leq \mathcal{U}_\tau\mathbb{1}_{\widetilde{T}>\tau} + \mathcal{U}_\tau\mathbb{1}_{\widetilde{T}\leq\tau} = \mathcal{U}_\tau.$$

Now that we established $\mathcal{U}_\tau$ is a supermartingale, Martingale maximal inequality (Revuz & Yor, 2013) implies that

$$\mathbb{P}\left\{\sup_{\tau\geq0}\mathcal{U}_\tau \geq 4\nu\|\boldsymbol{r}_0\|_{\ell_2}\right\} \leq \frac{\mathcal{U}_0}{4\nu\|\boldsymbol{r}_0\|_{\ell_2}} = \frac{\mathcal{V}_0}{4\nu\|\boldsymbol{r}_0\|_{\ell_2}} \leq \frac{14\left(\frac{\beta}{\alpha}\right)^{1/p}}{4\nu} \leq 4\frac{\left(\frac{\beta}{\alpha}\right)^{1/p}}{\nu}.$$

Hence, $\mathbb{P}\{T = \infty\} \geq \mathbb{P}\{\widetilde{T} = \infty\} \geq 1 - \frac{4}{\nu}\left(\frac{\beta}{\alpha}\right)^{\frac{1}{p}}$. ■

9.4.6. Putting everything together (completing the proof of Theorem 3.1)

In this Section we combine the results of the previous sections to complete the proof. First, note that Lemma 9.12 already establishes the result for $\mathbb{P}\{T=\infty\}$. We set the event E to be equal to $\{T=\infty\}$. To show the result on the convergence rate, we note that if $T=\infty$ then $\boldsymbol{\theta}_\tau\in\mathcal{B}\left(\frac{\nu}{2}\right)$ and hence (9.28) holds. This in turn implies that $\mathbb{E}[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2]\le\left(1-\frac{\eta}{2n}\alpha^2\right)\|\boldsymbol{r}_\tau\|_{\ell_2}^2$. Now recall the filteration $\mathcal{F}_\tau$ generated from random SGD updates in the proof of Lemma 9.12. We have

$$\mathbb{E}[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2\mathbb{1}_{T=\infty}\mid\mathcal{F}_\tau]\le\mathbb{E}[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2\mathbb{1}_{T>\tau}\mid\mathcal{F}_\tau].$$

To continue further note that $\mathbb{1}_{T>\tau}$ is measurable with respect to $\mathcal{F}_\tau$. Hence, applying (9.28) over the event $T>\tau$, we conclude that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2\mathbb{1}_{T>\tau}\mid\mathcal{F}_\tau]&=\mathbb{E}[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2\mid\mathcal{F}_\tau]\mathbb{1}_{T>\tau},\\&\le\left(1-\frac{\eta\alpha^2}{2n}\right)\|\boldsymbol{r}_\tau\|_{\ell_2}^2\mathbb{1}_{T>\tau},\\&\le\left(1-\frac{\eta\alpha^2}{2n}\right)\|\boldsymbol{r}_\tau\|_{\ell_2}^2\mathbb{1}_{T>(\tau-1)}.\end{aligned}$$

With this recursion established, we take conditional expectations to obtain

$$\mathbb{E}\left[\|\boldsymbol{r}_{\tau+1}\|_{\ell_2}^2\mathbb{1}_{T>\tau}\right]\le\left(1-\frac{\eta\alpha^2}{2n}\right)^\tau\mathbb{E}\left[\|\boldsymbol{r}_1\|_{\ell_2}^2\mathbb{1}_{T>0}\right]\le\left(1-\frac{\eta\alpha^2}{2n}\right)^{\tau+1}\|\boldsymbol{r}_0\|_{\ell_2}^2,$$

which completes the proof.

9.5. GLM proofs (Proof of Theorem 4.1)

First we prove that the is a globally optimal solution achieving zero training error. To see this note that any strictly increasing and differentiable activation ϕ is invertible on $\mathbb{R}$ by the implicit function theorem. Let $\Pi_{\mathcal{R}}$ and $\Pi_{\mathcal{N}}$ denote the projections onto the row space and null space of $\boldsymbol{X}$ respectively. By the assumptions of the theorem $\boldsymbol{X}$ has full row rank and pseudo-inverse solution $\boldsymbol{\theta}^\dagger$ is given by $\boldsymbol{\theta}^\dagger=\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\phi^{-1}(\boldsymbol{y})\}$. Hence the set of global optimal solutions is non-empty and all globally optimal solutions are characterized by the null space as follows

$$\mathcal{G}=\{\boldsymbol{\theta}\in\mathbb{R}^p:\boldsymbol{\theta}=\boldsymbol{\theta}^\dagger+\boldsymbol{v}\quad\text{where}\quad\boldsymbol{v}\in\text{null}(\boldsymbol{X})\}$$

Let $\boldsymbol{\theta}^*=\Pi_{\mathcal{N}}(\boldsymbol{\theta}_0)+\boldsymbol{\theta}^\dagger\in\mathcal{G}$. By construction $\boldsymbol{\theta}^*$ is the closest global minima to $\boldsymbol{\theta}_0$ as the null space projections match. We will argue that the gradient descent iterations linearly converge to $\boldsymbol{\theta}^*$.

Towards this goal, note that $\boldsymbol{y}=\phi(\boldsymbol{X}\boldsymbol{\theta}^*)$ and note that the gradient descent iterations are given by

$$\begin{aligned}\boldsymbol{\theta}_{\tau+1}=&\boldsymbol{\theta}_\tau+\eta\boldsymbol{X}^T\text{diag}\left(\phi'(\boldsymbol{X}\boldsymbol{\theta}_\tau)\right)\left(\phi(\boldsymbol{X}\boldsymbol{\theta}^\star)-\phi(\boldsymbol{X}\boldsymbol{\theta}_\tau)\right)&(9.50)\\&\boldsymbol{\theta}_\tau+\eta\boldsymbol{X}^T\text{diag}\left(\phi'(\boldsymbol{X}\boldsymbol{\theta}_\tau)\right)\left(\boldsymbol{y}-\phi(\boldsymbol{X}\boldsymbol{\theta}_\tau)\right).&(9.51)\end{aligned}$$

Now, for two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ obeying $\boldsymbol{a}\ne\boldsymbol{b}$ define $\phi'(\boldsymbol{a},\boldsymbol{b})=\frac{\phi(\boldsymbol{a})-\phi(\boldsymbol{b})}{\boldsymbol{a}-\boldsymbol{b}}$ (with the devision interpreted as entry by entry) and note that by the mean value theorem $\phi'(\boldsymbol{a},\boldsymbol{b})\ge\gamma$. Also note that, we can write $\phi(\boldsymbol{X}\boldsymbol{\theta}_\tau)-\phi(\boldsymbol{X}\boldsymbol{\theta}^\star)=\text{diag}(\phi'(\boldsymbol{X}\boldsymbol{\theta}_\tau,\boldsymbol{X}\boldsymbol{\theta}^\star))\boldsymbol{X}(\boldsymbol{\theta}_\tau-\boldsymbol{\theta}^\star)$. Consequently, setting $\boldsymbol{h}_\tau=\boldsymbol{\theta}_\tau-\boldsymbol{\theta}^\star$ and $\boldsymbol{D}_\tau=\text{diag}(\phi'(\boldsymbol{X}\boldsymbol{\theta}_\tau))\text{diag}(\phi'(\boldsymbol{X}\boldsymbol{\theta}_\tau,\boldsymbol{X}\boldsymbol{\theta}^\star))$, we have

$$\boldsymbol{h}_{\tau+1}=\boldsymbol{h}_\tau-\eta\boldsymbol{X}^T\boldsymbol{D}_\tau\boldsymbol{X}\boldsymbol{h}_\tau=\left(\boldsymbol{I}-\eta\boldsymbol{X}^T\boldsymbol{D}_\tau\boldsymbol{X}\right)\boldsymbol{h}_\tau.\tag{9.52}$$

Since gradient is an element of the row space $\mathcal{R}$, $\Pi_{\mathcal{N}}(\boldsymbol{\theta}_\tau)=\Pi_{\mathcal{N}}(\boldsymbol{\theta}_0)=\Pi_{\mathcal{N}}(\boldsymbol{\theta}^*)$ and $\boldsymbol{h}_\tau\in\mathcal{R}$. To proceed, let $\boldsymbol{V}\in\mathbb{R}^{n\times p}$ be an orthonormal basis (i.e. $\boldsymbol{V}\boldsymbol{V}^T=\boldsymbol{I}_n$) for the row space of $\boldsymbol{X}$ and define $\widetilde{\boldsymbol{h}}_\tau=\boldsymbol{V}\boldsymbol{h}_\tau$ and $\widetilde{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{V}^T$. (9.52) yields the following update rule for $\widetilde{\boldsymbol{h}}_\tau$

$$\begin{aligned}\widetilde{\boldsymbol{h}}_{\tau+1}&=\boldsymbol{V}(\boldsymbol{I}-\eta\boldsymbol{X}^T\boldsymbol{D}_\tau\boldsymbol{X})\boldsymbol{V}^T\widetilde{\boldsymbol{h}}_\tau,\\&=\left(\boldsymbol{I}-\eta\widetilde{\boldsymbol{X}}^T\boldsymbol{D}_\tau\widetilde{\boldsymbol{X}}\right)\widetilde{\boldsymbol{h}}_\tau.\end{aligned}$$

To continue further, we use the fact that $\boldsymbol{D}_\tau$ is diagonal with entries between γ^2 and Γ^2. This combined with the fact that the matrices $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{X}$ have the same eigenvalues allow us to conclude that $\gamma^2\sigma_{\min}^2(\boldsymbol{X})\boldsymbol{I} \le \widetilde{\boldsymbol{X}}^T\boldsymbol{D}_\tau\widetilde{\boldsymbol{X}} \le \Gamma^2\|\boldsymbol{X}\|^2\boldsymbol{I}$. Thus, for $\eta \le \frac{1}{\Gamma^2\|\boldsymbol{X}\|^2}$

$$\boldsymbol{0} \le \boldsymbol{I} - \eta\widetilde{\boldsymbol{X}}^T\boldsymbol{D}_\tau\widetilde{\boldsymbol{X}} \le \left(1-\eta\gamma^2\sigma_{\min}^2(\boldsymbol{X})\right)\boldsymbol{I}.$$

Thus, using the fact that $\left\|\widetilde{\boldsymbol{h}}_\tau\right\|_{\ell_2} = \|\boldsymbol{V}\boldsymbol{h}_\tau\|_{\ell_2} = \|\boldsymbol{h}_\tau\|_{\ell_2}$ (as $\boldsymbol{h}_\tau \in \mathcal{R}$) we conclude that

$$\|\boldsymbol{h}_{\tau+1}\|_{\ell_2} \le \left(1-\eta\gamma^2\sigma_{\min}^2(\boldsymbol{X})\right)\|\boldsymbol{h}_\tau\|_{\ell_2},$$

completing the proof of (4.2). Furthermore, note that

$$\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2} = \|\boldsymbol{h}_{\tau+1}-\boldsymbol{h}_\tau\|_{\ell_2} \le \|\eta\boldsymbol{X}^T\boldsymbol{D}_\tau\boldsymbol{X}\boldsymbol{h}_\tau\|_{\ell_2} \le \eta\Gamma^2\|\boldsymbol{X}\|^2\|\boldsymbol{h}_\tau\|_{\ell_2}.$$

Summing these up from $\tau = 0$ to ∞ and using $\|\boldsymbol{h}_\tau\|_{\ell_2} \le \left(1-\eta\gamma^2\sigma_{\min}^2(\boldsymbol{X})\right)^\tau\|\boldsymbol{h}_0\|_{\ell_2}$ we conclude that for $\eta = \frac{1}{\Gamma^2\|\boldsymbol{X}\|^2}$

$$\sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2} = \sum_{\tau=0}^{\infty}\|\boldsymbol{h}_\tau\|_{\ell_2} \le \frac{1}{1-\left(1-\eta\gamma^2\sigma_{\min}^2(\boldsymbol{X})\right)}\|\boldsymbol{h}_0\|_{\ell_2} = \frac{\Gamma^2}{\gamma^2}\frac{\lambda_{\max}\left(\boldsymbol{X}\boldsymbol{X}^T\right)}{\lambda_{\min}\left(\boldsymbol{X}\boldsymbol{X}^T\right)}\|\boldsymbol{h}_0\|_{\ell_2},$$

establishing (4.3).

9.6. Low-rank recovery proofs (Proof of Theorem 4.2)

To specialize Theorem 2.1 we begin by calculating the Jacobian $\mathcal{J}(\boldsymbol{\Theta}) := \mathcal{J}(\text{vect}(\boldsymbol{\Theta}))$ which is given by an $n \times dr$ matrix of the form

$$\mathcal{J}(\boldsymbol{\Theta}) = \left[\text{vect}(\boldsymbol{X}_1\boldsymbol{\Theta}) \quad \text{vect}(\boldsymbol{X}_2\boldsymbol{\Theta}) \quad \dots \quad \text{vect}(\boldsymbol{X}_n\boldsymbol{\Theta})\right]^T.$$

Here, for a matrix $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ we use $\text{vect}(\boldsymbol{M}) \in \mathbb{R}^{n_1n_2}$ to denote an n_1n_2 dimensional column vectors obtained by concatenating the columns of $\boldsymbol{M}$. Similarly, for a vector $\boldsymbol{v} \in \mathbb{R}^{n_1n_2}$ we use $\text{mat}(\boldsymbol{v}) \in \mathbb{R}^{n_1\times n_2}$ to denote a matrix obtained by reshaping the vector into an $n_1 \times n_2$ matrix.

9.6.1. Key Lemmas for low-rank recovery

In order to verify the assumptions of Theorem 2.1, in this section we gather some key lemmas related to the Jacobian matrix that building on top of each other play a crucial role in our proofs. We defer the proofs to Appendix A. The first key lemma which will play a crucial role in our proofs is that the nuclear norm $\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)\|_\star$ is uniformly bounded for all $\boldsymbol{v}$ and $\boldsymbol{\Theta}$ with unit Frobenius/Euclidean norms.

Lemma 9.13 *For $i = 1,2,\dots,n$, $\boldsymbol{X}_i \in \mathbb{R}^{d\times d}$ be i.i.d. matrices with i.i.d. entries distributed as $\mathcal{N}(0,1)$. Furthermore, assume $n \le dr$ and $r \le d$. Then*

$$\sup_{\boldsymbol{v}\in\mathbb{R}^k,\boldsymbol{\Theta}\in\mathbb{R}^{d\times r}:\|\boldsymbol{v}\|_{\ell_2}=\|\boldsymbol{\Theta}\|_F=1}\|mat\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)\|_\star \le 12\sqrt{dr},$$

holds with probability at least $1-e^{-2dr}$.

The next lemma concerns the average of the nuclear norm of a Gaussian matrix multiplied by a diagonal matrix.

Lemma 9.14 *Let $\boldsymbol{G} \in \mathbb{R}^{d\times r}$ with $d \le r$ be i.i.d. $\mathcal{N}(0,1)$ matrix and $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ be a diagonal matrix with entries obeying*

$$\vartheta \le \sigma_{\min}(\boldsymbol{\Sigma}) \le \sigma_{\max}(\boldsymbol{\Sigma}) \le 2\vartheta$$

Then,

$$\mathbb{E}[\|\boldsymbol{G}\boldsymbol{\Sigma}\|_\star] \ge \frac{1}{32}\vartheta\sqrt{dr}.$$

The next key lemma used in our proofs also concerns the nuclear norm $\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)\|_\star$, however this time we bound this quantity from both below and above for a fixed matrix $\boldsymbol{\Theta}$ that is well conditioned and for all vectors $\boldsymbol{v} \in \mathbb{R}^k$ with unit Euclidean norm.

Lemma 9.15 *Let $\mathbf{\Theta} \in \mathbb{R}^{d\times r}$ be a matrix with eigenvalues obeying*

$$\vartheta \le \sigma_{\min}\left(\mathbf{\Sigma}\right) \le \sigma_{\max}\left(\mathbf{\Sigma}\right) \le 2\vartheta.$$

Furthermore, assume $r \le d$ and $n \le cdr$ with c a fixed numerical constant. Then,

$$\frac{1}{40}\vartheta\sqrt{dr} \le \left\| mat\left(\mathcal{J}(\mathbf{\Theta})^T\boldsymbol{v}\right)\right\|_* \le 24\vartheta\sqrt{dr},$$

holds for all $\boldsymbol{v} \in \mathbb{S}^{n-1}$ with probability at least $1 - 2e^{-\gamma dr}$ with γ a fixed numerical constant.

Next we bound the spectrum of the Jacobian matrix in a ball around the initialization $\mathbf{\Theta}_0$.

Lemma 9.16 (Jacobian spectrum bounds) *Let $\mathbf{\Theta}_0 \in \mathbb{R}^{d\times r}$ with $r \le d$ be a matrix with singular values lying in the range $[\vartheta, 2\vartheta]$. Consider the Frobenius ball around $\mathbf{\Theta}_0$ given by $\mathcal{D} = \mathcal{B}\left(\mathbf{\Theta}_0, \frac{1}{2400}\vartheta\sqrt{r}\right)$. Then as long as $n \le Cdr$ with C a fixed numerical constant, then, with probability at least $1 - 3e^{-\gamma dr}$*

$$\frac{1}{50}\vartheta\sqrt{dr} \le \sigma_{\min}\left(\mathcal{J}(\mathbf{\Theta})\right) \le \sigma_{\max}\left(\mathcal{J}(\mathbf{\Theta})\right) \le 25\vartheta\sqrt{dr}. \tag{9.53}$$

Furthermore, the Jacobian matrix is $12\sqrt{dr}$-Lipschitz. That is, for all $\mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \mathbb{R}^{d\times r}$ we have

$$\|\mathcal{J}(\mathbf{\Theta}_2) - \mathcal{J}(\mathbf{\Theta}_1)\| \le 12\sqrt{dr}\,\|\mathbf{\Theta}_2 - \mathbf{\Theta}_1\|_F\,. \tag{9.54}$$

9.6.2. Completing the proof of Theorem 4.2

We will prove this theorem by a direct application of Theorem 2.1. To this aim we need to calculate the various parameters in this theorem.

We begin by calculating the size of the initial misfit. To this aim note that $\left\langle \boldsymbol{X}_i, \mathbf{\Theta}_0\mathbf{\Theta}_0^T\right\rangle \sim \mathcal{N}(0, \|\mathbf{\Theta}_0\mathbf{\Theta}_0^T\|_F^2)$ and $\|\mathbf{\Theta}_0\mathbf{\Theta}_0^T\|_F \le \sqrt{r}\left\|\mathbf{\Theta}_0\mathbf{\Theta}_0^T\right\| \le 4\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}$. Hence, $f(\mathbf{\Theta}_0)$ is an i.i.d. Gaussian random vector with standard deviation at most $4\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}$. Using Lipschitz concentration of Gaussians, this implies that

$$\mathbb{P}\left\{\frac{\|f(\mathbf{\Theta}_0)\|_{\ell_2}}{4\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{n}}} \ge 2\sqrt{n}\right\} \le e^{-\frac{n}{2}}.$$

Hence, with probability at least $1 - e^{-\frac{n}{2}}$, the following holds

$$\|f(\mathbf{\Theta}_0) - \boldsymbol{y}\|_{\ell_2} \le \|f(\mathbf{\Theta}_0)\|_{\ell_2} + \|\boldsymbol{y}\|_{\ell_2} \le 9\,\|\boldsymbol{y}\|_{\ell_2}\,, \tag{9.55}$$

Furthermore, applying Lemma 9.16 with $\vartheta = \sqrt{\frac{\|\boldsymbol{y}\|_{\ell_2}}{\sqrt{rn}}}$, Jacobian matrix satisfies

$$\alpha = \frac{1}{50}\sqrt{d\,\|\boldsymbol{y}\|_{\ell_2}\sqrt{\frac{r}{n}}}, \quad \beta = 25\sqrt{dr\,\|\boldsymbol{y}\|_{\ell_2}\sqrt{\frac{r}{n}}}, \quad \text{and} \quad L = 12\sqrt{dr}.$$

over the domain $\mathcal{D}' = \mathcal{B}(\boldsymbol{\theta}_0, \frac{1}{2400}\vartheta\sqrt{r})$ with probability $1 - 3e^{-\gamma dr} \ge 1 - 3e^{-n/2}$ (by picking $c \le \gamma$). On the other hand, for Theorem 2.1 to be applicable, we need the domain $\mathcal{D}$ radius to be $R := \frac{4\|f(\mathbf{\Theta}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha}$. The key idea is choosing $n \le cdr$ for a sufficiently small c to ensure that $\mathcal{D} \subset \mathcal{D}'$ and Theorem 2.1 applies. In particular, this follows from

$$R := \frac{4\,\|f(\mathbf{\Theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\alpha} \le 1800\sqrt{\frac{\|\boldsymbol{y}\|_{\ell_2}}{d\sqrt{\frac{r}{n}}}} = 1800\sqrt{\frac{n}{dr}}\sqrt{\|\boldsymbol{y}\|_{\ell_2}\sqrt{\frac{r}{n}}} \le \frac{1}{2400}\sqrt{\|\boldsymbol{y}\|_{\ell_2}\sqrt{\frac{r}{n}}} = \frac{1}{2400}\vartheta\sqrt{r},$$

Now that Theorem 2.1 applies, all that remains is to upper bound these quantities in the upper bound on the learning rate. Per Theorem 2.1 we need to ensure

$$\eta \le \frac{1}{2\beta^2}\cdot\min\left(1, \frac{\alpha^2}{L\,\|f(\mathbf{\Theta}_0) - \boldsymbol{y}\|_{\ell_2}}\right).$$

To do this note that

$$\frac{\alpha^2}{L\left\|f(\boldsymbol{\Theta}_0)-\boldsymbol{y}\right\|_{\ell_2}}\ge\frac{\alpha^2}{9L\left\|\boldsymbol{y}\right\|_{\ell_2}}=\frac{\alpha^2}{108\sqrt{dr}\left\|\boldsymbol{y}\right\|_{\ell_2}}=\frac{\frac{1}{2500}d\|\boldsymbol{y}\|_{\ell_2}\sqrt{r/n}}{108\sqrt{dr}\left\|\boldsymbol{y}\right\|_{\ell_2}}=\frac{1}{270000}\sqrt{\frac{d}{n}},$$

and use $\min(1,\sqrt{\frac{d}{n}})\ge\frac{1}{\sqrt{r}}$. Proceeding, we use this naive bound to simplify the final expressions. This yields the step size requirement of

$$\eta\le\frac{c'}{\beta^2\sqrt{r}}=\frac{c_1}{dr\|\boldsymbol{y}\|_{\ell_2}\sqrt{r/n}\sqrt{r}}=\frac{c_1\sqrt{n}}{r^2d\|\boldsymbol{y}\|_{\ell_2}}$$

Observing $\alpha^2/\beta^2=1/r$ and substituting η and convergence rate $1-\eta\alpha^2/2$ concludes the proof.

9.7. Neural net proofs (Proof of Theorem 4.3)

We begin by noting that the Jacobian matrix in this case is equal to

$$\mathcal{J}(\boldsymbol{W})=\begin{bmatrix}\boldsymbol{v}_1\mathcal{J}(\boldsymbol{w}_1) & \ldots & \boldsymbol{v}_k\mathcal{J}(\boldsymbol{w}_k)\end{bmatrix}\in\mathbb{R}^{n\times kd}\quad\text{with}\quad\mathcal{J}(\boldsymbol{w}_\ell):=\text{diag}(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell))\boldsymbol{X}.$$

To prove this theorem we use Theorem 2.1 with $R=\infty$. We just need to calculate the various parameters and verify that the assumptions hold.

Bounding the spectrum of $\mathcal{J}$. We begin by calculating α and β. To this aim note

$$\mathcal{J}(\boldsymbol{w}_{\boldsymbol{\ell}})\mathcal{J}^T(\boldsymbol{w}_{\boldsymbol{\ell}})=\text{diag}\left(\left(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)\right)\boldsymbol{X}\boldsymbol{X}^T\text{diag}\left(\left(\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)\right)\right.\right..$$

Thus, using the bounds on ϕ'

$$\gamma^2\sigma_{\min}^2(\boldsymbol{X})\boldsymbol{I}\le\mathcal{J}(\boldsymbol{w}_{\boldsymbol{\ell}})\mathcal{J}^T(\boldsymbol{w}_{\boldsymbol{\ell}})\le\Gamma^2\left\|\boldsymbol{X}\right\|^2\boldsymbol{I}.$$

This in turn implies that for $\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W})=\sum_{\ell=1}^k\boldsymbol{v}_k^2\mathcal{J}(\boldsymbol{w}_{\boldsymbol{\ell}})\mathcal{J}^T(\boldsymbol{w}_{\boldsymbol{\ell}})$ we have

$$\gamma^2\left\|\boldsymbol{v}\right\|_{\ell_2}^2\sigma_{\min}^2(\boldsymbol{X})\boldsymbol{I}\le\mathcal{J}(\boldsymbol{W})\mathcal{J}^T(\boldsymbol{W})\le\Gamma^2\left\|\boldsymbol{v}\right\|_{\ell_2}^2\left\|\boldsymbol{X}\right\|^2\boldsymbol{I},$$

so that we can use

$$\alpha=\gamma\sigma_{\min}(\boldsymbol{X})\quad\text{and}\quad\beta=\Gamma\left\|\boldsymbol{X}\right\|.$$

Bounding the Lipschitz parameter of $\mathcal{J}$. To calculate L note that

$$\mathcal{J}(\widetilde{\boldsymbol{W}})-\mathcal{J}(\boldsymbol{W})=\begin{bmatrix}\boldsymbol{v}_1\left(\mathcal{J}(\widetilde{\boldsymbol{w}}_1)-\mathcal{J}(\boldsymbol{w}_1)\right) & \ldots & \boldsymbol{v}_k\left(\mathcal{J}(\widetilde{\boldsymbol{w}}_k)-\mathcal{J}(\boldsymbol{w}_k)\right)\end{bmatrix}.$$

Thus

$$\begin{aligned}\left\|\mathcal{J}(\widetilde{\boldsymbol{W}})-\mathcal{J}(\boldsymbol{W})\right\|^2&\overset{(a)}{\le}\sum_{\ell=1}^k\left\|\boldsymbol{v}_\ell\left(\mathcal{J}(\widetilde{\boldsymbol{w}}_\ell)-\mathcal{J}(\boldsymbol{w}_\ell)\right)\right\|^2\\&=\sum_{\ell=1}^k\boldsymbol{v}_\ell^2\left\|\text{diag}\left(\phi'(\boldsymbol{X}\widetilde{\boldsymbol{w}}_\ell)-\phi'(\boldsymbol{X}\boldsymbol{w}_\ell)\right)\boldsymbol{X}\right\|^2\\&=\sum_{\ell=1}^k\boldsymbol{v}_\ell^2\left\|\text{diag}\left(\int_0^1\phi''\left(\boldsymbol{X}\left(t\widetilde{\boldsymbol{w}}_\ell+(1-t)\boldsymbol{w}_\ell\right)\right)dt\right)\text{diag}\left(\boldsymbol{X}\left(\widetilde{\boldsymbol{w}}-\boldsymbol{w}\right)\right)\boldsymbol{X}\right\|^2,\\&\le\sum_{\ell=1}^k\boldsymbol{v}_\ell^2M^2\|\boldsymbol{X}\|_{2,\infty}^2\left\|\boldsymbol{X}\right\|^2\left\|\widetilde{\boldsymbol{w}}_\ell-\boldsymbol{w}_\ell\right\|_{\ell_2}^2\\&=\left\|\boldsymbol{v}\right\|_{\ell_2}^2M^2\|\boldsymbol{X}\|_{2,\infty}^2\left\|\boldsymbol{X}\right\|^2\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}\right\|_F^2.\end{aligned}$$

In the above (a) follows from the fact the square of the spectral norm of concatenation of matrices is bounded by sum of squares of the spectral norms of the individual matrices. Thus we can use

$$L=M\|\boldsymbol{X}\|_{2,\infty}\left\|\boldsymbol{X}\right\|.$$

The proof is complete by applying Theorem 2.1.

9.8. PL proofs

9.8.1. PL convergence proof (Proof of Theorem 5.2)

Suppose (5.1) and (5.2) hold until step τ. This implies $\boldsymbol{\theta}_\tau \in \mathcal{D}$ and local PL is applicable. If $\mathcal{L}(\boldsymbol{\theta}_\tau) = 0$, then $\boldsymbol{\theta}_\tau$ is global minimizer and since $\mathcal{L}$ is differentiable $\nabla\mathcal{L}(\boldsymbol{\theta}_\tau) = 0$ which in turn implies that $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau$ and thus (5.2) holds for $\boldsymbol{\theta}_{\tau+1}$. Otherwise, $\mathcal{L}(\boldsymbol{\theta}_\tau) > 0$ and using the triangular inequality we can conclude that

$$\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\|_{\ell_2} \le \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} \le \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \eta\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}. \tag{9.56}$$

Since $\nabla\mathcal{L}(\cdot)$ is Lipschitz, we have $\mathcal{L}(\boldsymbol{\theta}_{\tau+1}) \le \mathcal{L}(\boldsymbol{\theta}_\tau) + (\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau)^T\nabla\mathcal{L}(\boldsymbol{\theta}_\tau) + \frac{L}{2}\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2}^2$ for $\eta \le \eta_{\max}$ where $\eta_{\max}$ is the largest step size ensuring $\boldsymbol{\theta}_{\tau+1} \in \mathcal{D}$. Hence, for any $\eta \le \widetilde{\eta}_{\max} = \min(1/L, \eta_{\max})$ we have

$$\mathcal{L}(\boldsymbol{\theta}_{\tau+1}) \le \mathcal{L}(\boldsymbol{\theta}_\tau) - \frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2. \tag{9.57}$$

Now define

$$\begin{aligned}
\varepsilon_\tau(\eta) :=& \left(\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} - \frac{\eta}{4\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)}}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2\right) - \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau) - \frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2},\\
\ge& \left(\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} - \frac{\eta}{4\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)}}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2\right) - \sqrt{\left(\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} - \frac{\eta}{4\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)}}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2\right)^2},\\
=&0,
\end{aligned}$$

so that $\epsilon_\tau(\eta) > 0$ for $\eta > 0$. Using this definition in (9.57) together with the PL condition for $\boldsymbol{\theta}_\tau \in \mathcal{D}$, we arrive at

$$\sqrt{\mathcal{L}(\boldsymbol{\theta}_{\tau+1})} \le \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} - \frac{\eta}{4\sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)}}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2 - \varepsilon_\tau(\eta) \le \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} - \frac{\eta\sqrt{2\mu}}{4}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} - \varepsilon_\tau(\eta). \tag{9.58}$$

To continue we define the potential/Lyapunov function $\mathcal{V}_\tau = \sqrt{\mathcal{L}(\boldsymbol{\theta}_\tau)} + \sqrt{\mu/8}\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2}$ to monitor the sum of the square root of the loss and the distance to initialization. Adding inequalities (9.56) and (9.58), we find that for all $\eta \le \widetilde{\eta}_{\max}$

$$\frac{1}{\eta}\left(\mathcal{V}_{\tau+1} + \varepsilon_\tau(\eta) - \mathcal{V}_\tau\right) \le \sqrt{\frac{\mu}{8}}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} - \frac{\sqrt{2\mu}}{4}\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2} \le 0 \quad\Longrightarrow\quad \mathcal{V}_{\tau+1} \le \mathcal{V}_\tau - \varepsilon_\tau(\eta). \tag{9.59}$$

Next, we argue that $\eta_{\max} \ge 1/L$ and thus $\widetilde{\eta}_{\max} = 1/L$. Note that $\eta_{\max} > 0$ since $\mathcal{L}(\boldsymbol{\theta}_\tau) > 0$ which implies $\boldsymbol{\theta}_\tau$ is strictly inside $\mathcal{D}$ via (5.2). To show that $\eta_{\max} \ge 1/L$, we proceed by contradiction and assume that $\eta_{\max} < 1/L$. Now define $\boldsymbol{\theta}_{\max} := \boldsymbol{\theta}_\tau - \eta_{\max}\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$ and note that by the definition of $\eta_{\max}$, we have $\|\boldsymbol{\theta}_{\max} - \boldsymbol{\theta}_0\|_{\ell_2} = R$. On the other hand, since $\eta_{\max} > 0$ we have $\varepsilon(\eta_{\max}) > 0$ so that applying the update inequality (9.59) (which holds if $\eta_{\max} < 1/L$) we can conclude that

$$\sqrt{\mu/8}\|\boldsymbol{\theta}_{\max} - \boldsymbol{\theta}_0\|_{\ell_2} \le \sqrt{\mathcal{L}(\boldsymbol{\theta}_{\max})} + \sqrt{\mu/8}\|\boldsymbol{\theta}_{\max} - \boldsymbol{\theta}_0\|_{\ell_2} \le \mathcal{V}_\tau - \varepsilon(\eta_{\max}) < \mathcal{V}_0 \implies \|\boldsymbol{\theta}_{\max} - \boldsymbol{\theta}_0\|_{\ell_2} < R.$$

This is in contradiction with $\|\boldsymbol{\theta}_{\max} - \boldsymbol{\theta}_0\|_{\ell_2} = R$ and therefore $\eta_{\max} \ge 1/L$ and $\widetilde{\eta}_{\max} = 1/L$.

The argument above shows that the recursion (9.59) is valid for $\eta \le 1/L$ which proves (5.2) and also in turn guarantees that all $\boldsymbol{\theta}_\tau$'s stay within the neighborhood $\mathcal{D}$ with the learning rate choice of $\eta \le 1/L$. To show convergence of the loss, we combine (9.57) with the PL condition $\|\nabla\mathcal{L}(\boldsymbol{\theta}_\tau)\|_{\ell_2}^2 \ge 2\mu\mathcal{L}(\boldsymbol{\theta}_\tau)$ to conclude that

$$\mathcal{L}(\boldsymbol{\theta}_{\tau+1}) \le (1 - \eta\mu)\mathcal{L}(\boldsymbol{\theta}_\tau) \le (1 - \eta\mu)^{\tau+1}\mathcal{L}(\boldsymbol{\theta}_0),$$

completing the proof of (5.1). To conclude with the result on the shortest path, we add (9.58) from $\tau = 0$ to ∞ to conclude that

$$\sum_{\tau=0}^{\infty}\frac{\eta\sqrt{2\mu}}{4}\|\nabla\mathcal{L}(\boldsymbol{\theta}_i)\|_{\ell_2} \le \sqrt{\mathcal{L}(\boldsymbol{\theta}_0)} \quad\Longrightarrow\quad \sum_{\tau=0}^{\infty}\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} \le \sqrt{\frac{8\mathcal{L}(\boldsymbol{\theta}_0)}{\mu}},$$

completing the proof of (5.3).

9.8.2. PL LOWER BOUND PROOF (PROOF OF THEOREM 5.4)

Proof Suppose there exists $\boldsymbol{\theta} \in \mathcal{D}$ satisfying $\mathcal{L}(\boldsymbol{\theta}) = 0$. Since $\mathcal{L}$ is differentiable and minimized at $\boldsymbol{\theta}$ the gradient must vanish, i.e. $\nabla\mathcal{L}(\boldsymbol{\theta}) = 0$. From smoothness of the loss we conclude that

$$\mathcal{L}(\boldsymbol{\theta}_0) \le \mathcal{L}(\boldsymbol{\theta}) + (\boldsymbol{\theta}_0 - \boldsymbol{\theta})^T \nabla\mathcal{L}(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}^2 = \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2}^2.$$

This implies $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \ge \sqrt{2\mathcal{L}(\boldsymbol{\theta}_0)/L}$ and contradicts with the choice of R.

The remaining proof is similar to that of Theorem 2.4. Consider the least squares problem where $\boldsymbol{X}$ is a matrix with orthogonal rows. The first row $\boldsymbol{x}_1$ of $\boldsymbol{X}$ has length $\sqrt{\mu}$ and the other rows have arbitrary lengths. Fix an arbitrary scaling $\gamma \ge 0$ and set $\boldsymbol{\theta}^\star = \gamma\boldsymbol{x}_1/\|\boldsymbol{x}_1\|_{\ell_2}$ and $\boldsymbol{\theta}_0 = 0$. Set labels $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^\star$ and loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2$. Gradient is $\|\boldsymbol{X}^T\boldsymbol{X}\|$ Lipschitz, which is same as ℓ_2^2 of the largest row, hence L can be set arbitrarily. For any $\boldsymbol{\theta}$, we have

$$\|\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})\|_{\ell_2}^2 = \|\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_{\ell_2}^2 \ge \mu\|\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\|_{\ell_2}^2 = 2\mu\mathcal{L}(\boldsymbol{\theta})$$

Next, observe that (i) $\mathcal{L}(0) = \gamma^2\mu/2$ and (ii) any global minimizer $\boldsymbol{\theta}$ satisfies $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^\star = \boldsymbol{X}\boldsymbol{\theta}$ hence we have that

$$\|\boldsymbol{\theta}\|_{\ell_2} \ge \frac{\boldsymbol{x}_1^T\boldsymbol{\theta}}{\|\boldsymbol{x}_1\|_{\ell_2}} = \frac{\boldsymbol{x}_1^T\boldsymbol{\theta}^\star}{\|\boldsymbol{x}_1\|_{\ell_2}} = \gamma.$$

This implies $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} = \|\boldsymbol{\theta}\|_{\ell_2} \ge \gamma$. Thus, there is no global minima within $R < \gamma = \sqrt{2\mathcal{L}(0)/\mu}$ neighborhood of $\boldsymbol{\theta}_0$. ■

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.

Alon Brutzkus, A. and Globerson, A. Globally optimal gradient descent for a convnet with Gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

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.

Bassily, R., Belkin, M., and Ma, S. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

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.

Candes, E. J., Li, X., and Soltanolkotabi, M. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

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.

Chen, Y., Chi, Y., Fan, J., and Ma, C. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *arXiv preprint arXiv:1803.07726*, 2018.

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.

De Sa, C. M., Zhang, C., Olukotun, K., and Ré, C. Taming the wild: A unified analysis of hogwild-style algorithms. In *Advances in neural information processing systems*, pp. 2674–2682, 2015.

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.

Fabian, Z., Li, M., Oymak, S., and Soltanolkotabi, M. Overparameterization without overfitting: Jacobian-based generalization guarantees for neural networks. *preprint*, 2019.

Ge, R., Lee, J. D., and Ma, T. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

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.

Hassani, S. H., Soltanolkotabi, M., and Karbasi, A. Gradient methods for submodular maximization. *CoRR*, abs/1708.03949, 2017. URL `http://arxiv.org/abs/1708.03949`.

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

Josz, C., Ouyang, Y., Zhang, R. Y., Lavaei, J., and Sojoudi, S. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. 05 2018. URL `https://arxiv.org/pdf/1805.08204`.

Kalan, S. M. M., Soltanolkotabi, M., and Avestimehr, A. S. Fitting relus via SGD and quantized SGD. *CoRR*, abs/1901.06587, 2019. URL `http://arxiv.org/abs/1901.06587`.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

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.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pp. 2348–2358, 2017.

Li, X., Ling, S., Strohmer, T., and Wei, K. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 2018a.

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. 12 2017. URL `https://arxiv.org/pdf/1712.09203`.

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, 2018b.

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

Lojasiewicz, S. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117: 87–89, 1963.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *arXiv preprint arXiv:1712.06559*, 2017.

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*, 2018a.

Oymak, S. Stochastic gradient descent learns state equations with nonlinear activations. *arXiv preprint arXiv:1809.03019*, 2018b.

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

Oymak, S., Recht, B., and Soltanolkotabi, M. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

Polyak, B. T. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3 (4):643–653, 1963.

Rakhlin, A., Shamir, O., Sridharan, K., et al. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, volume 12, pp. 1571–1578. Citeseer, 2012.

Revuz, D. and Yor, M. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

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. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. 02 2017a. URL https://arxiv.org/pdf/1702.06175.

Soltanolkotabi, M. Learning ReLUs via gradient descent. *arXiv preprint arXiv:1705.04591*, 2017b.

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.

Sun, J., Qu, Q., and Wright, J. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5): 1131–1198, 2018.

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.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

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.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

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.

Xu, R., Soltanolkotabi, M., Haldar, J. P., Unglaub, W., Zusman, J., Levi, A. F., and Leahy, R. M. Accelerated wirtinger flow: A fast algorithm for ptychography. *arXiv preprint arXiv:1806.05546*, 2018.

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

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.

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.

A. Proof of key lemmas for low-rank recovery

A.1. Uniform upper bounds on the nuclear norm (Proof of Lemma 9.13)

Given the random nature of the matrices $\boldsymbol{X}_i$, mat $\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)$ defines a random process $\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}$ indexed by $\boldsymbol{\Theta}$ and $\boldsymbol{v}$ that can be rewritten in the form

$$\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}} := \text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right) = \sum_{i=1}^n \boldsymbol{v}_i\boldsymbol{X}_i\boldsymbol{\Theta}.$$

Define $\mathbb{S}^{dr-1} = \{\boldsymbol{\Theta} \in \mathbb{R}^{d\times r} : \|\boldsymbol{\Theta}\|_F = 1\}$ as the space of matrices with unit Frobenius norm and $\mathbb{S}^{n-1}$ as the unit sphere in $\mathbb{R}^n$. The statement of the lemma can then be rephrased as bounding the supremum of this stochastic process over $\mathbb{S}^{dr-1}\times\mathbb{S}^{n-1}$, that is $\sup_{\boldsymbol{v}\in\mathbb{S}^{n-1},\boldsymbol{\Theta}\in\mathbb{S}^{dr-1}}\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_\star$. To establish such a bound, we first determine the behavior of $\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}$ for fixed $\boldsymbol{\Theta}\in\mathbb{S}^{dr-1}$ and $\boldsymbol{v}\in\mathbb{S}^{n-1}$. Assume $\boldsymbol{\Theta}$ has a singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ with $\boldsymbol{U},\boldsymbol{V}\in\mathbb{R}^{d\times r}$. Define $\boldsymbol{Y} = \sum_{i=1}^n \boldsymbol{v}_i\boldsymbol{X}_i\boldsymbol{U}$ and note that $\boldsymbol{Y}\in\mathbb{R}^{d\times r}$ is a matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Hence, using $\|\boldsymbol{\Sigma}\|_F = 1$ and $\|\boldsymbol{\Sigma}\|_\star \le \sqrt{r}$, we have

$$\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_* = \left\|\boldsymbol{Y}\boldsymbol{\Sigma}\boldsymbol{V}^T\right\|_* = \|\boldsymbol{Y}\boldsymbol{\Sigma}\|_* \le \|\boldsymbol{Y}\|\,\|\boldsymbol{\Sigma}\|_* \le \sqrt{r}\,\|\boldsymbol{Y}\|\,.$$

Note that, expectation of the spectral norm is known to be bounded by $\mathbb{E}[\|\boldsymbol{Y}\|] \le \sqrt{d}+\sqrt{r} \le 2\sqrt{d}$ via Gordon's lemma. This yields

$$\mathbb{E}[\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_*] \le \mathbb{E}[\sqrt{r}\,\|\boldsymbol{Y}\|] \le 2\sqrt{dr}. \tag{A.1}$$

Next, we also show that $\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_*$ concentrates well around this expectation. To show this we use the fact stated above that $\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_* = \|\boldsymbol{Y}\boldsymbol{\Sigma}\|_*$ is a function of a Gaussian matrix $\boldsymbol{Y}$. Furthermore, $\|\boldsymbol{Y}\boldsymbol{\Sigma}\|_*$ is Lipschitz as for any two matrices $\boldsymbol{Y}_1, \boldsymbol{Y}_2$ we have

$$\begin{aligned} \left|\|\boldsymbol{Y}_2\boldsymbol{\Sigma}\|_* - \|\boldsymbol{Y}_1\boldsymbol{\Sigma}\|_*\right| \le& \|(\boldsymbol{Y}_2 - \boldsymbol{Y}_1)\boldsymbol{\Sigma}\|_* \\ =&\langle \boldsymbol{V}, (\boldsymbol{Y}_2 - \boldsymbol{Y}_1)\boldsymbol{\Sigma}\rangle \\ =&\langle \boldsymbol{Y}_2 - \boldsymbol{Y}_1, \boldsymbol{V}\boldsymbol{\Sigma}^T\rangle \\ \le& \|\boldsymbol{Y}_2 - \boldsymbol{Y}_1\|_F \left\|\boldsymbol{V}\boldsymbol{\Sigma}^T\right\|_F \\ \le& \|\boldsymbol{Y}_2 - \boldsymbol{Y}_1\|_F \|\boldsymbol{V}\| \|\boldsymbol{\Sigma}\|_F \\ \le& \|\boldsymbol{Y}_2 - \boldsymbol{Y}_1\|_F . \end{aligned} \tag{A.2}$$

Here $\boldsymbol{V}$ follows from dual representation of the nuclear norm and is a matrix with spectral norm bounded by 1 maximizing $\langle \boldsymbol{V}, (\boldsymbol{Y}_2 - \boldsymbol{Y}_1)\boldsymbol{\Sigma}\rangle$. Thus for fixed $\boldsymbol{v}$ and $\boldsymbol{\Theta}$, $\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_*$ is a 1-Lipschitz function of a Gaussian matrix $\boldsymbol{Y}$. Thus utilizing concentration of Gaussian measure combined with (A.1) implies

$$\mathbb{P}\Big\{\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_* \ge 2\sqrt{dr} + t\Big\} \le \mathbb{P}\Big\{\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_* \ge \mathbb{E}\left[\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_*\right] + t\Big\} \le e^{-\frac{t^2}{2}}. \tag{A.3}$$

We will combine (A.3) above with an application of standard union bound. To this aim let $\mathcal{M} \subset \mathbb{S}^{dr-1}$ be an $\varepsilon = 1/4$ cover of $\mathbb{S}^{dr-1}$ and $\mathcal{S} \subset \mathbb{S}^{n-1}$ be a $\varepsilon = 1/4$ cover of $\mathbb{S}^{n-1}$ and note that based on standard covering bounds,

$$\log|\mathcal{S}| \le 3n \quad \text{and} \quad \log|\mathcal{M}| \le 3rd.$$

Using (A.3) with $t = 4\sqrt{dr}$ combined with the above covering bound we conclude that for $n \le dr$

$$\mathbb{P}\left\{\sup_{(\boldsymbol{v},\boldsymbol{\Theta})\in\mathcal{S}\times\mathcal{M}} \|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_* \ge 6\sqrt{dr}\right\} \le |\mathcal{S}|\cdot|\mathcal{M}|\cdot\mathbb{P}\Big\{\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_* \ge \mathbb{E}\left[\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_*\right] + 4\sqrt{dr}\Big\} \le e^{3n}\cdot e^{3rd}\cdot e^{-8rd} \le e^{-2rd}.$$

Thus for all $(\boldsymbol{v},\boldsymbol{\Theta}) \in \mathcal{S}\times\mathcal{M}$ we have $\|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_\star \le 6\sqrt{dr}$ with high probability. To extend this over the entire set $\mathbb{S}^{n-1}\times\mathbb{S}^{dr-1}$ define

$$(\boldsymbol{v}^\star, \boldsymbol{\Theta}^\star) := \underset{(\boldsymbol{v},\boldsymbol{\Theta})\in\mathbb{S}^{n-1}\times\mathbb{S}^{dr-1}}{\arg\sup} \|\Gamma_{\boldsymbol{v},\boldsymbol{\Theta}}\|_* \quad \text{and} \quad \text{OPT} = \|\Gamma_{\boldsymbol{\Theta}^\star,\boldsymbol{v}^\star}\|_* .$$

Now let $\widetilde{\boldsymbol{v}}$ and $\widetilde{\boldsymbol{\Theta}}$ be the closest points of the covers $\mathcal{S}$ and $\mathcal{M}$ to $\boldsymbol{v}^*$ and $\boldsymbol{\Theta}^*$ and note that $\|\widetilde{\boldsymbol{v}} - \boldsymbol{v}^*\|_{\ell_2} \le 1/4$ and $\left\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_F \le 1/4$. Thus, will probability at least $1 - e^{-2rd}$ we have

$$\begin{aligned} \text{OPT} =& \left\|\boldsymbol{\Gamma}_{\boldsymbol{v}^*,\boldsymbol{\Theta}^*-\widetilde{\boldsymbol{\Theta}}} + \boldsymbol{\Gamma}_{\boldsymbol{v}^*-\widetilde{v},\widetilde{\boldsymbol{\Theta}}} + \boldsymbol{\Gamma}_{\widetilde{\boldsymbol{v}},\widetilde{\boldsymbol{\Theta}}}\right\|_*, \\ \overset{(a)}{\le}& \left\|\boldsymbol{\Gamma}_{\boldsymbol{v}^*,\boldsymbol{\Theta}^*-\widetilde{\boldsymbol{\Theta}}}\right\|_* + \left\|\boldsymbol{\Gamma}_{\boldsymbol{v}^*-\widetilde{v},\widetilde{\boldsymbol{\Theta}}}\right\|_* + \left\|\boldsymbol{\Gamma}_{\widetilde{\boldsymbol{v}},\widetilde{\boldsymbol{\Theta}}}\right\|_*, \\ \overset{(b)}{\le}& \text{OPT}\left\|\boldsymbol{\Theta}^* - \widetilde{\boldsymbol{\Theta}}\right\|_F + \text{OPT}\|\boldsymbol{v}^* - \widetilde{\boldsymbol{v}}\|_{\ell_2} + \left\|\boldsymbol{\Gamma}_{\widetilde{\boldsymbol{v}},\widetilde{\boldsymbol{\Theta}}}\right\|_*, \\ \overset{(c)}{\le}& \frac{1}{2}\text{OPT} + 6\sqrt{dr}, \end{aligned}$$

which implies that $\text{OPT} = \|\Gamma_{\boldsymbol{\Theta}^\star,\boldsymbol{v}^\star}\|_* \le 12\sqrt{dr}$, completing the proof. In the above (a) follows from the triangular inequality, (b) from the linearity of $\boldsymbol{\Gamma}_{\boldsymbol{v},\boldsymbol{\Theta}}$ with respect to $\boldsymbol{v}$ and $\boldsymbol{\Theta}$ and the definition of OPT, and (c) from the bound on the cover.

A.2. Proof of Lemma 9.14

Note that for a Gaussian random vector $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ we have

$$\mathbb{E}\left[\|\boldsymbol{g}\|_{\ell_2}^4\right] = \mathbb{E}\left[\left(\sum_{g=1}^d \boldsymbol{g}_k^2\right)^2\right] = \sum_{k=1}^d \left(\mathbb{E}[\boldsymbol{g}_k^4] - (\mathbb{E}[\boldsymbol{g}_k^2])^2\right) + \left(\mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2]\right)^2 = d^2 + 2d$$

Using the above we can conclude that

$$\begin{aligned}
\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^4\,\right] =& \mathbb{E}\left[\left(\sum_{k=1}^{r}\boldsymbol{\Sigma}_{kk}^2\,\|\boldsymbol{G}_k\|_{\ell_2}^2\right)^2\right]\\
=& \sum_{k=1}^{r}\boldsymbol{\Sigma}_{kk}^4\left(\mathbb{E}\left[\,\|\boldsymbol{G}_k\|_{\ell_2}^4\,\right]-\mathbb{E}\left[\,\|\boldsymbol{G}_k\|_{\ell_2}^2\,\right]^2\right)+\left(\mathbb{E}\left[\,\sum_{k=1}^{r}\boldsymbol{\Sigma}_{kk}^2\,\|\boldsymbol{G}_k\|_{\ell_2}^2\,\right]\right)^2\\
=&2d\sum_{k=1}^{r}\boldsymbol{\Sigma}_{kk}^4+d^2\,\|\boldsymbol{\Sigma}\|_F^4\\
\le&(2d+d^2)\,\|\boldsymbol{\Sigma}\|_F^4\\
\le&3d^2\,\|\boldsymbol{\Sigma}\|_F^4\\
=&3\left(\mathbb{E}[\|\boldsymbol{G\Sigma}\|_F^2]\right)^2
\end{aligned}\tag{A.4}$$

Note that, using $\mathbb{E}[\|\boldsymbol{G}\|]\le\sqrt{d}+\sqrt{r}$,

$$\mathbb{P}\Big\{\,\|\boldsymbol{G\Sigma}\|\ge(\sqrt{d}+\sqrt{r}+t)\|\boldsymbol{\Sigma}\|\Big\}\le\mathbb{P}\Big\{\,\|\boldsymbol{G}\|\ge\sqrt{d}+\sqrt{r}+t\Big\}\le e^{-\frac{t^2}{2}}.$$

Define the event $\mathcal{E}=\left\{\boldsymbol{G}\in\mathbb{R}^{d\times r}:\|\boldsymbol{G\Sigma}\|\le2\vartheta\left(\sqrt{d}+\sqrt{r}\right)\right\}$. Using the above with $t=2\sqrt{r}$ we have

$$\mathbb{P}\Big\{\mathcal{E}^c\Big\}=\mathbb{P}\Big\{\,\|\boldsymbol{G\Sigma}\|\ge2\vartheta\left(\sqrt{d}+3\sqrt{r}\right)\Big\}=\mathbb{P}\Big\{\,\|\boldsymbol{G\Sigma}\|\ge2\vartheta\left(\sqrt{d}+\sqrt{r}+t\right)\Big\}\le e^{-\frac{t^2}{2}}=e^{-2r}.\tag{A.5}$$

Using these definitions we conclude that

$$\begin{aligned}
\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^2\,\right]\overset{(a)}{\le}&\,\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|\,\|\boldsymbol{G\Sigma}\|_*\,\right]\\
=&\,\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|\,(\mathbb{1}_{\mathcal{E}}+\mathbb{1}_{\mathcal{E}^c})\,\|\boldsymbol{G\Sigma}\|_*\,\right]\\
=&\,\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|\,\mathbb{1}_{\mathcal{E}}\,\|\boldsymbol{G\Sigma}\|_*\,\right]+\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|\,\mathbb{1}_{\mathcal{E}^c}\,\|\boldsymbol{G\Sigma}\|_*\,\right]\\
\overset{(b)}{\le}&2\vartheta\left(\sqrt{d}+3\sqrt{r}\right)\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*]+\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|\,\mathbb{1}_{\mathcal{E}^c}\,\|\boldsymbol{G\Sigma}\|_*\,\right]\\
\overset{(c)}{\le}&2\vartheta\left(\sqrt{d}+3\sqrt{r}\right)\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*]+\mathbb{E}\left[\sqrt{r}\mathbb{1}_{\mathcal{E}^c}\,\|\boldsymbol{G\Sigma}\|_F^2\,\right]\\
\overset{(d)}{\le}&2\vartheta\left(\sqrt{d}+3\sqrt{r}\right)\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*]+\sqrt{r}\sqrt{\mathbb{E}\left[\mathbb{1}_{\mathcal{E}^c}\right]}\sqrt{\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^4\,\right]}\\
\overset{(e)}{\le}&2\vartheta\left(\sqrt{d}+3\sqrt{r}\right)\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*]+\sqrt{3r}\sqrt{\mathbb{E}\left[\mathbb{1}_{\mathcal{E}^c}\right]}\,\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^2\,\right]\\
\overset{(f)}{\le}&2\vartheta\left(\sqrt{d}+3\sqrt{r}\right)\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*]+\sqrt{3re^{-2r}}\,\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^2\,\right]\\
\overset{(g)}{\le}&2\vartheta\left(\sqrt{d}+3\sqrt{r}\right)\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*]+\frac{3}{4}\,\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^2\,\right].
\end{aligned}$$

Here, (a) follows from Holder's inequality, (b) from (A.5), (c) from the fact that $\|\boldsymbol{G\Sigma}\|\le\|\boldsymbol{G\Sigma}\|_F$ and $\|\boldsymbol{G\Sigma}\|_*\le\sqrt{r}\,\|\boldsymbol{G\Sigma}\|_F$, (d) from Cauchy Schwarz, and (e) from (A.4), (f) from (A.5), and (g) from the fact that $\sqrt{3re^{-2r}}\le\frac{3}{4}$. The above chain of inequalities thus allow us to conclude that

$$\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^2\,\right]\le8\vartheta\left(\sqrt{d}+3\sqrt{r}\right).\,\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*]\le32\vartheta\sqrt{d}\,\mathbb{E}[\|\boldsymbol{G\Sigma}\|_*].$$

Combining the latter with the fact that $\mathbb{E}\left[\,\|\boldsymbol{G\Sigma}\|_F^2\,\right]=d\,\|\boldsymbol{\Sigma}\|_F^2\ge dr\vartheta^2$, concludes the proof.

A.3. Proof of Lemma 9.15

For the upper bound we use Lemma 9.13 together with the fact that $\|\boldsymbol{\Theta}\|_F\le2\vartheta\sqrt{r}$ to conclude that for all $\boldsymbol{v}\in\mathbb{S}^{n-1}$

$$\left\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)\right\|_*\le24\vartheta\sqrt{d}r,\tag{A.6}$$

holds with probability at least $1 - e^{-2dr}$.

We next turn our attention to the lower bound. Given the random nature of the matrices $\boldsymbol{X}_i$, $\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)$ defines a random process $\boldsymbol{\Gamma}_{\boldsymbol{v}}$ indexed by $\boldsymbol{v}$ which can be rewritten in the form

$$\boldsymbol{\Gamma}_{\boldsymbol{v}} := \text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right) = \sum_{i=1}^{n}\boldsymbol{v}_i\boldsymbol{X}_i\boldsymbol{\Theta}.$$

Thus, in this lemma we are interested in lower bounding $\inf_{\boldsymbol{v}\in\mathbb{S}^{n-1}}\|\Gamma_{\boldsymbol{v}}\|_*$ for a fixed $\boldsymbol{\Theta}$. To establish such bounds, we first determine the behavior of $\Gamma_{\boldsymbol{v}}$ for a fixed $\boldsymbol{v}$. Let $\boldsymbol{\Theta}$ have singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ with $\boldsymbol{U}\in\mathbb{R}^{d\times r}$ and set $\boldsymbol{Y} = \sum_{i=1}^n \boldsymbol{v}_i\boldsymbol{X}_i\boldsymbol{U}$ so that $\Gamma_{\boldsymbol{v}} = \boldsymbol{Y}\boldsymbol{\Sigma}\boldsymbol{V}^T$. By construction, for a fixed $\boldsymbol{v}\in\mathbb{S}^{n-1}$, the matrix $\boldsymbol{Y}\in\mathbb{R}^{d\times r}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Also note that by (A.2), $\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_* = \left\|\boldsymbol{Y}\boldsymbol{\Sigma}\boldsymbol{V}^T\right\|_*$ is a $\|\boldsymbol{\Sigma}\|_F \le 2\vartheta\sqrt{r}$ Lipschitz function of $\boldsymbol{Y}$. Also by Lemma 9.14, $\mathbb{E}[\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_*] \ge \frac{1}{32}\vartheta\sqrt{dr}$. Thus, by concentration of Lipschitz functions of Gaussian we have

$$\mathbb{P}\left\{\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_* \le \frac{1}{32}\vartheta\sqrt{dr} - t\right\} \le \mathbb{P}\left\{\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_* - \mathbb{E}[\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_*] \le -t\right\} \le e^{-\frac{t^2}{8r\vartheta^2}}.$$

Thus using $t = \frac{1}{288}\vartheta\sqrt{dr}$ we conclude that $\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_* \ge \frac{1}{36}\vartheta\sqrt{dr}$ holds with probability at least $1 - e^{-2\gamma dr}$ with γ a fixed numerical constant. Now pick a $\frac{1}{19000}$ cover $\mathcal{S}$ of $\mathbb{S}^{n-1}$. This cover size is at most $\log|\mathcal{S}| \le \log\left(\frac{3}{\frac{1}{19000}}\right)n \le 11n$. Thus using the union bound we conclude that for $n \le cdr := \frac{\gamma}{11}dr$ we have

$$\mathbb{P}\left\{\inf_{\boldsymbol{v}\in\mathcal{S}}\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_* \le \frac{1}{36}\vartheta\sqrt{dr}\right\} \le e^{11n}e^{-2\gamma dr} \le e^{-\gamma dr}.$$

To proceed, given any $\boldsymbol{v}\in\mathbb{S}^{n-1}$ denote the closest point from the cover $\mathcal{S}$ to this point by $\widetilde{\boldsymbol{v}}$. Using the fact that $\|\boldsymbol{v}-\widetilde{\boldsymbol{v}}\|_{\ell_2} \le \frac{1}{19000}$ combined with (A.6) we conclude that

$$\begin{aligned}\|\boldsymbol{\Gamma}_{\boldsymbol{v}}\|_* &\ge \|\boldsymbol{\Gamma}_{\widetilde{\boldsymbol{v}}}\|_* - \|\boldsymbol{\Gamma}_{\boldsymbol{v}-\widetilde{\boldsymbol{v}}}\|_*\\ &\ge \frac{1}{36}\vartheta\sqrt{dr} - \frac{24}{19000}\vartheta\sqrt{dr}\\ &\ge \frac{1}{40}\vartheta\sqrt{dr},\end{aligned}$$

holds with probability at least $1 - e^{-\gamma dr} - e^{-2dr} \ge 1 - 2e^{-\gamma dr}$.

A.4. Proof of Lemma 9.16

For any arbitrary $\boldsymbol{\Theta}\in\mathcal{D}$ and $\boldsymbol{v}\in\mathbb{S}^{n-1}$ using Lemma 9.13,

$$\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right) - \text{mat}\left(\mathcal{J}(\boldsymbol{\Theta}_0)\boldsymbol{v}\right)\|_\star = \|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0)\boldsymbol{v}\right)\|_\star \le 12\sqrt{dr}\,\|\boldsymbol{\Theta}-\boldsymbol{\Theta}_0\|_F\,,$$

holds with probability at least $1 - e^{-2dr}$. Using Lemma 9.15,

$$\frac{1}{40}\vartheta\sqrt{dr} \le \left\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta}_0)^T\boldsymbol{v}\right)\right\|_* \le 24\vartheta\sqrt{dr},$$

holds with probability at least $1 - 2e^{-\gamma dr}$. Combining the latter two bounds, using $\boldsymbol{\Theta}\in\mathcal{D}$ and definition of $\mathcal{D}$, and applying the triangle inequality we conclude that

$$\frac{1}{50}\vartheta\sqrt{dr} \le \left\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)\right\|_* \le 25\vartheta\sqrt{dr},$$

holds with probability at least $1 - 3e^{-\gamma dr}$. Using the fact that $\frac{\|\boldsymbol{A}\|_*}{\sqrt{r}} \le \|\boldsymbol{A}\|_F \le \|\boldsymbol{A}\|_*$ we thus have

$$\frac{1}{50}\vartheta\sqrt{dr} \le \left\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)\right\|_F \le 25\vartheta\sqrt{dr}.$$

Using the fact that $\left\|\text{mat}\left(\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\right)\right\|_F = \|\mathcal{J}(\boldsymbol{\Theta})^T\boldsymbol{v}\|_{\ell_2}$ and the result holds for all $\boldsymbol{v}$, completes the proof of (9.53).

To prove (9.54), note that on the same event applying Lemma 9.13, for any $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\in\mathbb{R}^{d\times r}$ we have

$$\|\mathcal{J}(\boldsymbol{\Theta}_2) - \mathcal{J}(\boldsymbol{\Theta}_1)\| = \sup_{\boldsymbol{v}\in\mathbb{S}^{n-1}}\|\text{mat}\left((\mathcal{J}(\boldsymbol{\Theta}_2) - \mathcal{J}(\boldsymbol{\Theta}_1))^T\boldsymbol{v}\right)\|_F \le 12\sqrt{dr}\,\|\boldsymbol{\Theta}_2-\boldsymbol{\Theta}_1\|_F\,,$$

concluding the proof of (9.54).