arXiv:1707.02444v1 [cs.LG] 8 Jul 2017

Global optimality conditions for deep neural networks

Chulhee Yun chulheey@mit.edu
Suvrit Sra suvrit@mit.edu
Ali Jadbabaie jadbabai@mit.edu
Massachusetts Institute of Technology

Abstract

We study the error landscape of deep linear and nonlinear neural networks with square error loss. We build on the recent results in the literature and present necessary and sufficient conditions for a critical point of the empirical risk function to be a global minimum in the deep *linear* network case. Our simple conditions can also be used to determine whether a given critical point is a global minimum or a saddle point. We further extend these results to deep *nonlinear* neural networks and prove similar sufficient conditions for global optimality in the function space.

1 Introduction

Since the advent of AlexNet [10], deep neural networks have surged in popularity, and have redefined the state-of-the-art across many application areas of machine learning and artificial intelligence, such as computer vision, speech recognition, and natural language processing.

Despite these successes, a concrete theoretical understanding of why deep neural networks work well in practice has remained elusive. From the perspective of optimization, a significant barrier is imposed by the nonconvexity of training neural networks. Moreover, it was proved by Blum and Rivest [3] that training even a simple 3-node neural network to global optimality is NP-complete in the worst case, so there is little hope that neural networks have properties that make global optimization tractable.

In spite of the difficulties of optimizing weights in neural networks, the empirical successes suggest that the local minima of their loss surfaces might be close to global minima. Several papers have recently appeared in the literature attempting to provide a theoretical justification for the success of these models. For example, by relating neural networks to spherical spin-glass models from statistical physics, Choromanska et al. [4] provided some empirical evidence that depth of neural networks makes the performance of local minima close to that of global minima.

Another line of recent results [13–15] provides conditions under which a critical point of the empirical risk is a global minimum. These type of results prove that if full rank conditions of some matrices (as well as some additional conditions) are satisfied, derivative of risk being zero implies error being zero. However, their results are obtained under restrictive assumptions; for example, Nguyen and Hein [13] requires the width of one of the hidden layers are as large as the number of training examples. Soudry and Carmon [14] and Xie et al. [15] also requires the product of widths of two adjacent layers are at least as large as the number of training examples, meaning that the number of parameters in the model must grow as we have more training data available.

A useful conceptual simplification of deep *nonlinear* networks is deep *linear* neural networks, in which all activation functions are linear and the output of the entire network is a chained product of weight matrices with the input vector. Although at first sight this model may look overly simplistic, even this problem is nonconvex, and only very recently theoretical results on it have started becoming available. Interestingly, already in 1989, Baldi and Hornik [1] showed that some shallow linear neural networks have no local minima. More recently, Kawaguchi [8] extended this result to deep networks and proved that any local minimum point is also a global minimum, and that any other critical point is a saddle point. Subsequently, Lu and Kawaguchi [11] provided a simpler proof that any local minimum is a global minimum, with fewer assumptions than [8]. Motivated by the success of deep residual networks [6, 7], Hardt and Ma [5] investigated loss surfaces of deep linear *residual networks* and show under certain assumptions on data distribution that every critical point is a global minimum in a near-identity region; very recently, Bartlett et al. [2] extended this result to a nonlinear function space setting.

1.1 Our contributions

Inspired by this recent line of work, we study deep linear and nonlinear networks, in settings either similar to or more general than existing work. Let us describe our main contributions in greater detail below to help place them in perspective.

Kawaguchi [8] considers a deep neural network with H hidden layers, where given a data matrix X, the output of the network is $W_{H+1}W_H \cdots W_1 X$. The author investigates the squared error risk function $L(W) = \frac{1}{2} \|W_{H+1}W_H \cdots W_1 X - Y\|_F^2$ and proves that every critical point of $L(W)$ is either a global minimum or a saddle point.

We generalize this result and provide necessary and sufficient conditions for a critical point of $L(W)$ to be a global minimum. In particular, in Theorem 2.1 we show that if the minimum-width layer of the deep linear network is either the input or the output layer, then a critical point of $L(W)$ is a global minimum if and only if the product $W_{H+1}W_H \cdots W_1$ is full-rank. This concise condition provides a checkable test whether a given critical point is a global minimum or a saddle point. Such efficiently checkable conditions on local optimality are in general impossible for nonconvex optimization; see e.g., [12] for an example where even checking whether a point is a local minimum is NP-complete.

In Theorem 2.2, we consider the case where some hidden layers can have smaller width than both the input and output layers, and provide similar necessary and sufficient conditions for global optimality.

While Kawaguchi [8] and our paper consider minimizing the empirical risk, Hardt and Ma [5] consider minimizing population risk, but under a simple linear model with Gaussian noise assumption on the data distribution. With similar assumptions, our Theorem 2.1 can also be modified to handle the population risk. Doing so, our result extends [5, Theorem 2.2] to a strictly *larger* set, while removing the assumption that the true underlying linear model has a positive determinant.

In the Computational Challenges in Machine Learning workshop at Simons Institute for the Theory of Computing, Peter Bartlett gave an interesting talk [2] about extending Hardt and Ma [5] to nonlinear neural networks, and outlined results on decomposition of a smooth nonlinear function into a composition of near-identity functions, as well as on global optimality of critical points for near-identity functions. Motivated by his talk, we extended our results on linear networks to obtain sufficient conditions for nonlinear networks; these are presented in Theorems 4.1 and 4.2.

2 Global optimality conditions for deep linear neural networks

In this section, we describe the problem formulation and notations for deep linear neural networks, state main results (Theorems 2.1 and 2.2), and explain their implication.

2.1 Problem formulation and notations

Suppose we have m input-output pairs, where the inputs are of dimension d_x and outputs of dimension d_y. Let $X \in \mathbb{R}^{d_x \times m}$ be the data matrix and $Y \in \mathbb{R}^{d_y \times m}$ be the output matrix. Suppose we have H hidden layers in the network, each having width $d_1, \dots, d_H$. For notational simplicity we let $d_0 = d_x$ and $d_{H+1} = d_y$. The weights between adjacent layers can be represented as matrices $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$, for $k = 1, \dots, H+1$, and the output $\hat{Y}$ of the network can be written as matrix multiplication with X:

$$\hat{Y} = W_{H+1} W_H \cdots W_1 X.$$

We consider minimizing the squared error loss over all data points,

$$L(W) = \frac{1}{2} \left\| \hat{Y} - Y \right\|_{\mathrm{F}}^2 = \frac{1}{2} \left\| W_{H+1} W_H \cdots W_1 X - Y \right\|_{\mathrm{F}}^2,$$

where W is a shorthand notation for $(W_1, \dots, W_{H+1})$. We are interested in understanding the loss surface of the empirical risk $L(W)$ by minimizing

$$\text{minimize} \quad \tfrac{1}{2} \left\| W_{H+1} \cdots W_1 X - Y \right\|_{\mathrm{F}}^2. \tag{2.1}$$

Assumptions. We assume that $d_x \leq m$ and $d_y \leq m$, and that XX^T and YX^T have full ranks. We also assume that the singular values of $YX^T(XX^T)^{-1}X$ are all distinct, which is made for notational simplicity and can be relaxed without too much difficulty.

Notations. Given a matrix A, let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote the largest and smallest singular values of A, respectively. Let $\mathrm{row}(A)$, $\mathrm{col}(A)$, $\mathrm{null}(A)$, and $\mathrm{rank}(A)$ be respectively the row space, column space, null space, and rank of matrix A. Given a subspace V of $\mathbb{R}^n$, we denote $V^\perp$ as its orthogonal complement.

Let us denote $p = \mathrm{argmin}_{i \in \{0,\dots,H+1\}} d_i$, and $k = \min_{i \in \{0,\dots,H+1\}} d_i$. That is, p is the layer with the smallest width, and $k = d_p$ is the width of that layer. Notice that the product $W_{H+1} \cdots W_1$ can have rank at most k.

Let $YX^T(XX^T)^{-1}X = U\Sigma V^T$ be the singular value decomposition of $YX^T(XX^T)^{-1}X \in \mathbb{R}^{d_y \times d_x}$. Let $\hat{U} \in \mathbb{R}^{d_y \times k}$ be a matrix consisting of the first k columns of U.

2.2 Necessary and sufficient conditions for global optimality

We now present two main theorems for deep linear neural networks. The theorems present two sets, one for the case of $k = \min\{d_x, d_y\}$ and the other for $k < \min\{d_x, d_y\}$, in which every critical point of $L(W)$ is a global minimum. Moreover, the sets have another desirable property that every critical point outside of these sets is a saddle point. Kawaguchi [8] and Lu and Kawaguchi [11] showed that any local minimum is a global minimum, and any other critical points are saddle points. In this paper, we are partitioning the domain of $L(W)$ into two sets clearly delineating one set which only contains global minima and the other set with only saddle points.

Theorem 2.1. *If $k = \min\{d_x, d_y\}$, define the following set*

$$\mathcal{V}_1 = \{(W_1, \dots, W_{H+1}) : \mathrm{rank}(W_{H+1} \cdots W_1) = k\}.$$

Then, every critical point of $L(W)$ in $\mathcal{V}_1$ is a global minimum. Moreover, every critical point of $L(W)$ in $\mathcal{V}_1^c$ is a saddle point.

Theorem 2.2. *If $k < \min\{d_x, d_y\}$, define the following set*

$$\mathcal{V}_2 = \left\{(W_1, \dots, W_{H+1}) : \mathrm{rank}(W_{H+1} \cdots W_1) = k, \mathrm{col}(W_{H+1} \cdots W_{p+1}) = \mathrm{col}(\hat{U})\right\}.$$

Then, every critical point of $L(W)$ in $\mathcal{V}_2$ is a global minimum. Moreover, every critical point of $L(W)$ in $\mathcal{V}_2^c$ is a saddle point.

Theorems 2.1 and 2.2 provide necessary and sufficient conditions for a critical point of $L(W)$ to be global optimal. From an algorithmic perspective, they provide easily checkable conditions, which we can use to determine if the critical point the algorithm encountered is a global optimum or not.

In Hardt and Ma [5], the authors consider minimizing population risk of deep linear residual networks, which in our notation can be written as

$$\text{minimize} \quad \tfrac{1}{2}\mathbb{E}_{x,y}\left[\|(I+W_{H+1})\cdots(I+W_1)x-y\|_{\text{F}}^2\right], \tag{2.2}$$

where $d_x = d_1 = \cdots = d_H = d_y = d$ and x, y are random vectors drawn from a joint distribution $\mathcal{D}$. They assume that x is drawn from a zero-mean distribution with a certain covariance matrix, and $y = Rx + \xi$ where ξ is iid standard Gaussian noise and R is the true underlying matrix with $\det(R) > 0$. With these assumptions they prove that whenever $\sigma_{\max}(W_i) < 1$ for all i, any critical point is a global minimum [5, Theorem 2.2].

Under the same assumptions on data distribution, we can slightly modify Theorem 2.1 into a population risk counterpart, and one can notice that the result proved in Hardt and Ma [5] is a corollary of this version because having $\sigma_{\max}(W_i) < 1$ for all i is a sufficient condition for $(I+W_{H+1})\cdots(I+W_1)$ having full rank. Moreover, notice that we can remove the assumption $\det(R) > 0$ on the true matrix. We state this special case as a corollary:

Corollary 2.3 (Theorem 2.2 of Hardt and Ma [5])**.** *Under assumptions on data distribution as described above, any critical point of* $\frac{1}{2}\mathbb{E}_{x,y}\left[\|(I+W_{H+1})\cdots(I+W_1)x-y\|_{\text{F}}^2\right]$ *is a global minimum if* $\sigma_{\max}(W_i) < 1$ *for all* i.

Remark. The previous result [8] assumed that $d_y \leq d_x$ in order to show that: (1) every local minimum is a global minimum, and (2) any other critical point is a saddle point. A subsequent paper by Lu and Kawaguchi [11] proved (1) without the assumption $d_y \leq d_x$, but as far as we know there is no result showing (2) in the case of $d_y > d_x$. We provide the proof for this case in Lemma B.3.

3 Analysis of deep linear networks

In this section, we provide proofs for Theorems 2.1 and 2.2.

3.1 Solutions of the relaxed problem

We first analyze the global optimal solution of a "relaxation" of $L(W)$, which turns out to be very useful while proving Theorems 2.1 and 2.2. Consider a relaxed risk function

$$L_0(R) = \frac{1}{2}\|RX - Y\|_{\text{F}}^2,$$

where $R \in \mathbb{R}^{d_y \times d_x}$ and $\text{rank}(R) \leq k$. For any W, the product $W_{H+1}W_H\cdots W_1$ has rank at most k and setting R to be this product gives the same loss values: $L_0(W_{H+1}W_H\cdots W_1) = L(W)$. Therefore, L_0 is a relaxation of L and

$$\inf_{R:\text{rank}(R)\leq k} L_0(R) \leq \inf_{W} L(W).$$

This means that if there exists W such that $L(W) = \inf_{R:\text{rank}(R)\leq k} L_0(R)$, then W is a global minimum of the function L. This observation is very important in proofs; we will show that inside certain sets, any critical point W of $L(W)$ must satisfy $R^* = W_{H+1} \cdots W_1$, where R^* is a global optimum of $L_0(R)$. This proves that $L(W) = L_0(R^*) = \inf_{R:\text{rank}(R) \le k} L_0(R)$, thus showing that W is a global minimum of L.

By restating this in term of an optimization problem, the solution of problem in Equation 2.1 is bounded below by the minimum value of the following:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} \|RX - Y\|_{\text{F}}^2 \\ \text{subject to} & \text{rank}(R) \le k. \end{array} \tag{3.1}$$

In case where $k = \min\{d_x, d_y\}$, Equation 3.1 is actually an unconstrained problem. Note that L_0 is a differentiable convex function of R, so any critical point is a global minimum. By differentiating and setting the derivative to zero, we can easily get the unique global optimal solution

$$R^* = YX^T(XX^T)^{-1}. \tag{3.2}$$

In case of $k < \min\{d_x, d_y\}$, the problem becomes non-convex because of the rank constraint, but the exact solution can be computed easily. We present the solution of this case as a proposition and defer the proof to Appendix C.

Proposition 3.1. *Suppose $k < \min\{d_x, d_y\}$. Then the optimal solution of Equation 3.1 is*

$$R^* = \hat{U}\hat{U}^T YX^T(XX^T)^{-1}, \tag{3.3}$$

which is the orthogonal projection of $YX^T(XX^T)^{-1}$ to the column space of $\hat{U}$.

3.2 Partial derivatives of $L(W)$

By straight-up matrix calculus, we can calculate the derivatives of $L(W)$ with respect to W_i's. We present the result as the following lemma, and defer the proof to Appendix C.

Lemma 3.2. *The partial derivative of $L(W)$ with respect to W_i is given as*

$$\frac{\partial L}{\partial W_i} = W_{i+1}^T \cdots W_{H+1}^T (W_{H+1}W_H \cdots W_1 X - Y)X^T W_1^T \cdots W_{i-1}^T, \tag{3.4}$$

for $i = 1, \ldots, H+1$.

In case of $i = 1$, let $W_1^T \cdots W_0^T$ be an identity matrix in $\mathbb{R}^{d_x \times d_x}$. Similarly, $W_{H+2}^T \cdots W_{H+1}^T$ is an identity matrix in $\mathbb{R}^{d_y \times d_y}$. This result will be used throughout the proof of Theorems 2.1 and 2.2.

3.3 Proof of Theorem 2.1

We prove Theorem 2.1, which addresses the case $k = \min\{d_x, d_y\}$. First off, recall the set defined in Theorem 2.1:

$$\mathcal{V}_1 = \{(W_1, \ldots, W_{H+1}) : \text{rank}(W_{H+1} \cdots W_1) = k\}.$$

As seen in Equation 3.2, the unique minimum point of L_0 has rank k. Therefore, any point $W \in \mathcal{V}_1^c$ cannot be a global minimum of L. By [8, Theorem 2.3.(iii)] and Lemma B.3, then any critical point in $\mathcal{V}_1^c$ is a saddle point.

For the rest of the proof, we need to consider two cases: $d_y \le d_x$ and $d_x \le d_y$. If $d_x = d_y$, both cases work. The outline of the proof is as follows: we show that any critical point in a set $\mathcal{W}_\epsilon$ is a global minimum, and then show that every $W \in \mathcal{V}_1$ is in $\mathcal{W}_\epsilon$ for some $\epsilon > 0$. This proves that any critical point of $L(W)$ in $\mathcal{V}_1$ is also in $\mathcal{W}_\epsilon$ for some $\epsilon > 0$, hence a global minimum.

The following proposition proves the first step:

Proposition 3.3. *Assume that $k = \min\{d_x, d_y\}$. For any $\epsilon > 0$, define the following set of W:*

$$\mathcal{W}_\epsilon = \begin{cases} \{(W_1, \dots, W_{H+1}) : \sigma_{\min}(W_{H+1} \cdots W_2) \geq \epsilon\}, & \text{if } d_y \leq d_x, \\ \{(W_1, \dots, W_{H+1}) : \sigma_{\min}(W_H \cdots W_1) \geq \epsilon\}, & \text{if } d_x \leq d_y. \end{cases}$$

Then any critical point of $L(W)$ in $\mathcal{W}_\epsilon$ is a global minimum point.

Proof. (If $d_y \leq d_x$) Consider Equation 3.4 in the case of $i = 1$. We can observe that $W_2^T \cdots W_{H+1}^T \in \mathbb{R}^{d_1 \times d_y}$ and that $d_1 \geq d_y$. Then by Lemma B.1,

$$\begin{aligned} \left\| \frac{\partial L}{\partial W_1} \right\|_F^2 &\geq \sigma_{\min}^2(W_{H+1} \cdots W_2) \left\| (W_{H+1} W_H \cdots W_1 X - Y) X^T \right\|_F^2 \\ &\geq \epsilon^2 \left\| (W_{H+1} W_H \cdots W_1 X - Y) X^T \right\|_F^2. \end{aligned}$$

By the above inequality, any critical point in $\mathcal{W}$ satisfies

$$\forall i, \frac{\partial L}{\partial W_i} = 0 \Rightarrow (W_{H+1} W_H \cdots W_1 X - Y) X^T = 0,$$

which means that $W_{H+1} W_H \cdots W_1 = Y X^T (X X^T)^{-1}$. The product is the unique global optimal solution (Equation 3.2) of the relaxed problem in Equation 3.1, so W is a global minimum point of L.

(If $d_x \leq d_y$) Consider Equation 3.4 now in the case of $i = H + 1$. We can observe that $W_1^T \cdots W_H^T \in \mathbb{R}^{d_x \times d_H}$ and that $d_x \leq d_H$. Then by Lemma B.2,

$$\left\| \frac{\partial L}{\partial W_{H+1}} \right\|_F^2 \geq \epsilon^2 \left\| (W_{H+1} W_H \cdots W_1 X - Y) X^T \right\|_F^2,$$

and the rest of the proof goes exactly the same way as the previous case. □

The next proposition proves the theorem:

Proposition 3.4. *For any point $W \in \mathcal{V}_1$, there exists an $\epsilon > 0$ such that $W \in \mathcal{W}_\epsilon$.*

Proof. Define a new set $\mathcal{W}$, a "limit" version (as $\epsilon \to 0$) of $\mathcal{W}_\epsilon$, as

$$\mathcal{W} = \begin{cases} \{(W_1, \dots, W_{H+1}) : \operatorname{rank}(W_{H+1} \cdots W_2) = d_y\}, & \text{if } d_y \leq d_x, \\ \{(W_1, \dots, W_{H+1}) : \operatorname{rank}(W_H \cdots W_1) = d_x\}, & \text{if } d_x \leq d_y. \end{cases}$$

We show that $\mathcal{V}_1 \subset \mathcal{W}$ by showing that $\mathcal{W}^c \subset \mathcal{V}_1^c$. Consider

$$\mathcal{W}^c = \begin{cases} \{(W_1, \dots, W_{H+1}) : \operatorname{rank}(W_{H+1} \cdots W_2) < d_y\}, & \text{if } d_y \leq d_x, \\ \{(W_1, \dots, W_{H+1}) : \operatorname{rank}(W_H \cdots W_1) < d_x\}, & \text{if } d_x \leq d_y. \end{cases}$$

Then any $W \in \mathcal{W}^c$ must have $\operatorname{rank}(W_{H+1} \cdots W_1) < \min\{d_x, d_y\} = k$, so $W \in \mathcal{V}_1^c$. Thus, any $W \in \mathcal{V}_1$ is also in $\mathcal{W}$, so either $\operatorname{rank}(W_{H+1} \cdots W_2) = d_y$ or $\operatorname{rank}(W_H \cdots W_1) = d_x$, depending on the cases. Then, take

$$\epsilon = \begin{cases} \sigma_{\min}(W_{H+1} \cdots W_2), & \text{if } d_y \leq d_x, \\ \sigma_{\min}(W_H \cdots W_1), & \text{if } d_x \leq d_y. \end{cases}$$

We have $\epsilon > 0$ because the matrices are full rank, and we can see that $W \in \mathcal{W}_\epsilon$. □

3.4 Proof of Theorem 2.2

In this section we prove Theorem 2.2, which considers the case $k < \min\{d_x, d_y\}$. Note that this assumption also implies that $1 \leq p \leq H$.

As done in the proof of Theorem 2.1, define

$$\mathcal{V}_1 = \{(W_1, \dots, W_{H+1}) : \operatorname{rank}(W_{H+1} \cdots W_1) = k\}.$$

The global optimal point of the relaxed problem (Equation 3.1) has rank k, as seen in Equation 3.3. Thus, any point outside of $\mathcal{V}_1$ cannot be a global minimum. Then, by [8, Theorem 2.3.(iii)] and Lemma B.3, it follows that any critical point in $\mathcal{V}_1^c$ is a saddle point. The remaining proof considers points in $\mathcal{V}_1$.

For this section, let us introduce some additional notations, for the sake of simplicity. Define

$$\begin{aligned} &E = (W_{H+1} \cdots W_1 X - Y) X^T \in \mathbb{R}^{d_y \times d_x}, \\ &A_i = W_{i+1}^T \cdots W_{H+1}^T \in \mathbb{R}^{d_i \times d_y},\ B_i = W_1^T \cdots W_{i-1}^T \in \mathbb{R}^{d_x \times d_{i-1}}, \qquad i = 1, \dots, H+1, \end{aligned}$$

so that $\frac{\partial L}{\partial W_i} = A_i E B_i$ as we can check by comparing with Equation 3.4. Notice that A_{H+1} and B_1 are identity matrices.

Now consider any $W \in \mathcal{V}_1$. Since the full product $W_{H+1} \cdots W_1$ has rank k, any partial products A_i and B_i must have $\operatorname{rank}(A_i) \geq k$ and $\operatorname{rank}(B_i) \geq k$, for all i. Then, consider $A_p \in \mathbb{R}^{k \times d_y}$ and $B_{p+1} \in \mathbb{R}^{d_x \times k}$. Since $\operatorname{rank}(A_p) \leq k$ and $\operatorname{rank}(B_{p+1}) \leq k$, we can see that $\operatorname{rank}(A_p) = \operatorname{rank}(B_{p+1}) = k$. Also, notice that $A_i = W_{i+1} A_{i+1}$ and $B_{i+1} = B_i W_i$, so

$$\operatorname{rank}(A_1) \leq \operatorname{rank}(A_2) \leq \cdots \leq \operatorname{rank}(A_p) \text{ and } \operatorname{rank}(B_{H+1}) \leq \operatorname{rank}(B_H) \leq \cdots \leq \operatorname{rank}(B_{p+1}),$$

but we have $k \leq \operatorname{rank}(A_1)$ and $k \leq \operatorname{rank}(B_{H+1})$, so the ranks are all identically k. Moreover,

$$\operatorname{row}(A_1) \subset \operatorname{row}(A_2) \subset \cdots \subset \operatorname{row}(A_p) \text{ and } \operatorname{col}(B_{H+1}) \subset \operatorname{col}(B_H) \subset \cdots \subset \operatorname{col}(B_{p+1}),$$

but it was just shown that the these spaces have the same dimensions, which is the rank k, meaning

$$\operatorname{row}(A_1) = \operatorname{row}(A_2) = \cdots = \operatorname{row}(A_p) \text{ and } \operatorname{col}(B_{H+1}) = \operatorname{col}(B_H) = \cdots = \operatorname{col}(B_{p+1}).$$

Using these notations and facts, we state a proposition showing necessary and sufficient conditions of $W \in \mathcal{V}_1$ being a critical point of $L(W)$.

Proposition 3.5. *A point $W \in \mathcal{V}_1$ is a critical point of $L(W)$ if and only if $A_p E = 0$ and $E B_{p+1} = 0$.*

Proof. (If part) $A_p E = 0$ implies that $\operatorname{col}(E) \subset \operatorname{row}(A_p)^\perp = \cdots = \operatorname{row}(A_1)^\perp$, so $\frac{\partial L}{\partial W_i} = A_i E B_i = 0 \cdot B_i = 0$, for $i = 1, \dots, p$. Similarly, $E B_{p+1} = 0$ implies $\operatorname{row}(E) \subset \operatorname{col}(B_{p+1})^\perp = \cdots = \operatorname{col}(B_{H+1})^\perp$, so $\frac{\partial L}{\partial W_i} = A_i E B_i = A_i \cdot 0 = 0$ for $i = p+1, \dots, H+1$.

(Only if part) We have $\frac{\partial L}{\partial W_i} = A_i E B_i = 0$ for all i. This means that

$$\begin{aligned} \operatorname{col}(E B_i) \subset \operatorname{row}(A_i)^\perp = \operatorname{row}(A_p)^\perp \text{ for } i = 1, \dots, p \\ \operatorname{row}(A_i E) \subset \operatorname{col}(B_i)^\perp = \operatorname{col}(B_{p+1})^\perp \text{ for } i = p+1, \dots, H+1. \end{aligned}$$

Now recall that B_1 and A_{H+1} are identity matrices, so $\operatorname{col}(E) \subset \operatorname{row}(A_p)^\perp$ and $\operatorname{row}(E) \subset \operatorname{col}(B_{p+1})^\perp$, which proves $A_p E = 0$ and $E B_{p+1} = 0$. □

Now we present a proposition that specifies the necessary and sufficient condition in which a critical point of $L(W)$ in $\mathcal{V}_1$ is a global minimum. Recall that when we take the SVD of $Y X^T (X X^T)^{-1} X = U \Sigma V^T$, $\hat{U} \in \mathbb{R}^{d_y \times k}$ is defined to be a matrix consisting of the first k columns of U.

Proposition 3.6. *A critical point $W \in \mathcal{V}_1$ of $L(W)$ is a global minimum point if and only if* $\operatorname{col}(W_{H+1} \cdots W_{p+1}) = \operatorname{row}(A_p) = \operatorname{col}(\hat{U})$.

Proof. Since W is a critical point, by Proposition 3.5 we have $A_p E = 0$. Also note from the definitions of A_i's and B_i's that $W_{H+1} \cdots W_1 = A_p^T B_{p+1}^T$, so

$$A_p E = A_p (A_p^T B_{p+1}^T X - Y) X^T = A_p A_p^T B_{p+1}^T X X^T - A_p Y X^T = 0.$$

Because $\operatorname{rank}(A_p) = k$, and $A_p A_p^T \in \mathbb{R}^{k \times k}$ is invertible, so B_{p+1} is determined uniquely as

$$B_{p+1}^T = (A_p A_p^T)^{-1} A_p Y X^T (X X^T)^{-1},$$

thus

$$W_{H+1} \cdots W_1 = A_p^T B_{p+1}^T = A_p^T (A_p A_p^T)^{-1} A_p Y X^T (X X^T)^{-1}.$$

Comparing this with Equation 3.3, W is a global minimum solution if and only if

$$\hat{U} \hat{U}^T Y X^T (X X^T)^{-1} = W_{H+1} \cdots W_1 = A_p^T (A_p A_p^T)^{-1} A_p Y X^T (X X^T)^{-1}.$$

This equation holds if and only if $A_p^T (A_p A_p^T)^{-1} A_p = \hat{U} \hat{U}^T$, meaning that they are projecting $Y X^T (X X^T)^{-1}$ onto the same subspace. The projection matrix $A_p^T (A_p A_p^T)^{-1} A_p$ is onto $\operatorname{row}(A_p)$, while $\hat{U} \hat{U}^T$ is onto $\operatorname{col}(\hat{U})$. From this, we conclude that W is a global minimum point if and only if $\operatorname{row}(A_p) = \operatorname{col}(\hat{U})$. □

From Proposition 3.6, we can define the set $\mathcal{V}_2$ that appeared in Theorem 2.2, and conclude that every critical point of $L(W)$ in $\mathcal{V}_2$ is a global minimum, and any other critical points are saddle points.

4 Extension to deep nonlinear neural networks

In this section, we build on results of the previous sections on deep linear networks and use the setup of a recent talk [2] where new results on the extension of Hardt and Ma [5] to nonlinear neural networks were presented. Given a smooth nonlinear function h that maps input to output, Bartlett et al. [2] describes a method to decompose it into a number of smooth nonlinear functions $h = h_{H+1} \circ \cdots \circ h_1$ where h_i's are close to identity. Using Fréchet derivatives of the population risk with respect to each function h_i, he shows that when all h_i's are close to identity, any critical point of the population risk is a global minimum. One can see that these results are direct generalization of Theorems 2.1 and 2.2 of Hardt and Ma [5] to nonlinear networks and utilize the classical "small gain" arguments often used in nonlinear analysis and control [9, 16]. Motivated by this result, we extended Theorem 2.1 to deep nonlinear neural networks and obtained some sufficient conditions for global optimality. This section describes the results and discusses their implication.

4.1 Problem formulation and notations

Suppose the data $X \in \mathbb{R}^{d_x}$ and its corresponding label $Y \in \mathbb{R}^{d_y}$ are drawn from some distribution. Notice that in this section, X and Y are random vectors instead of matrices. We want to predict Y given X with a deep nonlinear neural network that has H hidden layers. Each layer takes d_{i-1}-dimensional input from the previous layer and produces d_i-dimensional output, which we can express as functions $h_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$. So the entire neural network can be expressed as a composition of these layers: $h_{H+1} \circ h_H \circ \cdots \circ h_1$. Our goal is to obtain functions $h_1, \ldots, h_{H+1}$ that minimize the population risk functional:

$$L(h) = L(h_1, \ldots, h_{H+1}) = \frac{1}{2}\mathbb{E}\left[\|h_{H+1} \circ \cdots \circ h_1(X) - Y\|_2^2\right],$$

where h is a shorthand notation for $(h_1, \ldots, h_{H+1})$. It is well-known that the minimizer of squared error risk is the conditional expectation of Y given X, which we will denote $h^*(x) = \mathbb{E}[Y \mid X = x]$. With this, we can separate the risk functional into two terms:

$$L(h) = \frac{1}{2}\mathbb{E}\left[\|h_{H+1} \circ \cdots \circ h_1(X) - h^*(X)\|_2^2\right] + C,$$

where the constant C denotes the variance term that is independent of $h_1, \ldots, h_{H+1}$. Note that if $h_{H+1} \circ \cdots \circ h_1 = h^*$ almost surely, the first term in $L(h)$ vanishes and the optimal value L^* of $L(h)$ is C.

Assumptions. Define the function spaces as the following:

$$\mathcal{F} = \left\{h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y} \mid h \text{ is differentiable, } h(0) = 0, \text{ and } \sup_x \frac{\|h(x)\|_2}{\|x\|_2} < \infty\right\},$$
$$\mathcal{F}_i = \left\{h : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i} \mid h \text{ is differentiable, } h(0) = 0, \text{ and } \sup_x \frac{\|h(x)\|_2}{\|x\|_2} < \infty\right\},$$

where $\mathcal{F}_i$ are defined for all $i = 1, \ldots, H+1$. Assume that $h^* \in \mathcal{F}$, and that we are optimizing $L(h)$ with $h_1 \in \mathcal{F}_1, \ldots, h_{H+1} \in \mathcal{F}_{H+1}$. In other words, the functions in $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_{H+1}$ are globally Lipschitz and show sublinear growth starting from 0. Notice that $h_{H+1} \circ \cdots \circ h_1 \in \mathcal{F}$, because a composition of differentiable functions is also differentiable, and a composition of sublinear functions is also sublinear.

We assume that $d_i \geq \min\{d_x, d_y\}$ for all $i = 1, \ldots, H+1$, which is identical to the assumption $k = \min\{d_x, d_y\}$ in Theorem 2.1.

Notations. To simplify multiple composition of functions, we denote $h_{i:j} = h_i \circ h_{i-1} \circ \cdots \circ h_{j+1} \circ h_j$. As in the matrix case, let $h_{0:1}$ and $h_{H+1:H+2}$ be identity maps in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively.

Given a function f, let $J[f](x)$ be the Jacobian matrix of function f evaluated at x. Let $D_{h_i}[L(h)]$ be the Fréchet derivative of $L(h)$ with respect to h_i evaluated at h. The Fréchet derivative $D_{h_i}[L(h)]$ is a linear functional that maps a function (direction) $\eta \in \mathcal{F}_i$ to a real number (directional derivative).

4.2 Sufficient conditions for global optimality

Here, we present two theorems which give sufficient conditions for a critical point ($D_{h_i}[L(h)] = 0$ for all i) in the function space to be a global optimum.

Theorem 4.1. *Consider the case $d_x \geq d_y$. If there exists $\epsilon > 0$ such that*

1. *$J[h_{H+1:2}](z) \in \mathbb{R}^{d_y \times d_1}$ has $\sigma_{\min}(J[h_{H+1:2}](z)) \geq \epsilon$ for all $z \in \mathbb{R}^{d_1}$,*
2. *$h_{H+1:2}(z)$ is twice-differentiable,*

then any critical point of $L(h)$ is a global minimum.

Theorem 4.2. *Consider the case $d_x \leq d_y$. Assume that there exists some $j \in \{1, \ldots, H+1\}$ such that $d_x = d_{j-1}$ and $d_y \leq d_j$. If there exist $\epsilon_1, \epsilon_2 > 0$ such that*

1. *$h_{j-1:1} : \mathbb{R}^{d_x} \to \mathbb{R}^{d_{j-1}} = \mathbb{R}^{d_x}$ is invertible,*
2. *$h_{j-1:1}$ satisfies $\left\|h_{j-1:1}(u)\right\|_2 \geq \epsilon_1 \left\|u\right\|_2$ for all $u \in \mathbb{R}^{d_x}$,*
3. *$J[h_{H+1:j+1}](z) \in \mathbb{R}^{d_y \times d_j}$ has $\sigma_{\min}(J[h_{H+1:j+1}](z)) \geq \epsilon_2$ for all $z \in \mathbb{R}^{d_j}$,*
4. *$h_{H+1:j+1}(z)$ is twice-differentiable,*

then any critical point of $L(h)$ is a global minimum.

Note that these theorems give *sufficient* conditions, whereas Theorems 2.1 and 2.2 provide *necessary and sufficient* conditions. So, if the sets we are describing in Theorems 4.1 and 4.2 do not contain any critical point, the claims would be vacuous. We ensure that there are critical points in the sets, by proving the following proposition.

Proposition 4.3. *For each of Theorems 4.1 and 4.2, there exists at least one global minimum solution of $L(h)$ satisfying the conditions of the theorem.*

Theorems 4.1 and 4.2 state that in certain sets of $(h_1, \ldots, h_{H+1})$, any critical point in function space a global minimum. However, this does not imply that any critical point for a fixed sigmoid or arctan network is a global minimum. As noted in [2], there is a downhill direction in function space at any suboptimal points, but this direction might be orthogonal to the function space represented by a fixed network, hence result in local minima in the parameter space of the fixed architecture.

Bartlett et al. [2] made some assumptions on the function spaces including the following: the function is invertible and there exists a point where the Jacobian matrix has positive determinant, which correspond to the assumption that $\det(R) > 0$ in Hardt and Ma [5]. Please note that in our setup we do not require such assumptions on h^*.

The proof of Theorems 4.1, 4.2, and Proposition 4.3 are deferred to Appendix A.

References

[1] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[2] P. Bartlett, S. Evans, and P. Long. Deep residual networks: Representation and optimization properties, 2017. Talk by Peter Bartlett at the Computational Challenges in Machine Learning Workshop at Simons Institute for the Theory of Computing, Berkeley, CA, USA.

[3] A. Blum and R. L. Rivest. Training a 3-node neural network is np-complete. In *Proceedings of the 1st International Conference on Neural Information Processing Systems*, pages 494–501. MIT Press, 1988.

[4] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

[5] M. Hardt and T. Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[8] K. Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[9] H. K. Khalil. *Noninear Systems*. Prentice-Hall, New Jersey, 1996.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] H. Lu and K. Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

[12] K. G. Murty and S. N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[13] Q. Nguyen and M. Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

[14] D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[15] B. Xie, Y. Liang, and L. Song. Diverse neural network learns true target functions. *arXiv preprint arXiv:1611.03131*, 2016.

[16] G. Zames. On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity. *IEEE transactions on automatic control*, 11(2):228–238, 1966.

A Analysis of deep nonlinear networks

A.1 Notations

In this section, we introduce additional notations that are used in the proofs. To emphasize that the Fréchet derivative $D_{h_i}[L(h)]$ is a linear functional that outputs a real number, we will write $D_{h_i}[L(h)](\eta)$ in an inner-product form $\langle D_{h_i}[L(h)], \eta \rangle$. This notation also helps avoiding confusion coming from multiple parentheses and square brackets.

There are many different kinds of norms that appear in the proofs. Given a finite-dimensional real vector v, $\|v\|_2$ denotes its ℓ_2 norm. For a matrix A, its operator norm is defined as $\|A\|_{\text{op}} = \sup_x \frac{\|Ax\|_2}{\|x\|_2}$. Let $h \in \mathcal{F}$. Then define a "generalized" induced norm for nonlinear functions with sublinear growth: $\|h\|_{\text{nl}} = \sup_x \frac{\|h(x)\|_2}{\|x\|_2}$, where the subscript nl is used to emphasize that this norm is for nonlinear functions. The norm $\|\cdot\|_{\text{nl}}$ is defined in the same way for $\mathcal{F}_i$'s. Now, given a linear functional G that maps a function $f \in \mathcal{F}_i$ to a real number $\langle G, f \rangle$, define the operator norm $\|G\|_{\text{op}} = \sup_{f \in \mathcal{F}_i} \frac{\langle G, f \rangle}{\|f\|_{\text{nl}}}$.

A.2 Fréchet Derivatives

By definition of Fréchet derivatives, we have

$$\langle D_{h_i}[L(h)], \eta \rangle = \lim_{\epsilon \to 0} \frac{L(h_1, \ldots, h_i + \epsilon\eta, \ldots, h_{H+1}) - L(h)}{\epsilon},$$

where $\eta \in \mathcal{F}_i$ is the direction of perturbation and $\langle D_{h_i}[L(h)], \eta \rangle$ is the directional derivative along that direction η. From the definition of $L(h)$,

$$\begin{aligned}
&L(h_1, \ldots, h_i + \epsilon\eta, \ldots, h_{H+1}) \\
=&\frac{1}{2}\mathbb{E}\left[\|h_{H+1:i+1} \circ (h_i + \epsilon\eta) \circ h_{i-1:1}(X) - h^*(X)\|_2^2\right] + C \\
=&\frac{1}{2}\mathbb{E}\left[\|h_{H+1:i+1}(h_{i:1}(X) + \epsilon\eta(h_{i-1:1}(X))) - h^*(X)\|_2^2\right] + C \\
=&\frac{1}{2}\mathbb{E}\left[\left\|h_{H+1:1}(X) + \epsilon J[h_{H+1:i+1}](h_{i:1}(X))\eta(h_{i-1:1}(X)) + O(\epsilon^2) - h^*(X)\right\|_2^2\right] + C \\
=&L(h) + \epsilon\mathbb{E}\left[(h_{H+1:1}(X) - h^*(X))^T J[h_{H+1:i+1}](h_{i:1}(X))\eta(h_{i-1:1}(X))\right] + O(\epsilon^2).
\end{aligned}$$

Therefore,

$$\langle D_{h_i}[L(h)], \eta \rangle = \mathbb{E}\left[(h_{H+1:1}(X) - h^*(X))^T J[h_{H+1:i+1}](h_{i:1}(X))\eta(h_{i-1:1}(X))\right]. \tag{A.1}$$

Equation A.1 will be used in the proof of Theorems 4.1 and 4.2.

A.3 Proof of Theorem 4.1

From Equation A.1, consider $D_{h_1}[L(h)]$. For any $\eta \in \mathcal{F}_1$,

$$\langle D_{h_1}[L(h)], \eta \rangle = \mathbb{E}\left[(h_{H+1:1}(X) - h^*(X))^T J[h_{H+1:2}](h_1(X))\eta(X)\right].$$

Let $A(X) = J[h_{H+1:2}](h_1(X))$. Since $A(X)$ has full row rank by assumption, $A(X)A(X)^T$ is invertible. Then define a particular direction

$$\tilde{\eta}(X) = A(X)^T(A(X)A(X)^T)^{-1}(h_{H+1:1}(X) - h^*(X)),$$

so that

$$\left\langle D_{h_1}[L(h)], \tilde{\eta} \right\rangle = \mathbb{E}\left[\|h_{H+1:1}(X) - h^*(X)\|_2^2 \right].$$

It remains to check if $\tilde{\eta} \in \mathcal{F}_1$. It is easily checked that $\tilde{\eta}(0) = 0$ because $h_{H+1:1}(0) - h^*(0) = 0$. Since $J[h_{H+1:2}]$ is differentiable by assumption and $h_1 \in \mathcal{F}_1$, $A(X)$ is differentiable and $A(X)^T$, $(A(X)A(X)^T)^{-1}$ are differentiable functions. Also, $h_{H+1:1} - h^* \in \mathcal{F}$, so we can conclude that $\tilde{\eta}$ is differentiable.

Moreover, if we decompose $A(X)$ with SVD, $A(X) = U\Sigma V^T$, Σ is of the form $\Sigma = \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}$ and

$$\begin{aligned} A(X)^T(A(X)A(X)^T)^{-1} &= V\Sigma^T U^T (U\Sigma V^T V \Sigma^T U^T)^{-1} = V\Sigma^T U^T (U \Sigma_1^2 U^T)^{-1} \\ &= V\Sigma^T U^T U \Sigma_1^{-2} U^T = V \begin{bmatrix} \Sigma_1^{-1} \\ 0 \end{bmatrix} U^T, \end{aligned}$$

from which we can see that

$$\left\| A(X)^T(A(X)A(X)^T)^{-1} \right\|_{\mathrm{op}} = \sigma_{\max}(A(X)^T(A(X)A(X)^T)^{-1}) \le 1/\epsilon,$$

by our assumption. Note that, for any $X \in \mathbb{R}^{d_x}$,

$$\begin{aligned} \|\tilde{\eta}(X)\|_2 &= \left\| A(X)^T(A(X)A(X)^T)^{-1}(h_{H+1:1}(X) - h^*(X)) \right\|_2 \\ &\le \left\| A(X)^T(A(X)A(X)^T)^{-1} \right\|_{\mathrm{op}} \|h_{H+1:1}(X) - h^*(X)\|_2 \\ &\le \left\| A(X)^T(A(X)A(X)^T)^{-1} \right\|_{\mathrm{op}} \|h_{H+1:1} - h^*\|_{\mathrm{nl}} \|X\|_2 . \end{aligned}$$

Since this holds for any X, we have

$$\|\tilde{\eta}\|_{\mathrm{nl}} \le \left\| A(X)^T(A(X)A(X)^T)^{-1} \right\|_{\mathrm{op}} \|h_{H+1:1} - h^*\|_{\mathrm{nl}} \le \frac{\|h_{H+1:1} - h^*\|_{\mathrm{nl}}}{\epsilon},$$

which ensures that $\tilde{\eta} \in \mathcal{F}_1$. Finally,

$$\left\| D_{h_1}[L(h)] \right\|_{\mathrm{op}} \ge \frac{\left\langle D_{h_1}[L(h)], \tilde{\eta} \right\rangle}{\|\tilde{\eta}\|_{\mathrm{nl}}} \ge \frac{\epsilon \mathbb{E}\left[\|h_{H+1:1}(X) - h^*(X)\|_2^2 \right]}{\|h_{H+1:1} - h^*\|_{\mathrm{nl}}} = \frac{\epsilon(L(h) - L^*)}{\|h_{H+1:1} - h^*\|_{\mathrm{nl}}},$$

which yields

$$\left\| D_{h_1}[L(h)] \right\|_{\mathrm{op}} \|h_{H+1:1} - h^*\|_{\mathrm{nl}} \ge \epsilon(L(h) - L^*).$$

From this we can see that if we have a critical point of $L(h)$, then $\left\| D_{h_1}[L(h)] \right\|_{\mathrm{op}} = 0$ implies $L(h) = L^*$, which means that the critical point is a global minimum of $L(h)$.

A.4 Proof of Theorem 4.2

Recall that by assumption we have $j \in \{1, \dots, H+1\}$ such that $d_x = d_{j-1}$ and $d_y \le d_j$. Consider $D_{h_j}[L(h)]$, then for any $\eta \in \mathcal{F}_j$,

$$\left\langle D_{h_j}[L(h)], \eta \right\rangle = \mathbb{E}\left[(h_{H+1:1}(X) - h^*(X))^T J[h_{H+1:j+1}](h_{j:1}(X)) \eta(h_{j-1:1}(X)) \right].$$

As done in the previous theorem, for any $w \in \mathbb{R}^{d_{j-1}}$, let $A(w) = J[h_{H+1:j+1}](h_j(w))$. Since $A(w)$ has full row rank by assumption, $A(w)A(w)^T$ is invertible. Then define

$$\tilde{\eta}(w) = A(w)^T(A(w)A(w)^T)^{-1}(h_{H+1:1} - h^*) \circ h_{j-1:1}^{-1}(w),$$

so that

$$\left\langle D_{h_j}[L(h)], \tilde{\eta} \right\rangle = \mathbb{E}\left[\|h_{H+1:1}(X) - h^*(X)\|_2^2 \right].$$

We need to check if $\tilde{\eta} \in \mathcal{F}_j$. It is easily checked that $\tilde{\eta}(0) = 0$. Since $J[h_{H+1:j+1}]$ is differentiable by assumption and $h_j \in \mathcal{F}_j$, $A(w)$ is differentiable, and so are $A(w)^T$ and $(A(w)A(w)^T)^{-1}$. The inverse function of a differentiable and invertible function is also differentiable, so $(h_{H+1:1} - h^*) \circ h_{j-1:1}^{-1}$ is differentiable. Hence, we can conclude that $\tilde{\eta}$ is differentiable.

As seen in the previous section,

$$\left\| A(w)^T (A(w)A(w)^T)^{-1} \right\|_{\text{op}} = \sigma_{\max}(A(w)^T (A(w)A(w)^T)^{-1}) \leq 1/\epsilon_2.$$

By the assumption that $h_{j-1:1}$ is invertible and $\left\| h_{j-1:1}(u) \right\|_2 \geq \epsilon_1 \left\| u \right\|_2$,

$$\|v\|_2 \geq \epsilon_1 \left\| h_{j-1:1}^{-1}(v) \right\|_2,$$

for all $v \in \mathbb{R}^{d_{j-1}}$. From this, we can see that $\left\| h_{j-1:1}^{-1} \right\|_{\text{nl}} \leq 1/\epsilon_1$. For any $w \in \mathbb{R}^{d_{j-1}}$,

$$\begin{aligned}
\|\tilde{\eta}(w)\|_2 &= \left\| A(w)^T (A(w)A(w)^T)^{-1} (h_{H+1:1} - h^*) \circ h_{j-1:1}^{-1}(w) \right\|_2 \\
&\leq \left\| A(w)^T (A(w)A(w)^T)^{-1} \right\|_{\text{op}} \left\| (h_{H+1:1} - h^*) \circ h_{j-1:1}^{-1}(w) \right\|_2 \\
&\leq \left\| A(w)^T (A(w)A(w)^T)^{-1} \right\|_{\text{op}} \|h_{H+1:1} - h^*\|_{\text{nl}} \left\| h_{j-1:1}^{-1}(w) \right\|_2 \\
&\leq \left\| A(w)^T (A(w)A(w)^T)^{-1} \right\|_{\text{op}} \|h_{H+1:1} - h^*\|_{\text{nl}} \left\| h_{j-1:1}^{-1} \right\|_{\text{nl}} \|w\|_2 .
\end{aligned}$$

From this, we have

$$\|\tilde{\eta}\|_{\text{nl}} \leq \left\| A(X)^T (A(X)A(X)^T)^{-1} \right\|_{\text{op}} \|h_{H+1:1} - h^*\|_{\text{nl}} \left\| h_{j-1:1}^{-1} \right\|_{\text{nl}} \leq \frac{\|h_{H+1:1} - h^*\|_{\text{nl}}}{\epsilon_1 \epsilon_2}.$$

Finally,

$$\left\| D_{h_j}[L(h)] \right\|_{\text{op}} \geq \frac{\left\langle D_{h_j}[L(h)], \tilde{\eta} \right\rangle}{\|\tilde{\eta}\|_{\text{nl}}} \geq \frac{\epsilon_1 \epsilon_2 \mathbb{E}\left[\|h_{H+1:1}(X) - h^*(X)\|_2^2 \right]}{\|h_{H+1:1} - h^*\|_{\text{nl}}} = \frac{\epsilon_1 \epsilon_2 (L(h) - L^*)}{\|h_{H+1:1} - h^*\|_{\text{nl}}},$$

which yields

$$\left\| D_{h_j}[L(h)] \right\|_{\text{op}} \|h_{H+1:1} - h^*\|_{\text{nl}} \geq \epsilon_1 \epsilon_2 (L(h) - L^*).$$

A.5 Proof of Proposition 4.3

(Theorem 4.1) By assumption, we have $d_1 \geq d_y$. Set $h_1(x) = (h^*(x), 0, \ldots, 0)$ where for every $x \in \mathbb{R}^{d_x}$, the first d_y components of $h_1(x)$ are identical to $h^*(x)$, and all other components are zero. For the rest of h_i's, define $h_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ to be

$$h_i(w) = \begin{cases} (w_1, \ldots, w_{d_i}), & \text{if } d_i \leq d_{i-1}, \\ (w_1, \ldots, w_{d_{i-1}}, 0, \ldots, 0), & \text{if } d_i > d_{i-1}, \end{cases} \tag{A.2}$$

for all $w \in \mathbb{R}^{d_{i-1}}$. Since $d_i \geq d_y$ for all i, we can check that $h_{H+1} \circ \cdots \circ h_1 = h^*$, and $h_i \in \mathcal{F}_i$ for all i. Moreover, for all $z \in \mathbb{R}^{d_1}$, $J[h_{H+1:2}](z)$ is all 0 except 1's in diagonal entries, so $\sigma_{\min}(J[h_{H+1:2}](z)) \geq 1$ and $h_{H+1:2}(z)$ is twice-differentiable.

(Theorem 4.2) It is given that we have $j \in \{1, \dots, H+1\}$ such that $d_x = d_{j-1}$ and $d_y \leq d_j$. Set $h_j(x) = (h^*(x), 0, \dots, 0)$, where the first d_y components are $h^*(x)$ and the rest are zero. All the rest of h_i are set as in Equation A.2. Then, it can be easily checked that $h_i \in \mathcal{F}_i$ for all i and all the conditions of the theorem are satisfied.

B Deferred Lemmas

Lemma B.1. *For any $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times l}$ where $m \geq n$,*

$$\|AB\|_{\mathrm{F}}^2 \geq \sigma_{\min}^2(A) \, \|B\|_{\mathrm{F}}^2 .$$

Proof. Since $A^T A \succeq \sigma_{\min}^2(A) I$, $B^T A^T A B \succeq \sigma_{\min}^2(A) B^T B$. Then

$$\|AB\|_{\mathrm{F}}^2 = \mathrm{tr}(B^T A^T A B) \geq \sigma_{\min}^2(A) \, \mathrm{tr}(B^T B) = \sigma_{\min}^2(A) \, \|B\|_{\mathrm{F}}^2 .$$

□

Lemma B.2. *For any $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times l}$ where $n \leq l$,*

$$\|AB\|_{\mathrm{F}}^2 \geq \sigma_{\min}^2(B) \, \|A\|_{\mathrm{F}}^2 .$$

Proof. Since $B B^T \succeq \sigma_{\min}^2(B) I$, $A B B^T A^T \succeq \sigma_{\min}^2(B) A A^T$. Then

$$\|AB\|_{\mathrm{F}}^2 = \mathrm{tr}(B^T A^T A B) = \mathrm{tr}(A B B^T A^T) \geq \sigma_{\min}^2(B) \, \mathrm{tr}(A A^T) = \sigma_{\min}^2(B) \, \|A\|_{\mathrm{F}}^2 .$$

□

Lemma B.3. *For $d_x < d_y$, any critical point that is not a local minimum is a saddle point.*

Proof. For this lemma, we separate the proof into two cases: $W_H \cdots W_1 \neq 0$ and $W_H \cdots W_1 = 0$. The crux of the proof is to show that any critical point cannot be a local maximum. Then, any critical point is either a local minimum or a saddle point, so the conclusion of this lemma follows.

In case of $W_H \cdots W_1 \neq 0$, we use some of the results in Kawaguchi [8] and examine the Hessian of $L(W)$ with respect to $\mathrm{vec}(W_{H+1}^T)$, where $\mathrm{vec}(A)$ denotes vectorization of matrix A. Let $D_{\mathrm{vec}(W_{H+1}^T)} L(W)$ be the partial derivative of $L(W)$ with respect to $\mathrm{vec}(W_{H+1}^T)$ in numerator layout. It was shown by Kawaguchi [8, Lemma 4.3] that the Hessian matrix

$$\begin{aligned} \mathcal{H}(W) = D_{\mathrm{vec}(W_{H+1}^T)} \left(D_{\mathrm{vec}(W_{H+1}^T)} L(W) \right)^T &= \left(I \otimes (W_H \cdots W_1 X)(W_H \cdots W_1 X)^T \right) \\ &= \left(I \otimes W_H \cdots W_1 X X^T W_1^T \cdots W_H^T \right), \end{aligned}$$

where $\otimes$ denotes the Kronecker product of two matrices. Notice that $\mathcal{H}(W)$ is positive semidefinite. Since XX^T is full rank, whenever $W_H \cdots W_1 \neq 0$ there exists a strictly positive eigenvalue in $\mathcal{H}(W)$, which means that there exists an increasing direction. So W cannot be a local maximum.

The case where $W_H \cdots W_1 = 0$ requires a bit more careful treatment. For any arbitrary $\epsilon > 0$, we describe a procedure that perturbs the matrices $W_1, \dots, W_{H+1}$ by perturbations sampled from Frobenius norm balls of radius ϵ centered at 0, which we will denote as $\mathcal{B}_i(\epsilon)$, $i = 1, \dots, H+1$. Let $\mathcal{U}(\mathcal{B}_i(\epsilon))$ be the uniform distribution over the ball $\mathcal{B}_i(\epsilon)$. The algorithm goes as the following:

1. For $i \in \{1, \dots, H+1\}$

   1.1. Sample $\Delta_i \sim \mathcal{U}(\mathcal{B}_i(\epsilon))$, and define $V_i = W_i + \Delta_i$.

1.2. If $W_{H+1} \cdots W_{i+1} V_i \cdots V_1 \neq 0$, stop and return $i^* = i$.

First, recall that the set of rank-deficient matrices have Lebesgue measure zero, so for any sample $\Delta_i \sim \mathcal{U}(\mathcal{B}_i(\epsilon))$, $V_i = W_i + \Delta_i$ has full rank with probability 1. If we proceed the for loop until $i = H+1$, we have a full-rank $V_{H+1} \cdots V_1$ with probability 1, which means that the algorithm must return $i^* \in \{1, \dots, H+1\}$ with probability 1. Notice that before and after the i^*-th iteration, we have

$$\begin{aligned}
&W_{H+1} \cdots W_{i^*} V_{i^*-1} \cdots V_1 = 0,\\
&W_{H+1} \cdots W_{i^*+1} V_{i^*} \cdots V_1 = W_{H+1} \cdots W_{i^*+1} (W_{i^*} + \Delta_{i^*}) V_{i^*-1} \cdots V_1 \neq 0,
\end{aligned}$$

meaning that $W_{H+1} \cdots W_{i^*+1} \Delta_{i^*} V_{i^*-1} \cdots V_1 \neq 0$. Define $\hat{\Delta} = W_{H+1} \cdots W_{i^*+1} \Delta_{i^*} V_{i^*-1} \cdots V_1$, and then notice that

$$W_{H+1} \cdots W_{i^*+1} (W_{i^*} - \Delta_{i^*}) V_{i^*-1} \cdots V_1 = -\hat{\Delta}.$$

Now, define two points

$$\begin{aligned}
U^{(1)} &= (V_1, \dots, V_{i^*-1}, W_{i^*} + \Delta_{i^*}, W_{i^*+1}, \dots, W_{H+1}),\\
U^{(2)} &= (V_1, \dots, V_{i^*-1}, W_{i^*} - \Delta_{i^*}, W_{i^*+1}, \dots, W_{H+1}),
\end{aligned}$$

and notice that they are all in the neighborhood of W, that is, the Cartesian product of ϵ-radius balls centered at $W_1, \dots, W_{H+1}$. Moreover, we have

$$\begin{aligned}
L(W) &= \frac{1}{2} \|0 \cdot X - Y\|_{\mathrm{F}}^2 = \frac{1}{2} \|Y\|_{\mathrm{F}}^2,\\
L(U^{(1)}) &= \frac{1}{2} \left\|\hat{\Delta} X - Y\right\|_{\mathrm{F}}^2 = \frac{1}{2} \|Y\|_{\mathrm{F}}^2 + \frac{1}{2} \left\|\hat{\Delta} X\right\|_{\mathrm{F}}^2 - \left\langle \hat{\Delta} X, Y \right\rangle,\\
L(U^{(2)}) &= \frac{1}{2} \left\|-\hat{\Delta} X - Y\right\|_{\mathrm{F}}^2 = \frac{1}{2} \|Y\|_{\mathrm{F}}^2 + \frac{1}{2} \left\|\hat{\Delta} X\right\|_{\mathrm{F}}^2 + \left\langle \hat{\Delta} X, Y \right\rangle,
\end{aligned}$$

from which we can see that at least one of $L(W) < L(U^{(1)})$ or $L(W) < L(U^{(2)})$ must hold. This shows that for any $\epsilon > 0$, there is a point U in ϵ-neighborhood of W with a strictly greater function value $L(U)$. This proves that W cannot be a local maximum.

□

C Deferred Proofs

C.1 Proof of Proposition 3.1

In case of $k < \min\{d_x, d_y\}$, we can decompose the loss function in the following way:

$$\begin{aligned}
\|RX - Y\|_{\mathrm{F}}^2 &= \left\|RX - YX^T(XX^T)^{-1}X + YX^T(XX^T)^{-1}X - Y\right\|_{\mathrm{F}}^2\\
&= \left\|RX - YX^T(XX^T)^{-1}X\right\|_{\mathrm{F}}^2 + \left\|YX^T(XX^T)^{-1}X - Y\right\|_{\mathrm{F}}^2\\
&\quad + 2\,\mathrm{tr}((YX^T(XX^T)^{-1}X - Y)(RX - YX^T(XX^T)^{-1}X)^T).
\end{aligned}$$

Let us take a close look into the last term in the RHS. Note that $YX^T(XX^T)^{-1}X$ is the orthogonal projection of Y onto $\mathrm{row}(X)$, so each row of $YX^T(XX^T)^{-1}X - Y$ must be in $\mathrm{null}(X)$. Also,

$$(RX - YX^T(XX^T)^{-1}X)^T = X^T(R^T - (XX^T)^{-1}XY^T).$$

It is X^T right-multiplied with some matrix, so its columns must lie in $\mathrm{col}(X^T) = \mathrm{row}(X)$. By the fact that $\mathrm{null}(X)^\perp = \mathrm{row}(X)$,

$$(YX^T(XX^T)^{-1}X - Y)(RX - YX^T(XX^T)^{-1}X)^T = 0,$$

thus

$$L_0(R) = \frac{1}{2}\left\|RX - YX^T(XX^T)^{-1}X\right\|_{\mathrm{F}}^2 + \frac{1}{2}\left\|YX^T(XX^T)^{-1}X - Y\right\|_{\mathrm{F}}^2$$

holds.

Now, Equation 3.1 becomes a problem of minimizing $\left\|RX - YX^T(XX^T)^{-1}X\right\|_{\mathrm{F}}^2$ subject to the rank constraint $\mathrm{rank}(R) \leq k$. The optimal solution for this is obtained when RX is the k-rank approximation of $YX^T(XX^T)^{-1}X$. Then, k-rank approximation of $YX^T(XX^T)^{-1}X$ can be expressed as $\hat{U}\hat{U}^TYX^T(XX^T)^{-1}X$, where $\hat{U}$ is unique due to our assumption that all singular values are distinct. Therefore,

$$R^* = \hat{U}\hat{U}^TYX^T(XX^T)^{-1}$$

is the unique global minimum solution of Equation 3.1 when $k < \min\{d_x, d_y\}$.

C.2 Proof of Lemma 3.2

$$\begin{aligned}
&L(W_1, \ldots, W_{i-1}, W_i + \Delta_i, W_{i+1}, \ldots, W_{H+1}) \\
=&\frac{1}{2}\left\|W_{H+1}\cdots W_{i+1}(W_i + \Delta_i)W_{i-1}\cdots W_1X - Y\right\|_{\mathrm{F}}^2 \\
=&\frac{1}{2}\left\|W_{H+1}\cdots W_1X - Y + W_{H+1}\cdots W_{i+1}\Delta_iW_{i-1}\cdots W_1X\right\|_{\mathrm{F}}^2 \\
=&L(W) + \mathrm{tr}((W_{H+1}\cdots W_{i+1}\Delta_iW_{i-1}\cdots W_1X)^T(W_{H+1}\cdots W_1X - Y)) + O(\|\Delta_i\|_{\mathrm{F}}^2) \\
=&L(W) + \mathrm{tr}(W_{i+1}^T\cdots W_{H+1}^T(W_{H+1}\cdots W_1X - Y)X^TW_1^T\cdots W_{i-1}^T\Delta_i^T) + O(\|\Delta_i\|_{\mathrm{F}}^2).
\end{aligned}$$

From this, we can conclude that

$$\frac{\partial L}{\partial W_i} = W_{i+1}^T\cdots W_{H+1}^T(W_{H+1}\cdots W_1X - Y)X^TW_1^T\cdots W_{i-1}^T.$$