Deep Learning without Poor Local Minima

Kenji Kawaguchi
Massachusetts Institute of Technology
kawaguch@mit.edu

Abstract

In this paper, we prove a conjecture published in 1989 and also partially address an open problem announced at the Conference on Learning Theory (COLT) 2015. With no unrealistic assumption, we first prove the following statements for the squared loss function of deep linear neural networks with any depth and any widths: 1) the function is non-convex and non-concave, 2) every local minimum is a global minimum, 3) every critical point that is not a global minimum is a saddle point, and 4) there exist "bad" saddle points (where the Hessian has no negative eigenvalue) for the deeper networks (with more than three layers), whereas there is no bad saddle point for the shallow networks (with three layers). Moreover, for deep nonlinear neural networks, we prove the same four statements via a reduction to a deep linear model under the independence assumption adopted from recent work. As a result, we present an instance, for which we can answer the following question: how difficult is it to directly train a deep model in theory? It is more difficult than the classical machine learning models (because of the non-convexity), but not too difficult (because of the nonexistence of poor local minima). Furthermore, the mathematically proven existence of bad saddle points for deeper models would suggest a possible open problem. We note that even though we have advanced the theoretical foundations of deep learning and non-convex optimization, there is still a gap between theory and practice.

1 Introduction

Deep learning has been a great practical success in many fields, including the fields of computer vision, machine learning, and artificial intelligence. In addition to its practical success, theoretical results have shown that deep learning is attractive in terms of its generalization properties (Livni *et al.*, 2014; Mhaskar *et al.*, 2016). That is, deep learning introduces good function classes that may have a low capacity in the VC sense while being able to represent target functions of interest well. However, deep learning requires us to deal with seemingly intractable optimization problems. Typically, training of a deep model is conducted via non-convex optimization. Because finding a global minimum of a *general* non-convex function is an NP-complete problem (Murty & Kabadi, 1987), a hope is that a function induced by a deep model has some structure that makes the non-convex optimization tractable. Unfortunately, it was shown in 1992 that training a very simple neural network is indeed NP-hard (Blum & Rivest, 1992). In the past, such theoretical concerns in optimization played a major role in shrinking the field of deep learning. That is, many researchers instead favored classical machining learning models (with or without a kernel approach) that require only convex optimization. While the recent great practical successes have revived the field, we do not yet know what makes optimization in deep learning tractable in theory.

In this paper, as a step toward establishing the optimization theory for deep learning, we prove a conjecture noted in (Goodfellow *et al.*, 2016) for deep *linear* networks, and also address an open problem announced in (Choromanska *et al.*, 2015b) for deep *nonlinear* networks. Moreover, for

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

both the conjecture and the open problem, we prove more general and tighter statements than those previously given (in the ways explained in each section).

2 Deep linear neural networks

Given the absence of a theoretical understanding of deep nonlinear neural networks, Goodfellow *et al.* (2016) noted that it is beneficial to theoretically analyze the loss functions of simpler models, i.e., deep *linear* neural networks. The function class of a linear multilayer neural network only contains functions that are linear with respect to inputs. However, their loss functions are non-convex in the weight parameters and thus nontrivial. Saxe *et al.* (2014) empirically showed that the optimization of deep *linear* models exhibits similar properties to those of the optimization of deep *nonlinear* models. Ultimately, for theoretical development, it is natural to start with linear models before working with nonlinear models (as noted in Baldi & Lu, 2012), and yet even for linear models, the understanding is scarce when the models become *deep*.

2.1 Model and notation

We begin by defining the notation. Let H be the number of hidden layers, and let (X, Y) be the training data set, with $Y \in \mathbb{R}^{d_y \times m}$ and $X \in \mathbb{R}^{d_x \times m}$, where m is the number of data points. Here, $d_y \geq 1$ and $d_x \geq 1$ are the number of components (or dimensions) of the outputs and inputs, respectively. Let $\Sigma = YX^T(XX^T)^{-1}XY^T$. We denote the model (weight) parameters by W, which consists of the entries of the parameter matrices corresponding to each layer: $W_{H+1} \in \mathbb{R}^{d_y \times d_H}, \ldots, W_k \in \mathbb{R}^{d_k \times d_{k-1}}, \ldots, W_1 \in \mathbb{R}^{d_1 \times d_x}$. Here, d_k represents the width of the k-th layer, where the 0-th layer is the input layer and the $(H+1)$-th layer is the output layer (i.e., $d_0 = d_x$ and $d_{H+1} = d_y$). Let I_{d_k} be the $d_k \times d_k$ identity matrix. Let $p = \min(d_H, \ldots, d_1)$ be the smallest width of a hidden layer. We denote the (j, i)-th entry of a matrix M by $M_{j,i}$. We also denote the j-th row vector of M by $M_{j,\cdot}$ and the i-th column vector of M by $M_{\cdot,i}$.

We can then write the output of a feedforward deep linear model, $\overline{Y}(W, X) \in \mathbb{R}^{d_y \times m}$, as

$$\overline{Y}(W, X) = W_{H+1}W_H W_{H-1} \cdots W_2 W_1 X.$$

We consider one of the most widely used loss functions, squared error loss:

$$\bar{\mathcal{L}}(W) = \frac{1}{2}\sum_{i=1}^{m} \|\overline{Y}(W, X)_{\cdot,i} - Y_{\cdot,i}\|_2^2 = \frac{1}{2}\|\overline{Y}(W, X) - Y\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm. Note that $\frac{2}{m}\bar{\mathcal{L}}(W)$ is the usual *mean* squared error, for which all of our results hold as well, since multiplying $\bar{\mathcal{L}}(W)$ by a constant in W results in an equivalent optimization problem.

2.2 Background

Recently, Goodfellow *et al.* (2016) remarked that when Baldi & Hornik (1989) proved Proposition 2.1 for shallow linear networks, they stated Conjecture 2.2 without proof for deep linear networks.

Proposition 2.1 (Baldi & Hornik, 1989: *shallow* linear network) *Assume that $H = 1$ (i.e., $\overline{Y}(W, X) = W_2W_1X$), assume that XX^T and XY^T are invertible, assume that Σ has d_y distinct eigenvalues, and assume that $p < d_x$, $p < d_y$ and $d_y = d_x$ (e.g., an autoencoder). Then, the loss function $\bar{\mathcal{L}}(W)$ has the following properties:*

- *(i) It is convex in each matrix W_1 (or W_2) when the other W_2 (or W_1) is fixed.*
- *(ii) Every local minimum is a global minimum.*

Conjecture 2.2 (Baldi & Hornik, 1989: *deep* linear network) *Assume the same set of conditions as in Proposition 2.1 except for $H = 1$. Then, the loss function $\bar{\mathcal{L}}(W)$ has the following properties:*

- *(i) For any $k \in \{1, \ldots, H+1\}$, it is convex in each matrix W_k when for all $k' \neq k$, $W_{k'}$ is fixed.*
- *(ii) Every local minimum is a global minimum.*

Baldi & Lu (2012) recently provided a proof for Conjecture 2.2 *(i)*, leaving the proof of Conjecture 2.2 *(ii)* for future work. They also noted that the case of $p \geq d_x = d_x$ is of interest, but requires further analysis, even for a shallow network with $H = 1$. An informal discussion of Conjecture 2.2 can be found in (Baldi, 1989). In Appendix D, we provide a more detailed discussion of this subject.

2.3 Results

We now state our main theoretical results for deep linear networks, which imply Conjecture 2.2 *(ii)* as well as obtain further information regarding the critical points with more generality.

Theorem 2.3 (Loss surface of *deep* linear networks) *Assume that XX^T and XY^T are of full rank with $d_y \leq d_x$ and Σ has d_y distinct eigenvalues. Then, for any depth $H \geq 1$ and for any layer widths and any input-output dimensions $d_y, d_H, d_{H-1}, \ldots, d_1, d_x \geq 1$ (the widths can arbitrarily differ from each other and from d_y and d_x), the loss function $\bar{\mathcal{L}}(W)$ has the following properties:*

- *(i) It is non-convex and non-concave.*
- *(ii) Every local minimum is a global minimum.*
- *(iii) Every critical point that is not a global minimum is a saddle point.*
- *(iv) If $\mathrm{rank}(W_H \cdots W_2) = p$, then the Hessian at any saddle point has at least one (strictly) negative eigenvalue.*[1]

Corollary 2.4 (Effect of deepness on the loss surface) *Assume the same set of conditions as in Theorem 2.3 and consider the loss function $\bar{\mathcal{L}}(W)$. For three-layer networks (i.e., $H = 1$), the Hessian at any saddle point has at least one (strictly) negative eigenvalue. In contrast, for networks deeper than three layers (i.e., $H \geq 2$), there exist saddle points at which the Hessian does not have any negative eigenvalue.*

The assumptions of having full rank and distinct eigenvalues in the training data matrices in Theorem 2.3 are realistic and practically easy to satisfy, as discussed in previous work (e.g., Baldi & Hornik, 1989). In contrast to related previous work (Baldi & Hornik, 1989; Baldi & Lu, 2012), we do not assume the invertibility of XY^T, $p < d_x$, $p < d_y$ nor $d_y = d_x$. In Theorem 2.3, $p \geq d_x$ is allowed, as well as many other relationships among the widths of the layers. Therefore, we successfully proved Conjecture 2.2 *(ii)* and a more general statement. Moreover, Theorem 2.3 *(iv)* and Corollary 2.4 provide additional information regarding the important properties of saddle points.

Theorem 2.3 presents an instance of a deep model that would be tractable to train with direct greedy optimization, such as gradient-based methods. If there are "poor" local minima with large loss values everywhere, we would have to search the entire space,[2] the volume of which increases exponentially with the number of variables. This is a major cause of NP-hardness for non-convex optimization. In contrast, if there are no poor local minima as Theorem 2.3 *(ii)* states, then saddle points are the main remaining concern in terms of tractability.[3] Because the Hessian of $\bar{\mathcal{L}}(W)$ is Lipschitz continuous, if the Hessian at a saddle point has a negative eigenvalue, it starts appearing as we approach the saddle point. Thus, Theorem 2.3 and Corollary 2.4 suggest that for 1-hidden layer networks, training can be done in polynomial time with a second order method or even with a modified stochastic gradient decent method, as discussed in (Ge *et al.*, 2015). For deeper networks, Corollary 2.4 states that there exist "bad" saddle points in the sense that the Hessian at the point has no negative eigenvalue. However, we know exactly when this can happen from Theorem 2.3 *(iv)* in our deep models. We leave the development of efficient methods to deal with such a bad saddle point in general deep models as an open problem.

3 Deep nonlinear neural networks

Now that we have obtained a comprehensive understanding of the loss surface of deep *linear* models, we discuss deep *nonlinear* models. For a practical deep nonlinear neural network, our theoretical results so far for the deep linear models can be interpreted as the following: depending on the

[1] If $H = 1$, to be succinct, we define $W_H \cdots W_2 = W_1 \cdots W_2 \triangleq I_{d_1}$, with a slight abuse of notation.

[2] Typically, we do this by assuming smoothness in the values of the loss function.

[3] Other problems such as the ill-conditioning can make it difficult to obtain a fast convergence rate.

nonlinear activation mechanism and architecture, training would not be arbitrarily difficult. While theoretical formalization of this intuition is left to future work, we address a recently proposed open problem for deep nonlinear networks in the rest of this section.

3.1 Model

We use the same notation as for the deep linear models, defined in the beginning of Section 2.1. The output of deep nonlinear neural network, $\hat{Y}(W, X) \in \mathbb{R}^{d_y \times m}$, is defined as

$$\hat{Y}(W, X) = q\sigma_{H+1}(W_{H+1}\sigma_H(W_H\sigma_{H-1}(W_{H-1} \cdots \sigma_2(W_2\sigma_1(W_1X)) \cdots))),$$

where $q \in \mathbb{R}$ is simply a normalization factor, the value of which is specified later. Here, $\sigma_k : \mathbb{R}^{d_k \times m} \rightarrow \mathbb{R}^{d_k \times m}$ is the element-wise rectified linear function:

$$\sigma_k \left(\begin{bmatrix} b_{11} & \dots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{d_k 1} & \cdots & b_{d_k m} \end{bmatrix} \right) = \begin{bmatrix} \bar{\sigma}(b_{11}) & \dots & \bar{\sigma}(b_{1m}) \\ \vdots & \ddots & \vdots \\ \bar{\sigma}(b_{d_k 1}) & \cdots & \bar{\sigma}(b_{d_k m}) \end{bmatrix},$$

where $\bar{\sigma}(b_{ij}) = \max(0, b_{ij})$. In practice, we usually set σ_{H+1} to be an identity map in the last layer, in which case all our theoretical results still hold true.

3.2 Background

Following the work by Dauphin *et al.* (2014), Choromanska *et al.* (2015a) investigated the connection between the loss functions of deep nonlinear networks and a function well-studied via random matrix theory (i.e., the Hamiltonian of the spherical spin-glass model). They explained that their theoretical results relied on several *unrealistic* assumptions. Later, Choromanska *et al.* (2015b) suggested at the Conference on Learning Theory (COLT) 2015 that discarding these assumptions is an important open problem. The assumptions were labeled A1p, A2p, A3p, A4p, A5u, A6u, and A7p.

In this paper, we successfully discard most of these assumptions. In particular, we only use a weaker version of assumptions A1p and A5u. We refer to the part of assumption A1p (resp. A5u) that corresponds only to the *model* assumption as A1p-m (resp. A5u-m). Note that assumptions A1p-m and A5u-m are explicitly used in the previous work (Choromanska *et al.*, 2015a) and included in A1p and A5u (i.e., we are *not* making new assumptions here).

As the model $\hat{Y}(W, X) \in \mathbb{R}^{d_y \times m}$ represents a directed acyclic graph, we can express an output from one of the units in the output layer as

$$\hat{Y}(W, X)_{j,i} = q \sum_{p=1}^{\Psi} [X_i]_{(j,p)} [Z_i]_{(j,p)} \prod_{k=1}^{H+1} w_{(j,p)}^{(k)}. \tag{1}$$

Here, Ψ is the total number of paths from the inputs to each j-th output in the directed acyclic graph. In addition, $[X_i]_{(j,p)} \in \mathbb{R}$ represents the entry of the i-th sample input datum that is used in the p-th path of the j-th output. For each layer k, $w_{(j,p)}^{(k)} \in \mathbb{R}$ is the entry of W_k that is used in the p-th path of the j-th output. Finally, $[Z_i]_{(j,p)} \in \{0, 1\}$ represents whether the p-th path of the j-th output is active ($[Z_i]_{(j,p)} = 1$) or not ($[Z_i]_{(j,p)} = 0$) for each sample i as a result of the rectified linear activation.

Assumption A1p-m assumes that the Z's are Bernoulli random variables with the same probability of success, $\Pr([Z_i]_{(j,p)} = 1) = \rho$ for all i and (j, p). Assumption A5u-m assumes that the Z's are independent from the input X's and parameters w's. With assumptions A1p-m and A5u-m, we can write $\mathbb{E}_Z[\hat{Y}(W, X)_{j,i}] = q \sum_{p=1}^{\Psi} [X_i]_{(j,p)} \rho \prod_{k=1}^{H+1} w_{(j,p)}^{(k)}$.

Choromanska *et al.* (2015b) noted that A6u is unrealistic because it implies that the inputs are not shared among the paths. In addition, Assumption A5u is unrealistic because it implies that the activation of any path is independent of the input data. To understand all of the seven assumptions (A1p, A2p, A3p, A4p, A5u, A6u, and A7p), we note that Choromanska *et al.* (2015b,a) used these seven assumptions to reduce their loss functions of nonlinear neural networks to:

$$\mathcal{L}_{\text{previous}}(W) = \frac{1}{\lambda^{H/2}} \sum_{i_1, i_2, \dots, i_{H+1}=1}^{\lambda} X_{i_1, i_2, \dots, i_{H+1}} \prod_{k=1}^{H+1} w_{i_k} \text{ subject to } \frac{1}{\lambda} \sum_{i=1}^{\lambda} w_i^2 = 1,$$

where $\lambda \in \mathbb{R}$ is a constant related to the size of the network. For our purpose, the detailed definitions of the symbols are not important (X and w are defined in the same way as in equation 1). Here, we point out that *the target function Y has disappeared in the loss $\mathcal{L}_{previous}(W)$* (i.e., the loss value does not depend on the target function). That is, whatever the data points of Y are, their loss values are the same. Moreover, *the nonlinear activation function has disappeared in $\mathcal{L}_{previous}(W)$* (and the nonlinearity is not taken into account in X or w). In the next section, by using only a strict subset of the set of these seven assumptions, we reduce our loss function to a more realistic loss function of an actual deep model.

Proposition 3.1 (High-level description of a main result in Choromanska *et al.*, 2015a) *Assume A1p (including A1p-m), A2p, A3p, A4p, A5u (including A5u-m), A6u, and A7p (Choromanska et al., 2015b). Furthermore, assume that $d_y = 1$. Then, the expected loss of each sample datum, $\mathcal{L}_{previous}(W)$, has the following property: above a certain loss value, the number of local minima diminishes exponentially as the loss value increases.*

3.3 Results

We now state our theoretical result, which partially address the aforementioned open problem. We consider loss functions for all the data points and all possible output dimensionalities (i.e., vectored-valued output). More concretely, we consider the squared error loss with expectation, $\mathcal{L}(W) = \frac{1}{2}\|E_Z[\hat{Y}(W, X) - Y]\|_F^2$.

Corollary 3.2 (Loss surface of deep nonlinear networks) *Assume A1p-m and A5u-m. Let $q = \rho^{-1}$. Then, we can reduce the loss function of the deep nonlinear model $\mathcal{L}(W)$ to that of the deep linear model $\bar{\mathcal{L}}(W)$. Therefore, with the same set of conditions as in Theorem 2.3, the loss function of the deep nonlinear model has the following properties:*

(i) It is non-convex and non-concave.

(ii) Every local minimum is a global minimum.

(iii) Every critical point that is not a global minimum is a saddle point.

(iv) The saddle points have the properties stated in Theorem 2.3 (iv) and Corollary 2.4.

Comparing Corollary 3.2 and Proposition 3.1, we can see that we successfully discarded assumptions A2p, A3p, A4p, A6u, and A7p while obtaining a tighter statement in the following sense: Corollary 3.2 states with fewer unrealistic assumptions that there is no poor local minimum, whereas Proposition 3.1 roughly asserts with more unrealistic assumptions that the number of poor local minimum may be not too large. Furthermore, our model $\hat{Y}$ is strictly more general than the model analyzed in (Choromanska *et al.*, 2015a,b) (i.e., this paper's model class contains the previous work's model class but not vice versa).

4 Proof Idea and Important lemmas

In this section, we provide overviews of the proofs of the theoretical results. Our proof approach largely differs from those in previous work (Baldi & Hornik, 1989; Baldi & Lu, 2012; Choromanska *et al.*, 2015a,b). In contrast to (Baldi & Hornik, 1989; Baldi & Lu, 2012), we need a different approach to deal with the "bad" saddle points that start appearing when the model becomes deeper (see Section 2.3), as well as to obtain more comprehensive properties of the critical points with more generality. While the previous proofs heavily rely on the first-order information, the main parts of our proofs take advantage of the second order information. In contrast, Choromanska *et al.* (2015a,b) used the seven assumptions to relate the loss functions of deep models to a function previously analyzed with a tool of random matrix theory. With no reshaping assumptions (A3p, A4p, and A6u), we cannot relate our loss function to such a function. Moreover, with no distributional assumptions (A2p and A6u) (except the activation), our Hessian is deterministic, and therefore, even random matrix theory itself is insufficient for our purpose. Furthermore, with no spherical constraint assumption (A7p), the number of local minima in our loss function can be uncountable.

One natural strategy to proceed toward Theorem 2.3 and Corollary 3.2 would be to use the first-order and second-order necessary conditions of local minima (e.g., the gradient is zero and the Hessian is

positive semidefinite).[4] However, are the first-order and second-order conditions sufficient to prove Theorem 2.3 and Corollary 3.2? Corollaries 2.4 show that the answer is negative for *deep* models with $H \geq 2$, while it is affirmative for shallow models with $H = 1$. Thus, for deep models, a simple use of the first-order and second-order information is insufficient to characterize the properties of each critical point. In addition to the complexity of the Hessian of the *deep* models, this suggests that we must strategically extract the second order information. Accordingly, in section 4.2, we obtain an organized representation of the Hessian in Lemma 4.3 and strategically extract the information in Lemmas 4.4 and 4.6. With the extracted information, we discuss the proofs of Theorem 2.3 and Corollary 3.2 in section 4.3.

4.1 Notations

Let $M \otimes M'$ be the Kronecker product of M and M'. Let $\mathcal{D}_{\text{vec}(W_k^T)} f(\cdot) = \frac{\partial f(\cdot)}{\partial_{\text{vec}(W_k^T)}}$ be the partial derivative of f with respect to $\text{vec}(W_k^T)$ in the numerator layout. That is, if $f : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$, we have $\mathcal{D}_{\text{vec}(W_k^T)} f(\cdot) \in \mathbb{R}^{d_{out} \times (d_k d_{k-1})}$. Let $\mathcal{R}(M)$ be the range (or the column space) of a matrix M. Let M^- be any generalized inverse of M. When we write a generalized inverse in a condition or statement, we mean it for any generalized inverse (i.e., we omit the universal quantifier over generalized inverses, as this is clear). Let $r = (\overline{Y}(W, X) - Y)^T \in \mathbb{R}^{m \times d_y}$ be an error matrix. Let $C = W_{H+1} \cdots W_2 \in \mathbb{R}^{d_y \times d_1}$. When we write $W_k \cdots W_{k'}$, we generally intend that $k > k'$ and the expression denotes a product over W_j for integer $k \geq j \geq k'$. For notational compactness, two additional cases can arise: when $k = k'$, the expression denotes simply W_k, and when $k < k'$, it denotes I_{d_k}. For example, in the statement of Lemma 4.1, if we set $k := H + 1$, we have that $W_{H+1} W_H \cdots W_{H+2} \triangleq I_{d_y}$.

In Lemma 4.6 and the proofs of Theorems 2.3, we use the following additional notation. We denote an eigendecomposition of Σ as $\Sigma = U \Lambda U^T$, where the entries of the eigenvalues are ordered as $\Lambda_{1,1} > \cdots > \Lambda_{d_y, d_y}$ with corresponding orthogonal eigenvector matrix $U = [u_1, \ldots, u_{d_y}]$. For each $k \in \{1, \ldots d_y\}$, $u_k \in \mathbb{R}^{d_y \times 1}$ is a column eigenvector. Let $\bar{p} = \text{rank}(C) \in \{1, \ldots, \min(d_y, p)\}$. We define a matrix containing the subset of the $\bar{p}$ largest eigenvectors as $U_{\bar{p}} = [u_1, \ldots, u_{\bar{p}}]$. Given any ordered set $\mathcal{I}_{\bar{p}} = \{i_1, \ldots, i_{\bar{p}} \mid 1 \leq i_1 < \cdots < i_{\bar{p}} \leq \min(d_y, p)\}$, we define a matrix containing the subset of the corresponding eigenvectors as $U_{\mathcal{I}_{\bar{p}}} = [u_{i_1}, \ldots, u_{i_{\bar{p}}}]$. Note the difference between $U_{\bar{p}}$ and $U_{\mathcal{I}_{\bar{p}}}$.

4.2 Lemmas

As discussed above, we extracted the first-order and second-order conditions of local minima as the following lemmas. The lemmas provided here are also intended to be our additional theoretical results that may lead to further insights. The proofs of the lemmas are in the appendix.

Lemma 4.1 (Critical point necessary and sufficient condition) *W is a critical point of $\bar{\mathcal{L}}(W)$ if and only if for all $k \in \{1, ..., H+1\}$,*

$$\left(\mathcal{D}_{\text{vec}(W_k^T)} \bar{\mathcal{L}}(W)\right)^T = \left(W_{H+1} W_H \cdots W_{k+1} \otimes (W_{k-1} \cdots W_2 W_1 X)^T\right)^T \text{vec}(r) = 0.$$

Lemma 4.2 (Representation at critical point) *If W is a critical point of $\bar{\mathcal{L}}(W)$, then*

$$W_{H+1} W_H \cdots W_2 W_1 = C(C^T C)^- C^T Y X^T (X X^T)^{-1}.$$

Lemma 4.3 (Block Hessian with Kronecker product) *Write the entries of $\nabla^2 \bar{\mathcal{L}}(W)$ in a block form as*

$$\nabla^2 \bar{\mathcal{L}}(W) = \begin{bmatrix} \mathcal{D}_{\text{vec}(W_{H+1}^T)} \left(\mathcal{D}_{\text{vec}(W_{H+1}^T)} \bar{\mathcal{L}}(W)\right)^T & \cdots & \mathcal{D}_{\text{vec}(W_1^T)} \left(\mathcal{D}_{\text{vec}(W_{H+1}^T)} \bar{\mathcal{L}}(W)\right)^T \\ \vdots & \ddots & \vdots \\ \mathcal{D}_{\text{vec}(W_{H+1}^T)} \left(\mathcal{D}_{\text{vec}(W_1^T)} \bar{\mathcal{L}}(W)\right)^T & \cdots & \mathcal{D}_{\text{vec}(W_1^T)} \left(\mathcal{D}_{\text{vec}(W_1^T)} \bar{\mathcal{L}}(W)\right)^T \end{bmatrix}.$$

[4] For a non-convex and *non-differentiable* function, we can still have a first-order and second-order necessary condition (e.g., Rockafellar & Wets, 2009, theorem 13.24, p. 606).

Then, for any $k \in \{1, ..., H+1\}$,

$$\mathcal{D}_{\mathrm{vec}(W_k^T)} \left(\mathcal{D}_{\mathrm{vec}(W_k^T)} \bar{\mathcal{L}}(W)\right)^T$$
$$= \left((W_{H+1}\cdots W_{k+1})^T (W_{H+1}\cdots W_{k+1}) \otimes (W_{k-1}\cdots W_1 X)(W_{k-1}\cdots W_1 X)^T\right),$$

and*, for any* $k \in \{2, ..., H+1\}$,

$$\mathcal{D}_{\mathrm{vec}(W_k^T)} \left(\mathcal{D}_{\mathrm{vec}(W_1^T)} \bar{\mathcal{L}}(W)\right)^T$$
$$= \left(C^T (W_{H+1}\cdots W_{k+1}) \otimes X(W_{k-1}\cdots W_1 X)^T\right) +$$
$$\left[(W_{k-1}\cdots W_2)^T \otimes X\right] \left[I_{d_{k-1}} \otimes (rW_{H+1}\cdots W_{k+1})_{\cdot,1} \quad \ldots \quad I_{d_{k-1}} \otimes (rW_{H+1}\cdots W_{k+1})_{\cdot,d_k}\right].$$

Lemma 4.4 (Hessian semidefinite necessary condition) *If* $\nabla^2 \bar{\mathcal{L}}(W)$ *is positive semidefinite or negative semidefinite at a critical point, then for any* $k \in \{2, ..., H+1\}$,

$$\mathcal{R}((W_{k-1}\cdots W_3 W_2)^T) \subseteq \mathcal{R}(C^T C) \quad \textbf{\textit{or}} \quad XrW_{H+1}W_H \cdots W_{k+1} = 0.$$

Corollary 4.5 *If* $\nabla^2 \bar{\mathcal{L}}(W)$ *is positive semidefinite or negative semidefinite at a critical point, then for any* $k \in \{2, ..., H+1\}$,

$$\mathrm{rank}(W_{H+1}W_H \cdots W_k) \geq \mathrm{rank}(W_{k-1}\cdots W_3 W_2) \quad \textbf{\textit{or}} \quad XrW_{H+1}W_H \cdots W_{k+1} = 0.$$

Lemma 4.6 (Hessian positive semidefinite necessary condition) *If* $\nabla^2 \bar{\mathcal{L}}(W)$ *is positive semidefinite at a critical point, then*

$$C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}^T \quad \textbf{\textit{or}} \quad Xr = 0.$$

4.3 Proof sketches of theorems

We now provide the proof sketch of Theorem 2.3 and Corollary 3.2. We complete the proofs in the appendix.

4.3.1 Proof sketch of Theorem 2.3 *(ii)*

By case analysis, we show that any point that satisfies the necessary conditions and the definition of a local minimum is a global minimum.

Case I: $\mathrm{rank}(W_H \cdots W_2) = p$ and $d_y \leq p$: If $d_y < p$, Corollary 4.5 with $k = H+1$ implies the necessary condition of local minima that $Xr = 0$. If $d_y = p$, Lemma 4.6 with $k = H+1$ and $k = 2$, combined with the fact that $\mathcal{R}(C) \subseteq \mathcal{R}(YX^T)$, implies the necessary condition that $Xr = 0$. Therefore, we have the necessary condition of local minima, $Xr = 0$. Interpreting condition $Xr = 0$, we conclude that W achieving $Xr = 0$ is indeed a global minimum.

Case II: $\mathrm{rank}(W_H \cdots W_2) = p$ and $d_y > p$: From Lemma 4.6, we have the necessary condition that $C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}^T$ or $Xr = 0$. If $Xr = 0$, using the exact same proof as in Case I, it is a global minimum. Suppose then that $C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}^T$. From Lemma 4.4 with $k = H+1$, we conclude that $\bar{p} \triangleq \mathrm{rank}(C) = p$. Then, from Lemma 4.2, we write $W_{H+1}\cdots W_1 = U_p U_p^T YX^T (XX^T)^{-1}$, which is the orthogonal projection onto the subspace spanned by the p eigenvectors corresponding to the p largest eigenvalues following the ordinary least square regression matrix. This is indeed the expression of a global minimum.

Case III: $\mathrm{rank}(W_H \cdots W_2) < p$: We first show that if $\mathrm{rank}(C) \geq \min(p, d_y)$, every local minimum is a global minimum. Thus, we consider the case where $\mathrm{rank}(W_H \cdots W_2) < p$ and $\mathrm{rank}(C) < \min(p, d_y)$. In this case, by induction on $k = \{1, \ldots, H+1\}$, we prove that we can have $\mathrm{rank}(W_k \cdots W_1) \geq \min(p, d_y)$ with arbitrarily small perturbation of each entry of $W_k, \ldots, W_1$ without changing the value of $\bar{\mathcal{L}}(W)$. Once this is proved, along with the results of Case I and Case II, we can immediately conclude that any point satisfying the definition of a local minimum is a global minimum.

We first prove the statement for the base case with $k = 1$ by using an expression of W_1 that is obtained by a first-order necessary condition: for an arbitrary L_1,

$$W_1 = (C^T C)^- C^T YX^T (XX^T)^{-1} + (I - (C^T C)^- C^T C) L_1.$$

By using Lemma 4.6 to obtain an expression of C, we deduce that we can have $\mathrm{rank}(W_1) \geq \min(p, d_y)$ with arbitrarily small perturbation of each entry of W_1 without changing the loss value.

For the inductive step with $k \in \{2, \ldots, H+1\}$, from Lemma 4.4, we use the following necessary condition for the Hessian to be (positive or negative) semidefinite at a critical point: for any $k \in \{2, \ldots, H+1\}$,

$$\mathcal{R}((W_{k-1} \cdots W_2)^T) \subseteq \mathcal{R}(C^T C) \quad \textbf{or} \quad XrW_{H+1} \cdots W_{k+1} = 0.$$

We use the inductive hypothesis to conclude that the first condition is false, and thus the second condition must be satisfied at a candidate point of a local minimum. From the latter condition, with extra steps, we can deduce that we can have $\mathrm{rank}(W_k W_{k-1} \cdots W_1) \geq \min(p, d_x)$ with arbitrarily small perturbation of each entry of W_k while retaining the same loss value.

We conclude the induction, proving that we can have $\mathrm{rank}(C) \geq \mathrm{rank}(W_{H+1} \cdots W_1) \geq \min(p, d_x)$ with arbitrarily small perturbation of each parameter without changing the value of $\bar{\mathcal{L}}(W)$. Upon such a perturbation, we have the case where $\mathrm{rank}(C) \geq \min(p, d_y)$, for which we have already proven that every local minimum is a global minimum. Summarizing the above, any point that satisfies the definition (and necessary conditions) of a local minimum is indeed a global minimum. Therefore, we conclude the proof sketch of Theorem 2.3 *(ii)*.

4.3.2 Proof sketch of Theorem 2.3 *(i), (iii)* and *(iv)*

We can prove the non-convexity and non-concavity of this function simply from its Hessian (Theorem 2.3 *(i)*). That is, we can show that in the domain of the function, there exist points at which the Hessian becomes indefinite. Indeed, the domain contains uncountably many points at which the Hessian is indefinite.

We now consider Theorem 2.3 (*iii*): every critical point that is not a global minimum is a saddle point. Combined with Theorem 2.3 *(ii)*, which is proven independently, this is equivalent to the statement that there are no local maxima. We first show that if $W_{H+1} \cdots W_2 \neq 0$, the loss function always has some strictly increasing direction with respect to W_1, and hence there is no local maximum. If $W_{H+1} \cdots W_2 = 0$, we show that at a critical point, if the Hessian is negative semidefinite (i.e., a necessary condition of local maxima), we can have $W_{H+1} \cdots W_2 \neq 0$ with arbitrarily small perturbation without changing the loss value. We can prove this by induction on $k = 2, \ldots, H+1$, similar to the induction in the proof of Theorem 2.3 *(ii)*. This means that there is no local maximum.

Theorem 2.3 *(iv)* follows Theorem 2.3 *(ii)*-*(iii)* and the analyses for Case I and Case II in the proof of Theorem 2.3 *(ii)*; when $\mathrm{rank}(W_H \cdots W_2) = p$, if $\nabla^2 \bar{\mathcal{L}}(W) \succeq 0$ at a critical point, W is a global minimum.

4.3.3 Proof sketch of Corollary 3.2

Since the activations are assumed to be random and independent, the effect of nonlinear activations disappear by taking expectation. As a result, the loss function $\mathcal{L}(W)$ is reduced to $\bar{\mathcal{L}}(W)$.

5 Conclusion

In this paper, we addressed some open problems, pushing forward the theoretical foundations of deep learning and non-convex optimization. For deep *linear* neural networks, we proved the aforementioned conjecture and more detailed statements with more generality. For deep *nonlinear* neural networks, when compared with the previous work, we proved a tighter statement (in the way explained in section 3) with more generality (d_y can vary) and with strictly weaker model assumptions (only two assumptions out of seven). However, our theory does not yet directly apply to the practical situation. To fill the gap between theory and practice, future work would further discard the remaining two out of the seven assumptions made in previous work. Our new understanding of the deep linear models at least provides the following theoretical fact: the bad local minima would arise in a deep nonlinear model but *only as an effect of adding nonlinear activations* to the corresponding *deep* linear model. Thus, depending on the nonlinear activation mechanism and architecture, we would be able to efficiently train *deep* models.

Acknowledgments

The author would like to thank Prof. Leslie Kaelbling, Quynh Nguyen, Li Huan and Anirbit Mukherjee for their thoughtful comments on the paper. We gratefully acknowledge support from NSF grant 1420927, from ONR grant N00014-14-1-0486, and from ARO grant W911NF1410433.

References

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

Baldi, Pierre, & Hornik, Kurt. 1989. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, **2**(1), 53–58.

Baldi, Pierre, & Lu, Zhiqin. 2012. Complex-valued autoencoders. *Neural Networks*, **33**, 136–147.

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

Choromanska, Anna, Henaff, MIkael, Mathieu, Michael, Ben Arous, Gerard, & LeCun, Yann. 2015a. The Loss Surfaces of Multilayer Networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. pp. 192–204.

Choromanska, Anna, LeCun, Yann, & Arous, Gérard Ben. 2015b. Open Problem: The landscape of the loss surfaces of multilayer networks. In *Proceedings of The 28th Conference on Learning Theory*. pp. 1756–1760.

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

Ge, Rong, Huang, Furong, Jin, Chi, & Yuan, Yang. 2015. Escaping From Saddle Points—Online Stochastic Gradient for Tensor Decomposition. In *Proceedings of The 28th Conference on Learning Theory*. pp. 797–842.

Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron. 2016. *Deep Learning*. Book in preparation for MIT Press. `http://www.deeplearningbook.org`.

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

Mhaskar, Hrushikesh, Liao, Qianli, & Poggio, Tomaso. 2016. Learning Real and Boolean Functions: When Is Deep Better Than Shallow. *Massachusetts Institute of Technology CBMM Memo No. 45*.

Murty, Katta G, & Kabadi, Santosh N. 1987. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical programming*, **39**(2), 117–129.

Rockafellar, R Tyrrell, & Wets, Roger J-B. 2009. *Variational analysis*. Vol. 317. Springer Science & Business Media.

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

Zhang, Fuzhen. 2006. *The Schur complement and its applications*. Vol. 4. Springer Science & Business Media.