Global optimality conditions for deep neural networks

Chulhee Yun
Mass. Institute of Technology
Cambridge, MA 02139
chulheey@mit.edu

Suvrit Sra
Mass. Institute of Technology
Cambridge, MA 02139
suvrit@mit.edu

Ali Jadbabaie
Mass. Institute of Technology
Cambridge, MA 02139
jadbabai@mit.edu

Abstract

We study the error landscape of deep linear neural networks with the squared error loss. Minimizing the loss of a deep linear neural network is a nonconvex problem, and despite recent progress, our understanding of this loss surface is still incomplete. For deep linear networks, we present necessary and sufficient conditions for a critical point of the risk function to be a global minimum. Our conditions provide an efficiently checkable test for global optimality, which is remarkable because such tests are typically intractable in nonconvex optimization.

1 Introduction

Since the advent of AlexNet [9], 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. Despite these successes, a concrete theoretical understanding of why deep neural networks work well in practice remains 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 3-node neural network to global optimality is NP-Hard in the worst case, so there is little hope that global optimization of neural networks is tractable. However, recent papers such as [4] provide theoretical and empirical evidence that the local minima of the loss surfaces could be close to global minima.

Towards obtaining a more precise characterization of the loss surfaces, a valuable 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. Its optimization is nonconvex, and only recently theoretical results on this problem have started emerging. 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 linear networks and proved that any local minimum is also global while any other critical point is a saddle point. Subsequently, Lu and Kawaguchi [10] provided a simpler proof that any local minimum is also global, under 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 showed every critical point is a global minimum in a near-identity region; subsequently, 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 networks, in settings either similar to or more general than existing work. We summarize our main contributions below.

- We provide necessary and sufficient conditions for a critical point of the empirical risk to be a global minimum (in comparison, Kawaguchi [8] only proves that every critical point of the risk is

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

either a global minimum or a saddle). Specifically, Theorem 2.1 shows that if the hidden layers are wide enough, then a critical point of the risk function is a global minimum *if and only if* the product of all parameter matrices is full-rank. This concise condition provides an efficient test on whether a given critical point is a global minimum or a saddle; it is worth noting such tests are intractable for general nonconvex optimization [11]. In Theorem 2.2, we consider the case where some hidden layers have smaller width than both the input and output layers, and again provide necessary and sufficient conditions for global optimality.

- Under the same assumption as [5] on the data distribution, namely, a linear model with Gaussian noise, we can modify Theorem 2.1 to handle the population risk. As a corollary, we not only recover Theorem 2.2 in [5], but also extend it to a strictly *larger* set, while removing their assumption that the true underlying linear model has a positive determinant.

If we were to also take a similar approach as [2], then building on our linear results (Theorem 2.1), we can also obtain sufficient conditions for global optimality in deep nonlinear networks, albeit in a function space view. We omit elaboration of these results due to the lack of space.

2 Global optimality conditions for deep linear neural networks

2.1 Problem formulation and notation

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 of the network can be written as the product of weight matrices $W_{H+1}, \dots, W_1$ and data matrix X: $W_{H+1}W_H \cdots W_1 X$. We consider minimizing the summation of squared error loss over all data points (i.e. empirical risk),

$$\text{minimize} \quad L(W) := \tfrac{1}{2} \left\| W_{H+1}W_H \cdots W_1 X - Y \right\|_{\mathrm{F}}^2 , \tag{1}$$

where W is a shorthand notation for the tuple $(W_1, \dots, W_{H+1})$.

Assumptions. We assume that $d_y \le d_x \le m$, and that XX^T and YX^T have full ranks. These assumptions are common when we consider supervised learning problems with deep neural networks (e.g. Kawaguchi [8]). 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.

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

Let us denote $k := \min_{i \in \{0,\dots,H+1\}} d_i$, and define $p \in \mathrm{argmin}_{i \in \{0,\dots,H+1\}} d_i$. That is, p is any layer with the smallest width, and $k = d_p$ is the width of that layer. 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 describe two sets, one for the case $k = d_y$ and the other for $k < d_y$, inside which every critical point of $L(W)$ is a global minimum, and outside of which every critical point is a saddle point. Previous works as Kawaguchi [8] and Lu and Kawaguchi [10] showed that any critical point is a global minimum or a saddle. 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 saddles.

Theorem 2.1. *If $k = d_y$, define the following set*

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

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 $k < d_y$, define the following set*

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

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 globally 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. It is worth noting that such checkable conditions for optimality are not typical in nonconvex problems.

In Hardt and Ma [5], the authors consider minimizing population risk of linear residual networks:

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

where $d_x = d_1 = \cdots = d_H = d_y = d$. They assume that x is drawn from a zero-mean distribution with a fixed 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 to derive a population risk counterpart, and in fact notice that the result proved in Hardt and Ma [5] is a corollary of this modification 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 $\det(R) > 0$ which was required by Hardt and Ma [5]. 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.*

3 Proofs of theorems

3.1 Unique solutions of the relaxed problem

Consider a "relaxed" version of the original problem (1):

$$\begin{array}{ll} \text{minimize}_{R \in \mathbb{R}^{d_y \times d_x}} & L_0(R) := \frac{1}{2}\|RX - Y\|_{\text{F}}^2 \\ \text{subject to} & \operatorname{rank}(R) \le k. \end{array} \tag{2}$$

We can observe that if $k = d_y$, (2) is an unconstrained optimization problem, and that the unique globally optimal solution is

$$R^* = YX^T(XX^T)^{-1}. \tag{3}$$

In case of $k < d_y$, the problem becomes non-convex because of the rank constraint, but the unique solution can be computed easily and it is the k-rank approximation of (3), namely

$$R^* = \hat{U}\hat{U}^T YX^T(XX^T)^{-1}. \tag{4}$$

Now note that for any $W = (W_1, \dots, W_{H+1})$, the product $W_{H+1} \cdots W_1$ has rank at most k and $L_0(W_{H+1} \cdots W_1) = L(W)$. Therefore, L_0 is a relaxation of L and $L_0(R^*) \le \inf_W L(W)$. This means that if there exists W such that $R^* = W_{H+1} \cdots W_1$, then $L(W) = L_0(R^*)$ and thus W is a global minimum of the function L.

3.2 Proof of Theorem 2.1

We now prove Theorem 2.1, which handles the case when $k = d_y$. First note that whenever the product $W_{H+1} \cdots W_1$ is not full rank ($W \in \mathcal{V}_1^c$), the product cannot be the global optimum R^* of $L_0(R)$, thus W cannot be a global optimum of (1). By [8, Theorem 2.3.(iii)], any critical point in $\mathcal{V}_1^c$ is a saddle point of $L(W)$.

We now focus on $W \in \mathcal{V}_1$. Observe that by simple matrix calculus,

$$\frac{\partial L}{\partial W_i} = W_{i+1}^T \cdots W_{H+1}^T (W_{H+1} \cdots W_1 X - Y) X^T W_1^T \cdots W_{i-1}^T, \tag{5}$$

and define $E := (W_{H+1} \cdots W_1 X - Y)X^T \in \mathbb{R}^{d_y \times d_x}$ for simplicity. For a fixed $\epsilon > 0$, define a set $\mathcal{W}_\epsilon$ as $\mathcal{W}_\epsilon := \{(W_1, \dots, W_{H+1}) : \sigma_{\min}(W_{H+1} \cdots W_2) \geq \epsilon\}$. Then, for any tuple $W \in \mathcal{W}_\epsilon$,

$$\left\| \frac{\partial L}{\partial W_1} \right\|_{\mathrm{F}}^2 \geq \sigma_{\min}^2(W_{H+1} \cdots W_2) \left\| E \right\|_{\mathrm{F}}^2 \geq \epsilon^2 \left\| E \right\|_{\mathrm{F}}^2,$$

so any critical point ($\frac{\partial L}{\partial W_1} = 0$) in $\mathcal{W}_\epsilon$ must satisfy $E := (W_{H+1} \cdots W_1 X - Y)X^T = 0$ and thus $W_{H+1} \cdots W_1 = YX^T(XX^T)^{-1} = R^*$. Therefore, any critical point in $\mathcal{W}_\epsilon$ is a global minimum.

Now, note that if $W \in \mathcal{V}_1$, $W_{H+1} \cdots W_2$ must also have full rank, so W is a member of $\mathcal{W}_\epsilon$ for any ϵ satisfying $0 < \epsilon \leq \sigma_{\min}(W_{H+1} \cdots W_2)$. So any critical point in $\mathcal{V}_1$ is a critical point in $\mathcal{W}_\epsilon$ for some $\epsilon > 0$. This finishes the proof.

3.3 Proof of Theorem 2.2

First, define the set $\mathcal{V}_1$ as in the previous theorem, i.e. the set of tuples W with $\mathrm{rank}(W_{H+1} \cdots W_1) = k$. For now, we focus on $W \in \mathcal{V}_1$. In order to simplify notation, define

$$A_i := W_{i+1}^T \cdots W_{H+1}^T \in \mathbb{R}^{d_i \times d_y},\ B_i := W_1^T \cdots W_{i-1}^T \in \mathbb{R}^{d_x \times d_{i-1}}, i = 1, \dots, H+1,$$

so that $\frac{\partial L}{\partial W_i} = A_i E B_i$. Notice that A_{H+1} and B_1 are identity matrices.

Since the full product $W_{H+1} \cdots W_1$ has rank k, any partial products A_i and B_i must have $\mathrm{rank}(A_i) \geq k$ and $\mathrm{rank}(B_i) \geq k$, for all i. Then, consider $A_p \in \mathbb{R}^{k \times d_y}$ and $B_{p+1} \in \mathbb{R}^{d_x \times k}$, and note that $\mathrm{rank}(A_p) \leq k$ and $\mathrm{rank}(B_{p+1}) \leq k$ by the dimension of matrices. Also, $\mathrm{rank}(A_i) \leq \mathrm{rank}(A_{i+1})$ and $\mathrm{rank}(B_{i+1}) \leq \mathrm{rank}(B_i)$ holds for all $i \in \{1, \dots, H\}$, so we have

$$\mathrm{rank}(A_i) = k \text{ for } i = 1, \dots, p, \text{ and } \mathrm{rank}(B_i) = k \text{ for } i = p+1, \dots, H+1.$$

More importantly, this implies that $\mathrm{row}(A_1) = \cdots = \mathrm{row}(A_p)$ and $\mathrm{col}(B_{H+1}) = \cdots = \mathrm{col}(B_{p+1})$. Using this observation, we can now state a proposition showing necessary and sufficient conditions for a tuple $W \in \mathcal{V}_1$ to be a critical point of $L(W)$.

Proposition 3.1. *A tuple $W \in \mathcal{V}_1$ is a critical point of L if and only if $A_p E = 0$ and $EB_{p+1} = 0$.*

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

$$\mathrm{col}(E) = \mathrm{col}(EB_1) \subset \mathrm{row}(A_1)^\perp = \mathrm{row}(A_p)^\perp \implies A_p E = 0,$$
$$\mathrm{row}(E) = \mathrm{row}(A_{H+1}E) \subset \mathrm{col}(B_{H+1})^\perp = \mathrm{col}(B_{p+1})^\perp \implies EB_{p+1} = 0.$$

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

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 $YX^T(XX^T)^{-1}X = U\Sigma V^T$, $\hat{U} \in \mathbb{R}^{d_y \times k}$ is defined to be a matrix consisting of the first k columns of U.

Proposition 3.2. *A critical point $W \in \mathcal{V}_1$ of $L(W)$ is a global minimum point if and only if* $\mathrm{col}(W_{H+1} \cdots W_{p+1}) = \mathrm{row}(A_p) = \mathrm{col}(\hat{U})$.

Proof. If W is a critical point, $A_p E = 0$ by Proposition 3.1. Note that $W_{H+1} \cdots W_1 = A_p^T B_{p+1}^T$, so

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

Thus, B_{p+1} is determined uniquely as $B_{p+1}^T = (A_p A_p^T)^{-1} A_p YX^T(XX^T)^{-1}$. From (4), W is a global minimum solution if and only if

$$W_{H+1} \cdots W_1 = A_p^T B_{p+1}^T = A_p^T (A_p A_p^T)^{-1} A_p YX^T(XX^T)^{-1} = \hat{U}\hat{U}^T YX^T(XX^T)^{-1}.$$

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 these matrices are projecting onto subspaces that are identical: $\mathrm{row}(A_p) = \mathrm{col}(\hat{U})$. □

From Proposition 3.2, 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.

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 *Artificial Intelligence and Statistics*, pages 192–204, 2015.

[5] M. Hardt and T. Ma. Identity matters in deep learning. In *International Conference on Learning Representations*, 2017.

[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] 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.

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

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