Depth Creates No Bad Local Minima

Anonymous Author(s)
Affiliation
Address
email

Abstract

In deep learning, *depth*, as well as *nonlinearity*, create non-convex loss surfaces. Then, does depth alone create bad local minima? In this paper, we prove that without nonlinearity, depth alone does not create bad local minima, although it induces non-convex loss surface. Using this insight, we greatly simplify a recently proposed proof to show that all of the local minima of feedforward deep linear neural networks are global minima. Our theoretical results generalize previous results with fewer assumptions, and this analysis provides a method to show similar results beyond square loss in deep linear models.

1 Introduction

Deep learning has recently had a profound impact on the machine learning, computer vision, and artificial intelligence communities. In addition to its practical successes, previous studies have revealed several reasons why deep learning has been successful from the viewpoint of its *model classes*. An (over-)simplified explanation is the harmony of its great expressivity and *big data*: because of its great expressivity, deep learning can have less *bias*, while a large training dataset leads to less *variance*. The great expressivity can be seen from an aspect of representation learning as well: whereas traditional machine learning makes use of features designed by human users or experts as a type of prior, deep learning tries to learn features from the data as well. More accurately, a key aspect of the model classes in deep learning is the *generalization* property; despite its great expressivity, deep learning model classes can maintain great generalization properties (Livni et al., 2014; Mhaskar et al., 2016; Poggio et al., 2016). This would distinguish deep learning from other possibly too flexible methods, such as shallow neural networks with too many hidden units, and traditional kernel methods with a too powerful kernel. Therefore, the practical success of deep learning seems to be supported by the great quality of its model classes.

However, having a great model class is not so useful if we cannot find a good model in the model class via training. Training a deep model is typically framed as non-convex optimization. Because of its non-convexity and high dimensionality, it has been unclear whether we can *efficiently* train a deep model. Note that the difficulty comes from the combination of non-convexity and high dimensionality in weight parameters. If we can reformulate the training problem into several decoupled training problems, with each having a small number of weight parameters, we can effectively train a model via non-convex optimization as theoretically shown in Bayesian optimization and global optimization literatures (Kawaguchi et al., 2015; Wang et al., 2016; Kawaguchi et al., 2016). As a result of non-convexity and high-dimensionality, it was shown that training a general neural network model is NP-hard (Blum & Rivest, 1992). However, such a hardness-result in a worst case analysis would not tightly capture what is going on in practice, as we seem to be able to efficiently train deep models in practice.

To understand its practical success beyond worst case analysis, theoretical and practical investigations on the training of deep models have recently become an active research area (Saxe et al., 2014; Dauphin et al., 2014; Choromanska et al., 2015; Haeffele & Vidal, 2015; Shamir, 2016;

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

Kawaguchi, 2016; Swirszcz et al., 2016; Arora et al., 2016; Freeman & Bruna, 2016; Soudry & Hoffer, 2017).

An important property of a deep model is that the non-convexity comes from *depth*, as well as *nonlinearity*: indeed, depth by itself creates highly non-convex optimization problems. One way to see a property of the non-convexity induced by depth is the non-uniqueness owing to *weight–space symmetries* (Krkova & Kainen, 1994): the model represents the same function mapping from the input to the output with different distinct settings in the weight space. Accordingly, there are many distinct globally optimal points and many distinct points with the same loss values due to weight–space symmetries, which would result in a non-convex epigraph (i.e., non-convex function) as well as non-convex sublevel sets (i.e., non-quasiconvex function). Thus, it has been unclear whether *depth* by itself can create a difficult non-convex loss surface. The recent work (Kawaguchi, 2016) indirectly showed, as a consequence of its main theoretical results, that depth does not create bad local minima of deep linear model with Frobenius norm although it creates potentially bad saddle points.

In this paper, we directly prove that all local minima of deep linear model corresponds to local minima of shallow model. Building upon this new theoretical insight, we propose a simpler proof for one of the main results in the recent work (Kawaguchi, 2016); all of the local minima of feedforward deep linear neural networks with Frobenius norm are global minima. The power of this proof can go beyond Frobenius norm: as long as the loss function satisfies Theorem 3.2, all local minima of deep linear model corresponds to local minimum of shallow model.

2 Main Result

To examine the effect of depth alone, we consider the following optimization problem of feedforward deep linear neural networks with the square error loss:

$$\underset{W}{\text{minimize}} \quad L(W) = \frac{1}{2}\|W_H W_{H-1} \cdots W_1 X - Y\|_F^2, \tag{1}$$

where $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ is the weight matrix, $X \in \mathbb{R}^{d_0 \times m}$ is the input training data, and $Y \in \mathbb{R}^{d_H \times m}$ is the target training data. Let $p = \arg\min_{0 \leq i \leq H} d_i$ be the index corresponding to the smallest width. Note that for any W, we have $\text{rank}(W_H W_{H-1} \cdots W_1) \leq d_p$. To analyze optimization problem (1), we also consider the following optimization problem with a "shallow" linear model, which is equivalent to problem (1) in terms of the global minimum value:

$$\underset{R}{\text{minimize}} \quad F(R) = \|RX - Y\|_F^2 \quad \text{s.t.} \quad \text{rank}(R) \leq d_p, \tag{2}$$

where $R \in \mathbb{R}^{d_H \times d_0}$. Note that problem (2) is non-convex, unless $d_p = \min(d_H, d_0)$, whereas problem (1) is non-convex, even when $d_p \geq \min(d_H, d_0)$ with $H > 1$. In other words, deep parameterization creates a non-convex loss surface even without nonlinearity.

Though we only consider the Frobenius loss here, the proof holds for general cases. As long as the loss function satisfies Theorem 3.2, all local minima of deep linear model corresponds to local minimum of shallow model.

Our first main result states that even though deep parameterization creates a non-convex loss surface, it does not create new bad local minima. In other words, every local minimum in problem (1) corresponds to a local minimum in problem (2).

Theorem 2.1. (Depth creates no new bad local minima) *Assume that X and Y have full row rank. If $\bar{W} = \{\bar{W}_1, \ldots, \bar{W}_H\}$ is a local minimum of problem (1), then $\bar{R} = \bar{W}_H \bar{W}_{H-1} \cdots \bar{W}_1$ achieves the value of a local minimum of problem (2).*

Therefore, we can deduce the property of the local minima in problem (1) from those in problem (2). Accordingly, we first analyze the local minima in problem (2), and obtain the following statement.

Theorem 2.2. (No bad local minima for rank restricted shallow model) *If X has full row rank, all local minima of optimization problem* (2) *are global minima.*

By combining Theorems 2.1 and 2.2, we conclude that every local minimum is a global minimum for feedforward deep linear networks with a square error loss.

Theorem 2.3. (No bad local minima for deep linear neural networks) *If X and Y have full row rank, then all local minima of problem (1) are global minima.*

Theorem 2.3 generalizes one of the main results in (Kawaguchi, 2016) with fewer assumptions. Following the theoretical work with a random matrix theory (Dauphin et al., 2014; Choromanska et al., 2015), the recent work (Kawaguchi, 2016) showed that under some strong assumptions, all of the local minima are global minima for a class of nonlinear deep networks. Furthermore, the recent work (Kawaguchi, 2016) proved the following properties for a class of general deep linear networks with arbitrary depth and width: 1) the objective function is non-convex and non-concave; 2) all of the local minima are global minima; 3) every other critical point is a saddle point; and 4) there is no saddle point with the Hessian having no negative eigenvalue for shallow networks with one hidden layer, whereas such saddle points exist for deeper networks. Theorem 2.3 generalizes the second statement with fewer assumptions; the previous papers (Baldi, 1989; Kawaguchi, 2016) assume that the data matrix $YX^T(XX^T)^{-1}XY^T$ has distinct eigenvalues, whereas we do not assume that.

3 Proof

In this section, we provide the proofs of Theorems 2.1, 2.2, and 2.3.

3.1 Proof of Theorem 2.1

In order to deduce the proof of Theorem 2.1, we need some fundamental facts in linear algebra. The next two lemmas recall some basic facts of perturbation theory for singular value decomposition (SVD).

Let M and $\bar{M}$ be two $m \times n$ ($m \geq n$) matrices with SVDs

$$B = U\Sigma V^T = (U_1, U_2)\begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \\ & \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$$

$$\bar{B} = \bar{U}\bar{\Sigma}\bar{V}^T = (\bar{U}_1, \bar{U}_2)\begin{pmatrix} \bar{\Sigma}_1 & \\ & \bar{\Sigma}_2 \\ & \end{pmatrix}\begin{pmatrix} \bar{V}_1^T \\ \bar{V}_2^T \end{pmatrix},$$

where $\Sigma_1 = \mathrm{diag}(\sigma_1, \cdots, \sigma_k)$, $\Sigma_2 = \mathrm{diag}(\sigma_{k+1}, \cdots, \sigma_n)$, $\Sigma_1 = \mathrm{diag}(\bar{\sigma}_1, \cdots, \bar{\sigma}_k)$, $\Sigma_2 = \mathrm{diag}(\bar{\sigma}_{k+1}, \cdots, \bar{\sigma}_n)$, U, V, $\bar{U}$ and $\bar{V}$ are orthogonal matrices.

Lemma 3.1. Continuity of Singular Value *The singular value σ_i of a matrix is a continuous map of entries of the matrix.*

Lemma 3.2. (Wedin, 1972) Continuity of Singular Space

If

$$\rho := \min\left\{ \min_{1\leq i\leq k, 1\leq j\leq n-k} |\sigma_i - \bar{\sigma}_{k+j}|, \min_{1\leq i\leq k} \sigma_i \right\} > 0,$$

then:

$$\sqrt{\|\sin(U_1, \bar{U}_1)\|_F^2 + \|\sin(V_1, \bar{V}_1)\|_F^2} \leq \frac{\sqrt{\|\left(\bar{M} - M\right)V_1\|_F^2 + \|\left(\bar{M}^* - M^*\right)U_1\|_F^2}}{\rho}.$$

For a fixed matrix B, we say "matrix A is a perturbation of matrix B" if $\|A - B\|_\infty$ is $o(1)$, which means that the difference between A and B is much smaller than any non-zero number in matrix B.

Lemma 3.2 implies that any SVD for a perturbed matrix is a perturbation of some SVD for the original matrix under full rank condition. More formally:

Lemma 3.3. *Let $\bar{M}$ be a full-rank matrix with singular value decomposition $\bar{M} = \bar{U}\bar{\Sigma}\bar{V}^T$. M is a perturbation of $\bar{M}$. Then, there exists one SVD of M, $M = U\Sigma V^T$, such that U is a perturbation of $\bar{U}$, Σ is a perturbation of $\bar{\Sigma}$ and V is a perturbation of $\bar{V}$.(Notice that SVD of a matrix may not be unique due to rotation of the eigen-space corresponding to the same eigenvalue)*

Proof: With the small perturbation of matrix $\bar{M}$, Lemma 3.1 shows that the singular values does not change much. Thus, if $\|\bar{M} - M\|_\infty$ is small enough, $|\sigma_i - \bar{\sigma}_i|$ is also small for all i. Remember that all singular values of $\bar{M}$ are positive. By letting Σ_1 contain only the singular value σ_i (which may be multiple, and hence U_1 and V_1 are the singular spaces corresponding to the singular value σ_i), we have $\rho > 0$ in Lemma 3.2, thus Lemma 3.2 implies that the singular space of the perturbed matrix corresponding to singular value σ_i in the initial matrix does not change much. The statement of the lemma follows by combining this result for the different singular values together (i.e., consider each index i for different σ_i in the above argument). □

We say that W satisfies the rank condition, if $\operatorname{rank}(W_H \cdots W_1) = d_p$. Any perturbation of the products of matrices is the product of the perturbed matrices, when the original matrix satisfies the rank constraint. More formally:

Theorem 3.1. *Let $\bar{R} = \bar{W}_H\bar{W}_{H-1}\cdots\bar{W}_1$ with $\operatorname{rank}(\bar{R}) = d_p$. Then, for any R, such that R is a perturbation of $\bar{R}$ and $\operatorname{rank}(R) \le d_p$, there exists $\{W_1, W_2, \ldots, W_H\}$, such that W_i is perturbation of $\bar{W}_i$ for all $i \in \{1, \ldots, H\}$ and $R = W_H W_{H-1}\cdots W_1$.*

We will prove the theorem by induction. When $H = 2$, we can easily show that the perturbation of the product of two matrices is the product of one matrix and the perturbation of the other matrix. When $H = k >= 3$, we let M be the product of two specific matrices, and by induction the perturbation of the product (R) is the product of a perturbation of M and perturbations of the other $H - 2$ matrix. And a perturbation of M is also the product of perturbations of those two specific matrices, which proves the statement when $H = k$.

Proof: The case with $H = 1$ holds by setting $W_1 = R$. We prove the lemma with $H \ge 2$ by induction.

We first consider the base case where $H = 2$ with $\bar{R} = \bar{W}_2\bar{W}_1$.

Let $\bar{R} = \bar{U}\bar{\Sigma}\bar{V}^T$ be the SVD of $\bar{R}$. It follows Lemma 3.3 that there exists an SVD of R, $R = U\Sigma V^T$, such that U is a perturbation of $\bar{U}$, Σ is a perturbation of $Sig\bar{m}a$ and V is a perturbation of $\bar{V}$. Because $\operatorname{rank}(\bar{R}) = d_p$, with a small perturbation, the positive singular values remain strictly positive, whereby, $\operatorname{rank}(R) \ge d_p$. Together with the assumption $\operatorname{rank}(R) \le d_p$, we have $\operatorname{rank}(R) = d_p$. Let $\bar{S}_2 = \bar{U}^T\bar{W}_2$ and $\bar{S}_1 = \bar{W}_1\bar{V}$. Note that $\bar{U}\bar{\Sigma}\bar{V}^T = \bar{R} = \bar{W}_2\bar{W}_1$. Hence, $\bar{S}_2\bar{S}_1 = \bar{\Sigma}$ is a diagonal matrix. Remember Σ is a perturbation of $\bar{\Sigma}$, thus there is an S_2, which is a perturbation of $\bar{S}_2$ (each row of S_2 is a scale of the corresponding row of $\bar{S}_2$), such that $S_2\bar{S}_1 = \Sigma$. Let $W_2 = US_2$ and $W_1 = \bar{S}_1 V$. Then, W_1 is a perturbation of $\bar{W}_1$, W_2 is a perturbation of $\bar{W}_2$, and $W_1W_2 = R$, which proves the case when $H = 2$.

For the inductive step, given that the lemma holds for the case with $H = k \ge 2$, let us consider the case when $H = k + 1 \ge 3$ with $\bar{R} = \bar{W}_{k+1}\bar{W}_k\cdots\bar{W}_1$. Let $\mathcal{I}$ be an index set defined as $\mathcal{I} = \{p, p-1\}$ if $p \ge 2$, $\mathcal{I} = \{p+2, p+1\}$ if $p = 0$ or $p = 1$. We denote the i-th element of a set $\mathcal{I}$ by $\mathcal{I}_i$. Then, $\bar{M} = \bar{W}_{\mathcal{I}_2}\bar{W}_{\mathcal{I}_1}$ exists as $k + 1 \ge 3$. Note that $\bar{R}$ can be written as a product of k matrices with $\bar{M}$ (for example, $\bar{R} = \bar{W}_H \cdots \bar{W}_{I_1+1}\bar{M}\bar{W}_{I_2-1}\cdots\bar{W}_1$). Thus, from the inductive hypothesis, for any R, such that R is a perturbation of $\bar{R}$ and $\operatorname{rank}(R) \le d_p$, there exists a set of desired k matrices M and W_i for $i \in \{1, \ldots, k+1\} \setminus \mathcal{I}$, such that W_i is perturbation of $\bar{W}_i$ for all $i \in \{1, \ldots, k+1\} \setminus \mathcal{I}$, M is perturbation of $\bar{M}$, and the product is equal to R. Meanwhile, because $\bar{M}$ is either a d_p by d_{p-2} matrix or a d_{p+2} by d_p matrix, we have $\operatorname{rank}(\bar{M}) \le d_p$ and $\operatorname{rank}(M) \le d_p$, and it follows $\operatorname{rank}(\bar{R}) = d_p$ that $\operatorname{rank}(\bar{M}) = d_p$. Thus, by setting $\bar{R} \leftarrow \bar{M}$ and $R \leftarrow M$ (note that d_p in $\bar{R} = \bar{W}_{k+1}\bar{W}_k\cdots\bar{W}_1$ is equal to d_p in $\bar{M} = \bar{W}_{\mathcal{I}_2}\bar{W}_{\mathcal{I}_1}$), we can apply the proof for the case of $H = 2$ to conclude: there exists $\{W_{\mathcal{I}_2}, W_{\mathcal{I}_1}\}$, such that W_i is perturbation of $\bar{W}_i$ for all $i \in \mathcal{I}$, and $M = W_{\mathcal{I}_2}W_{\mathcal{I}_1}$. Combined with the above statement from the inductive hypothesis, this implies the lemma with $H = k + 1$, whereby we finish the proof by induction. □

The next two theorems show that, for any local minimum of $L(\cdot)$, there is another local minimum of $L(\cdot)$, whose function value is the same as the original and it satisfies the rank constraint.

Theorem 3.2. *Let $W = \{W_1, \cdots, W_H\}$ be a local minimum of problem (1) and $R \triangleq W_H W_{H-1} \cdots W_1$. If W_i is not of full rank, then there exists a $\bar{W}_i$, such that $\bar{W}_i$ is of full rank, $\bar{W}_i$ is a perturbation of W_i, $\bar{W} = \{W_1, \cdots, W_{i-1}, \bar{W}_i, W_{i+1}, \cdots, W_H\}$ is a local minimum of problem (1), and $L(W) = L(\bar{W})$.*

The idea of the proof is that if we just change one weight W_i and keep all other weights, it becomes a convex least square problem. Then we are able to perturb W_i to maintain the objective value as well as the perturbation is full rank.

Proof of Theorem 3.2 For notational convenience, let $A = W_{i-1} \cdots W_1 X$ and $B = W_{i+1} \cdots W_H$, and let $L_i(W_i) = \frac{1}{2}\|B^T W_i A - Y\|_F^2$. Because W is a local minimum of L, W_i is a local minimum of L_i. Let $A = U_1^T D_1 V_1$ and $B = U_2^T D_2 V_2$ are the SVDs of A and B, respectively, where D_i is a diagonal matrix with the first s_i terms being strictly positive, $i = 1, 2$. Minimizing L_i over W_i is a least square problem, and the normal equation is

$$BB^T W_i AA^T = BYA^T, \tag{3}$$

hence

$$\begin{aligned} W_i &\in (BB^T)^+ BYA^T (AA^T)^+ + \left\{M | BB^T M AA^T = 0\right\} \\ &= U_2 D_2^+ V_2^T Y V_1 D_1^+ U_1^T + \left\{U_2 K U_1^T | K_{1:s_2, 1:s_1} = 0\right\}, \end{aligned}$$

where $(\cdot)^+$ is a Moore–Penrose pseudo-inverse and K is a matrix with suitable dimension with the entries in the top left $s_2 \times s_1$ rectangular being 0.

Since $V_2^T Y V_1$ is of full rank,

$$\operatorname{rank}(D_2^+ V_2^T Y V_1 D_1^+) \geq \max\{0, s_2 + s_1 - \max\{d_i, d_{i-1}\}\}$$

Thus, we can choose a proper K (which contains $d_i + d_{i-1} - s_2 - s_1$ 1s at proper positions with all other terms being 0s) such that $D_2^+ V_2^T Y V_1 D_1^+ + K$ is of full rank, whereby $U_2 \left(D_2^+ V_2^T Y V_1 D_1^+ + K\right) U_1^T$ is of full rank. Therefore, there is a full rank $\hat{W}_i$ that satisfies the normal equation (3).

Let $\bar{W}_i(\mu) = W_i + \mu\left(\hat{W}_i - W_i\right)$. Then, $\bar{W}_i(\mu)$ also satisfies the normal equation, and $L(\bar{W}(\mu)) = L_i(\bar{W}_i(\mu)) = L_i(W_i) = L(W)$, for any $\mu > 0$.

Note that W is a local minimum of $L(W)$. Thus, there exists a $\delta > 0$, such that for any W^0 satisfying $\|W^0 - W\|_\infty \leq \delta$, we have $L(W^0) \geq L(W)$. It follows from $\hat{W}_i$ being full rank that there exists a small enough μ, such that $\bar{W}_i(\mu)$ is full rank and $\|\bar{W}_i(\mu) - W_i\|_\infty$ is arbitrarily small (in particular, $\|\bar{W}_i(\mu) - W_i\|_\infty \leq \frac{\delta}{2}$), because the non-full-rank matrices are discrete on the line of $\bar{W}_i(\mu)$ with parameter $\mu > 0$ by considering the determine of $W_i^T(\mu) W_i(\mu)$ or $W_i(\mu) W_i^T(\mu)$ as a polynomial of λ. Therefore, for any W^0, such that $\|W^0 - \bar{W}(\mu)\|_\infty \leq \frac{\delta}{2}$, we have

$$\|W^0 - W\|_\infty \leq \|W^0 - \bar{W}(\mu)\|_\infty + \|\bar{W}_i(\mu) - W_i\|_\infty \leq \delta\ ,$$

whereby

$$L(W^0) \geq L(W) = L(\bar{W}(\mu))\ .$$

This shows that $\bar{W}(\mu) = \left\{W_1, \cdots, W_{i-1}, \bar{W}_i(\mu), W_{i+1}, \cdots, W_H\right\}$ is also a local minimum of problem (1) for some small enough μ. □

Lemma 3.4. *Let $R = AB$ for two given matrices $A \in R^{d_1 \times d_2}$ and $B \in R^{d_2 \times d_3}$. If $d_1 \leq d_2$, $d_1 \leq d_3$ and $rank(A) = d_1$, then any perturbation of R is the product of A and perturbation of B.*

Proof: Let $A = UDV^T$ be the SVD of A, then, $R = UDV^T B$. Let $\bar{R}$ be a perturbation of R and let $\bar{B} = B + VD^+U^T(\bar{R} - R)$. Then, $\bar{B}$ is a perturbation of B and $A\bar{B} = \bar{R}$ by noticing $DD^+ = I$, as A has full row rank. □

Theorem 3.3. *If $\bar{W} = \{\bar{W}_1, \cdots, \bar{W}_H\}$ is a local minimum with $\bar{W}_i$ being full rank, then, there exists $\hat{W} = \left\{\hat{W}_1, \cdots, \hat{W}_H\right\}$, such that $\hat{W}_i$ is a perturbation of $\bar{W}_i$ for all $i \in \{1, \ldots, H\}$, $\hat{W}$ is a local minimum, $L(\hat{W}) = L(\bar{W})$, and $\operatorname{rank}(\hat{W}_H \hat{W}_{H-1} \cdots \hat{W}_1) = d_p$.*

In the proof of Theorem 3.3, we will use Theorem 3.2 and Lemma 3.4 to show that we can perturb $\bar{W}_{p-1},\ \bar{W}_{p-2}\ldots,\bar{W}_1$ in sequence to make sure the perturbed weight is still the optimal solution and $\mathrm{rank}(\hat{W}_p\hat{W}_{p-1}) = d_p$. Similar strategy can make sure $\mathrm{rank}(\hat{W}_H\hat{W}_{H-1}\cdots\hat{W}_{p+1}) = d_p$, which then proves the whole theorem.

Proof of Theorem 3.3 : If $p \neq 1$, consider

$$L_1(T) := \|\bar{W}_H\cdots\bar{W}_{p+1}T\bar{W}_{p-2}\cdots\bar{W}_1X - Y\|_F^2.$$

Then, it follows from Lemma 3.4 and $\bar{W}$ is a local minimum of $L(W)$ that $\bar{T}$ is a local minimum of L_1, where $\bar{T} = \bar{W}_p\bar{W}_{p-1}$. It follows from Theorem 3.2 that there exists $\hat{T}$, such that $\hat{T}$ is close enough to $\bar{T}$, $\hat{T}$ is a local minimum of $L_1(T)$, $L_1(\hat{T}) = L_1(\bar{T})$, and $\mathrm{rank}(\hat{T}) = d_p$. Note $\hat{T}$ is a perturbation of $\bar{T}$, whereby, from Lemma 3.4, there exists $\hat{W}_p$, $\hat{W}_{p-1}$, which are perturbations of $\bar{W}_p$ and $\bar{W}_{p-1}$, respectively, such that $\hat{W}_p\hat{W}_{p-1} = \hat{T}$. Thus, $\hat{W}^0 = \left(\bar{W}_H,\cdots,\bar{W}_{p+1},\hat{W}_p,\hat{W}_{p-1},\bar{W}_{p-2}\cdots,\bar{W}_1\right)$ is a local minimum of $L(W)$, $L(\hat{W}) = L(\bar{W})$ and $\mathrm{rank}(\hat{W}_p\hat{W}_{p-1}) = d_p$.

By that analogy, we can find $\hat{W}_p\cdots\hat{W}_1$, such that $\hat{W}^1 = \left(\bar{W}_H,\cdots,\bar{W}_{p+1},\hat{W}_p,\hat{W}_{p-1},\cdots,\hat{W}_1\right)$ is a local minimum of $L(W)$, $\hat{W}_i$ is a perturbation of $\bar{W}_i$ for $i = 1,\cdots,p$, $L(\hat{W}^1) = L(\bar{W})$ and $\mathrm{rank}(\hat{W}_p\hat{W}_{p-1}\cdots\hat{W}_1) = d_p$.

Similarly, we can find $\hat{W}_H\cdots\hat{W}_{p+1}$, such that $\hat{W}^2 = \left(\hat{W}_H,\cdots,\hat{W}_{p+1},\hat{W}_p,\hat{W}_{p-1},\cdots,\hat{W}_1\right)$ is a local minimum of $L(W)$, $\hat{W}_i$ is a perturbation of $\bar{W}_i$ for $i = p+1,\cdots,H$, $L(\hat{W}^2) = L(\hat{W}^1) = L(\bar{W})$ and $\mathrm{rank}(\hat{W}_H\hat{W}_{H-1}\cdots\hat{W}_{p+1}) = d_p$.

Noticing that

$$\mathrm{rank}(\hat{W}_H\cdots\hat{W}_1) \geq \mathrm{rank}(\hat{W}_H\hat{W}_{H-1}\cdots\hat{W}_{p+1}) + \mathrm{rank}(\hat{W}_p\hat{W}_{p-1}\cdots\hat{W}_1) - d_p = d_p$$

and $\mathrm{rank}(\hat{W}_H\cdots\hat{W}_1) \leq \min_{i=0,\ldots,H} d_i = d_p$, we have $\mathrm{rank}(\hat{W}_H\cdots\hat{W}_1) = d_p$, which completes the proof. □

Proof of Theorem 2.1: It follows from Theorem 3.2 and Theorem 3.3 that there exists another local minimum $\hat{W} = \hat{W} = \left\{\hat{W}_1,\cdots,\hat{W}_H\right\}$, such that $L(\hat{W}) = L(\bar{W})$ and $\mathrm{rank}(\hat{W}_H\hat{W}_{H-1}\cdots\hat{W}_1) = d_p$. Remember that $\hat{R} = \hat{W}_H\hat{W}_{H-1}\cdots\hat{W}_1$. It then follows from Theorem 3.1 that for any R, such that R is a perturbation of $\hat{R}$ and $\mathrm{rank}(R) \leq d_p$, we have $R = W_HW_{H-1}\cdots W_1$, where W_i is a perturbation of $\hat{W}_i$. Therefore, by noticing $\hat{W}$ is a local minimum of (1), we have

$$F(R) = L(W) \geq L(\hat{W}) = F(\hat{R})\ ,$$

which shows that $\hat{R}$ is a local minimum of (2). □

In the proof of Theorem 2.2, we at first show that we just need to consider the case where X is an identity matrix and Y is a diagonal matrix by noticing rotation is invariant under Frobenius norm. Then we show that the local minimum must be a block diagonal and symmetric matrix, and each block term is a projection matrix on the space corresponding to the same eigenvalue of the diagonal matrix Y. Finally, we show that those projection matrices must be onto the eigenspace of Y corresponding to the as large as possible eigenvalues, which then shows that the local minimum shares the same function value.

3.2 Proof of Theorem 2.2

Let $X = U_1\Sigma_1V_1^T$ be the SVD decomposition of X, where Σ_1 is a diagonal matrix with full row rank. Then,

$$\begin{aligned} F(R) &= \|RU_1\Sigma_1V_1^T - Y\|_F^2 = \|RU_1\Sigma_1 - YV_1\|_F^2 \\ &= \|(RU_1)(\Sigma_1)_{1:d_1,1:d_1} - (YV_1)_{1:d_2,1:d_1}\|_F^2 + \text{Const}, \end{aligned}$$

where Const is a constant in R and $(\cdot)_{t_1:t_2,t_3:t_4}$ is a submatrix of $(\cdot)$, which contains the t_1 to t_2 row and t_3 to t_4 column of $(\cdot)$. If R is a local minimum of (2), then $S = RU_1$ is a local minimum of

$$\begin{array}{rl} \min_S & G(S) = \|S\hat{\Sigma}_1 - \hat{Y}\|_F^2 \\ s.t. & \text{rank}(S) \leq k \ , \end{array} \tag{4}$$

where $\hat{\Sigma}_1 := (\Sigma_1)_{1:d_1,1:d_1}$, $\hat{Y} := (YV_1)_{1:d_2,1:d_1}$ and the difference of objective function values of (2) and (4) is a constant. Let $\hat{Y} := U_2\Sigma_2V_2^T$ be the SVD of $\hat{Y}$, then

$$G(S) = \|S\hat{\Sigma}_1 - U_2\Sigma_2V_2^T\|_F^2 = \|U_2^T S\hat{\Sigma}_1 V_2 - \hat{\Sigma}_2\|_F^2 \ ,$$

and if S is a local minimum of $G(S)$, we have $T := U_2^T S\hat{\Sigma}_1 V_2$ is a local minimum of

$$\begin{array}{rl} \min_T & H(T) = \|T - \Sigma_2\|_F^2 \\ s.t. & \text{rank}(T) \leq k \ , \end{array} \tag{5}$$

and the objective function values of (4) and (5) are the same at corresponding points. Let Σ_2 have r distinct positive diagonal terms $\lambda_1 > \cdots > \lambda_r \geq 0$ with multiplicities $m_1, \cdots, m_r$. Let T^* be a local minimum of (5), and

$$T^* = U^*\Sigma^*V^{*T} = [U_S^* U_N^*] \begin{bmatrix} \Sigma_S^* & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_S^{*T} \\ V_N^{*T} \end{bmatrix}$$

be the SVD of T, where Σ_S^* are positive singular values. Let $P_L := U_S^* \left(U_S^{*T} U_S^*\right)^{-1} U_S^{*T}$ and $P_R := V_S^*(V_S^{*T}V_S^*)^{-1}V_S^{*T}$ be the projection matrix to the space spanned by U_S^* and V_S^*, respectively. Note that $\{T|P_LT = T\} \subseteq \{T|rank(T) \leq k\}$, thus, T^* is also a local minimum of

$$\begin{array}{l} \min\|T - \Sigma_2\|_F^2 \\ s.t. P_L T = T, \end{array} \tag{6}$$

which is a convex problem, and it can be shown by the first order optimality condition that the only local minimum of (6) is $T^* = P_L\Sigma_2$. Similarly, we have $T^* = \Sigma_2 P_R$. Then, $D := \Sigma_2\Sigma_2^T$ is a diagonal matrix, with r distinct non-zero diagonal terms $\lambda_1^2 > \cdots > \lambda_r^2 > 0$ with multiplicities $m_1, \cdots, m_r$. Therefore,

$$\begin{aligned} P_L D P_L &= P_L\Sigma_2\Sigma_2^T P_L^T = T^*T^{*T} = \Sigma_2 P_R P_R^T \Sigma_2^T \\ &= \Sigma_2 P_R \Sigma_2^T = \Sigma_2 T^{*T} = \Sigma_2\Sigma_2^T P_L^T = DP_L. \end{aligned}$$

Note that the left hand is a symmetric matrix, thus, DP_L is also a symmetric matrix. Meanwhile, P_L is a symmetric matrix, whereby P_L is a r-block diagonal matrix with each block corresponding to the same diagonal terms of D. Therefore, $T^* = P_L\Sigma_2$ is also a r-block diagonal matrix.

Let

$$T^* = \begin{bmatrix} T_1^* & & & \\ & \ddots & & \\ & & T_r^* & \\ & & & 0 \end{bmatrix},$$

where T_i^* is a $m_i \times m_i$ matrix, then $T^*T^{*T} = \Sigma_2 T^{*T}$ implies $T_i^* T_i^{*T} = \lambda_i T_i^{*T}$. Thus, T_i^* is a symmetric matrix and $\frac{T_i^*}{\lambda_i}$ is a projection matrix. Let $rank(T_i^*) = d_{p_i}$, then, $\sum_{i=1}^r d_{p_i} \leq p$ and $tr(T_i^*) = \lambda_i d_{p_i}$, whereby,

$$\begin{aligned} H(T^*) &= \sum_{i=1}^r \|T_i^* - \lambda_i I_{m_i}\|_F^2 \\ &= \sum_{i=1}^r tr(T_i^2) - 2\lambda_i tr(T_i) + m_i\lambda_i^2 \\ &= \sum_{i=1}^r (m_i - d_{p_i})\,\lambda_i^2. \end{aligned}$$

Let j be the largest number that $\sum_{i=1}^{j} m_i < d_p$. Then, it is easy to find that the global minima of (6) satisfy $d_{p_i} = m_i$ for $i \le j$, $d_{p_{j+1}} = d_p - \sum_{i=1}^{j} m_i$ and $d_{p_i} = 0$ for $i > j + 1$ which gives all of the global minima.

Now, let us show that all local minima must be global minima. As local minima T^* is a block diagonal matrix, thus, we can assume without loss of generality that both Σ_2 and T^* are square matrices, because the all 0 rows and columns in Σ_2 and T do not change anything. Thus, it follows T_i^* is symmetric that T^* is a symmetric matrix. Remember that $\frac{T_i^*}{\lambda_i}$ is a projection matrix, thus the eigenvalues of T_i^* are either 0 or λ_i, whereby

$$T^* = \sum_{i=1}^{r} \sum_{j=1}^{d_{p_i}} \lambda_i u_{ij} u_{ij}^T,$$

where u_{ij} is the jth normalized orthogonal eigen-vector of T^* corresponding to eigenvalue λ_i.

It is easy to see that, at a local minimum, we have $\sum_{i=1}^{r} d_{p_i} = d_p$, otherwise, there is a descent direction by adding a rank 1 matrix to T^* corresponding to one positive eigenvalue. If there exists i_1, i_2, such that $i_1 < i_2$, $d_{p_{i_1}} < m_{i_1}$, and $d_{p_{i_2}} \ge 1$, then, there exists $\bar{u}_{i_1}$, such that $\bar{u}_{i_1} \perp u_{i_1 j}$ for $j = 1, \cdots d_{p_{i_1}}$. Let

$$\begin{aligned} T(\theta) :=& T^* - \lambda_{i_2} u_{i_2 1} u_{i_2 1}^T + \left(\lambda_{i_1} \sin^2\theta + \lambda_{i_2} \cos^2\theta\right) \\ & \left(u_{i_2 1}\cos\theta + \bar{u}_{i_1}\sin\theta\right)\left(u_{i_2 1}\cos\theta + \bar{u}_{i_1}\sin\theta\right)^T. \end{aligned}$$

Then, $\operatorname{rank}(T(\theta)) = \operatorname{rank}(T^*) = d_p$, $T(0) = T^*$ and

$$H(T(\theta)) = H(T^*) + \lambda_1^2 + \lambda_2^2 - \left(\lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta\right)^2.$$

It is easy to check that $H(T(\theta))$ is monotonically decreasing with θ, which gives a descent direction at T^*, contradicting with that T^* is a local minimum. Therefore, there is no such i_1 and i_2, which shows that T^* is a global minimum. □

3.3 Proof of Theorem 2.3

The statement follows from Theorem 2.1 and 2.2.

4 Conclusion

We have proven that, even though depth creates a non-convex loss surface, it does not create new bad local minima. Based on this new insight, we have successfully proposed a new simple proof for the fact that all of the local minima of feedforward deep linear neural networks are global minima as a corollary.

The benefits of this new results are not limited to the simplification of the previous proof. For example, our results apply to problems beyond square loss. Let us consider the shallow problem (S) $\text{minimize}\, L(R)$ s.t. $rank(R) \le d_p$, and and the deep parameterization counterpart (D) $\text{minimize}\, L(W_H W_{H-1} \cdots W_1)$. Our analysis shows that for any function L, as long as L satisfies Theorem 3.2, any local minimum of (D) corresponds to a local minimum of (S). This is not limited to when L is least square loss, and this is why we say depth creates no bad local minima.

In addition, our analysis can directly apply to matrix completion unlike previous results. Ge et al. (2016) show that local minima of the symmetric matrix completion problem are global with high probability. This should be able to extend to asymmetric case. Denote $f(W) := \sum_{i,j\in\Omega}(Y - W_2 W_1)_{i,j}$, then local minimum of $f(W)$ is global with high probability, where Ω is the observed entries. Then, our analysis here can directly show that the result can be extended for deep linear parameterization: for $h(W) := \sum_{i,j\in\Omega}(Y - W_H W_{H-1} \cdots W_1)_{i,j}$, any local minimum of $h(W)$ is global with high probability.

References

Arora, Raman, Basu, Amitabh, Mianjy, Poorya, and Mukherjee, Anirbit. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Baldi, Pierre. Linear learning: Landscapes and algorithms. In *Advances in neural information processing systems*, pp. 65–72, 1989.

Blum, Avrim L and Rivest, Ronald L. Training a 3-node neural network is np-complete. *Neural Networks*, 5(1):117–127, 1992.

Choromanska, Anna, Henaff, MIkael, Mathieu, Michael, Ben Arous, Gerard, and LeCun, Yann. The loss surfaces of multilayer networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 192–204, 2015.

Dauphin, Yann N, Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, Ganguli, Surya, and Bengio, Yoshua. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pp. 2933–2941, 2014.

Freeman, C Daniel and Bruna, Joan. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Ge, Rong, Lee, Jason D, and Ma, Tengyu. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

Haeffele, Benjamin D and Vidal, René. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

Kawaguchi, Kenji. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Kawaguchi, Kenji, Kaelbling, Leslie Pack, and Lozano-Pérez, Tomás. Bayesian optimization with exponential convergence. In *In Advances in Neural Information Processing (NIPS)*, 2015.

Kawaguchi, Kenji, Maruyama, Yu, and Zheng, Xiaoyu. Global continuous optimization with error bound and fast convergence. *Journal of Artificial Intelligence Research*, 56:153–195, 2016.

Krkova, Vera and Kainen, Paul C. Functionally equivalent feedforward neural networks. *Neural Computation*, 6(3):543–558, 1994.

Livni, Roi, Shalev-Shwartz, Shai, and Shamir, Ohad. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pp. 855–863, 2014.

Mhaskar, Hrushikesh, Liao, Qianli, and Poggio, Tomaso. Learning functions: When is deep better than shallow. *arXiv preprint arXiv:1603.00988*, 2016.

Poggio, Tomaso, Mhaskar, Hrushikesh, Rosasco, Lorenzo, Miranda, Brando, and Liao, Qianli. Why and when can deep–but not shallow–networks avoid the curse of dimensionality: a review. *arXiv preprint arXiv:1611.00740*, 2016.

Saxe, Andrew M, McClelland, James L, and Ganguli, Surya. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Shamir, Ohad. Distribution-specific hardness of learning neural networks. *arXiv preprint arXiv:1609.01037*, 2016.

Soudry, Daniel and Hoffer, Elad. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

Swirszcz, Grzegorz, Czarnecki, Wojciech Marian, and Pascanu, Razvan. Local minima in training of deep networks. *arXiv preprint arXiv:1611.06310*, 2016.

Wang, Zi, Zhou, Bolei, and Jegelka, Stefanie. Optimization as estimation with gaussian processes in bandit settings. In *International Conf. on Artificial and Statistics (AISTATS)*, 2016.

Wedin, Per-Åke. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.