Optimizing Neural Networks: A Review of "Deep Learning without Poor Local Minima" and "Identity Matters in Deep Learning"

Nicholas K. Branigan, Akash Velu

December 3, 2021

1 "Deep Learning without Poor Local Minima" [6]

1.1 Problem Statement

1.1.1 The non-convexity of neural network optimization

Optimizing a neural network, even in a simple case, is a non-convex optimization problem. Finding the global minimum for a general non-convex function is NP-Complete, and prior work has demonstrated that, even for simple neural networks, the global optimization problem is NP-Hard [3]. Consequently, major optimization methods for training neural networks, like gradient descent, are not guaranteed to converge to global minima, or even local minima due to the presence of saddle points [7]. (This is in contrast with other popular machine learning models, e.g., support vector machines, which are convex in their objectives). Interestingly, in spite of these apparent hurdles, deep neural networks have achieved remarkable empirical performance. Explaining the success of these networks and characterizing their optimization landscapes is an area of active research.

To simplify the analysis of the optimization of neural networks, prior work has proposed analyzing the optimization of *linear* neural networks [2]. These networks are a composition of linear maps, so they can only parameterize linear functions. Despite this, the analysis of linear neural networks appears to be a reasonable starting point. Linear neural network optimization is not trivial: the loss function remains non-convex with respect to the model's weights, and, importantly, there are similarities in the optimization of linear and general neural networks [2].

1.1.2 Notation

We follow the notation in the paper we are reviewing, which subsequently we will refer to simply as Kawaguchi. Let $\boldsymbol{x} \in \mathbb{R}^{d_x}$, $\boldsymbol{y} \in \mathbb{R}^{d_y}$. Suppose we have m observations $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^m$, and define $\boldsymbol{X} \in \mathbb{R}^{d_x \times m}$ to be a design matrix containing all input training points, and similarly define $\boldsymbol{Y} \in \mathbb{R}^{d_y \times m}$ to be a matrix containing all corresponding output training points. Let $\boldsymbol{\Sigma} = \boldsymbol{Y}\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{X}^\intercal)^{-1}\boldsymbol{X}\boldsymbol{Y}^\intercal$. The hypothesis class of interest is linear feed-forward neural networks with H hidden layers. Consider the input layer to be the 0-th layer and the output layer to be the $(H+1)$-th layer, and define d_i be width of the i-th layer. The neural network is then parameterized by $H+1$ weight matrices $\boldsymbol{W} = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{H+1}\}$, where $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$. The model's prediction on all the input points is $\overline{\boldsymbol{Y}}(\boldsymbol{W}, \boldsymbol{X}) = \boldsymbol{W}_{H+1}\boldsymbol{W}_H ... \boldsymbol{W}_1\boldsymbol{X}$. To train the model, we use the maximum likelihood loss function and minimize the empirical risk,

$$\bar{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{X}) = \frac{1}{2}\sum_{i=1}^{m} \left\| \overline{\boldsymbol{Y}}(\boldsymbol{W}, \boldsymbol{X})_i - \boldsymbol{y}_i \right\|^2 = \frac{1}{2} \left\| \overline{\boldsymbol{Y}}(\boldsymbol{W}, \boldsymbol{X}) - \boldsymbol{Y} \right\|_F^2 .$$

1.1.3 Prior work

The analysis of the optimization properties and loss landscapes of linear neural networks has been the focus of several papers. Most relevant to Kawaguchi is [1], which proves the following result about *shallow* linear neural networks.

Proposition 1.1.1. *Suppose $H = 1$, $\boldsymbol{X}\boldsymbol{X}^\intercal$ and $\boldsymbol{X}\boldsymbol{Y}^\intercal$ are invertible, and $\boldsymbol{\Sigma}$ has d_y distinct eigenvalues. Also assume that $d_x = d_y$ and $d_1 < d_x$. Then $\bar{\mathcal{L}}$ is convex with respect to one weight matrix when the other weight matrix is kept constant (e.g., if $\boldsymbol{W}_2$ is fixed, $\bar{\mathcal{L}}$ is convex in $\boldsymbol{W}_1$), and every local minimum is a global minimum.*

A natural follow on question is whether a similar result holds for deep networks. [1] state the following conjecture without proof.

Conjecture 1.1.1. *Suppose all assumptions from Proposition 1.1.1, with the exception of $H = 1$, hold. Then, for $k \in \{1, \dots, H+1\}$, if $\boldsymbol{W}_j$ is kept fixed for all $j \neq k$, $\bar{\mathcal{L}}$ is convex with respect to $\boldsymbol{W}_k$, and every local minimum is a global minimum.*

1.1.4 This paper's contributions

[2] proved the first part of Conjecture 1.1.1. Kawaguchi focuses on proving a result which will give the second part, and more. We discuss the significance and implications of this result after stating the theorem.

1.2 Main Result

The result at the heart of Kawaguchi is Theorem 2.3.

Theorem 2.3. *Suppose $\boldsymbol{X}\boldsymbol{X}^\intercal$ and $\boldsymbol{X}\boldsymbol{Y}^\intercal$ are of full rank, $d_y \leq d_x$, and $\boldsymbol{\Sigma}$ has d_y distinct eigenvalues. Then the following results hold regarding the loss function $\bar{\mathcal{L}}$:*

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

(ii) Every local minimum is a global minimum.

(iii) If a critical point is not a global minimum, it is a saddle point.

(iv) If $H \geq 2$ and $\text{rank}(\boldsymbol{W}_H \cdots \boldsymbol{W}_2) = \min(d_H, \dots, d_1)$ or if $H = 1$, then at any saddle point the Hessian has at least one negative eigenvalue.

Theorem 2.3 establishes that for deep linear networks, under certain reasonable[1] assumptions regarding the training data, there are no bad local minima. For gradient based methods, this leaves only saddle-points as potential problems where optimizers could get trapped. However, for special networks, Theorem 2.3 has an answer to this problem. For very shallow networks ($H = 1$), saddle points are guaranteed to have at least one negative eigenvalue. In order for a point to be a local minimum, the Hessian at the point must be positive semidefinite. Thus, a critical point with at least one Hessian negative eigenvalue cannot be a local minimum. Then, by Theorem 2.3 (iii), such a point must be a saddle point. So under this regime, if our optimization algorithm is approaching a saddle point, we are guaranteed to see this since the Hessian is Lipschitz continuous.

1.3 Examples

In this section we probe Theorem 2.3, first by instantiating it, then by picking apart its conditions.

Claim 1.3.1. *When $d_y = 1$, $\boldsymbol{\Sigma}$ has d_y distinct eigenvalues.*

Proof. $\boldsymbol{\Sigma} = \boldsymbol{Y}\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{X}^\intercal)^{-1}\boldsymbol{X}\boldsymbol{Y}^\intercal$, so $\boldsymbol{\Sigma} \in \mathbb{R}^{d_y \times d_y}$. For any $\boldsymbol{A} \in \mathbb{R}^{1\times 1}$, $\boldsymbol{A}$ is a diagonal matrix. The diagonal entries of a diagonal matrix are its eigenvalues. So, $\boldsymbol{A}$ has one eigenvalue: $\boldsymbol{A}_{11}$. □

Example 1.3.1. Let

$$\boldsymbol{X} = \begin{bmatrix} 1 & 2 \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} 2 & 4 \end{bmatrix}, \quad \boldsymbol{W}_2 = \boldsymbol{A} = \begin{bmatrix} a_1 & a_2 \end{bmatrix}, \quad \boldsymbol{W}_1 = \boldsymbol{B} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

We check

$$\boldsymbol{X}\boldsymbol{X}^\intercal = \begin{bmatrix} 5 \end{bmatrix} \quad \text{and} \quad \boldsymbol{X}\boldsymbol{Y}^\intercal = \begin{bmatrix} 10 \end{bmatrix}.$$

[1] See Remark 1.3.1.

Then, by Claim 1.3.1, all of the conditions on Theorem 2.3 are met.

$$\bar{\mathcal{L}}(\boldsymbol{AB}) = \tfrac{1}{2}\left((-2 + a_1b_1 + a_2b_2)^2 + (-4 + 2(a_1b_1 + a_2b_2))^2\right).$$

$$\nabla\bar{\mathcal{L}}(\boldsymbol{AB}) = \begin{bmatrix} \frac{\partial\bar{\mathcal{L}}(\boldsymbol{AB})}{\partial a_1} \\ \frac{\partial\bar{\mathcal{L}}(\boldsymbol{AB})}{\partial a_2} \\ \frac{\partial\bar{\mathcal{L}}(\boldsymbol{AB})}{\partial b_1} \\ \frac{\partial\bar{\mathcal{L}}(\boldsymbol{AB})}{\partial b_2} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} (2b_1(-2 + a_1b_1 + a_2b_2) + 4b_1(-4 + 2(a_1b_1 + a_2b_2))) \\ (2b_2(-2 + a_1b_1 + a_2b_2) + 4b_2(-4 + 2(a_1b_1 + a_2b_2))) \\ (2a_1(-2 + a_1b_1 + a_2b_2) + 4a_1(-4 + 2(a_1b_1 + a_2b_2))) \\ (2a_2(-2 + a_1b_1 + a_2b_2) + 4a_2(-4 + 2(a_1b_1 + a_2b_2))) \end{bmatrix}.$$

$$\nabla^2\bar{\mathcal{L}}(\boldsymbol{AB}) = \begin{bmatrix} 5b_1^2 & 5b_1b_2 & \mathcal{E}_1 & 5a_2b_1 \\ 5b_1b_2 & 5b_2^2 & 5a_1b_2 & \mathcal{E}_2 \\ \mathcal{E}_1 & 5a_1b_2 & 5a_1^2 & 5a_1a_2 \\ 5a_2b_1 & \mathcal{E}_2 & 5a_1a_2 & 5a_2^2 \end{bmatrix},$$

where $\mathcal{E}_1 = 1/2\left(10a_1b_1 + 2(-2 + a_1b_1 + a_2b_2) + 4(-4 + 2(a_1b_1 + a_2b_2))\right),$

$\mathcal{E}_2 = 1/2\left(10a_2b_2 + 2(-2 + a_1b_1 + a_2b_2) + 4(-4 + 2(a_1b_1 + a_2b_2))\right).$

Letting $\boldsymbol{\Lambda}(\boldsymbol{AB})$ be a vector whose entries are the eigenvalues of $\nabla^2\bar{\mathcal{L}}(\boldsymbol{AB})$,

$$\boldsymbol{\Lambda}(\boldsymbol{AB}) = \begin{bmatrix} 5(2 - a_1b_1 - a_2b_2) \\ 5(-2 + a_1b_1 + a_2b_2) \\ 5/2\left(a_1^2 + a_2^2 + b_1^2 + b_2^2 - \mathcal{S}\right) \\ 5/2\left(a_1^2 + a_2^2 + b_1^2 + b_2^2 + \mathcal{S}\right) \end{bmatrix},$$

where $\mathcal{S} = \sqrt{(a_1^2 + a_2^2 + b_1^2 + b_2^2)^2 + 4(4 - 8a_1b_1 + 3a_1^2b_1^2 - 8a_2b_2 + 6a_1a_2b_1b_2 + 3a_2^2b_2^2)}.$

Solving $\nabla\bar{\mathcal{L}}(\boldsymbol{AB}) = \mathbf{0}$, we find that the critical points of $\bar{\mathcal{L}}(\boldsymbol{AB})$ are

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ (2 - a_1b_1)/a_2 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} a_1 \\ 0 \\ 2/a_1 \\ b_2 \end{bmatrix} \quad \text{for} \quad a_1, a_2, b_1, b_2 \in \mathbb{R}.$$

We can now verify the guarantees of Theorem 2.3. First, it is immediately clear from $\boldsymbol{\Lambda}$ that $\bar{\mathcal{L}}(\boldsymbol{AB})$ has positive and negative eigenvalues on its domain, implying that it is non-convex and non-concave as Theorem 2.3 (i) says it should be. Next, we examine the eigenvalues of the Hessian at these critical points.

$$\boldsymbol{\Lambda}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} -10 \\ -10 \\ 10 \\ 10 \end{bmatrix},$$

$$\boldsymbol{\Lambda}\left(\begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ (2 - a_1b_1)/a_2 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 5(4 + a_1^4 + a_1^2b_2^2)/a_1^2 \end{bmatrix}, \quad \boldsymbol{\Lambda}\left(\begin{bmatrix} a_1 \\ 0 \\ 2/a_1 \\ b_2 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 5(4 + a_1^2a_2^2 + a_2^4 - 4a_1b_1 + a_1^2b_1^2 + a_2^2b_1^2)/a_2^2 \end{bmatrix}.$$

The second derivative test indicates that $\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^\intercal$ is a saddle point, but is inconclusive on the other critical points. Nevertheless, we can show that these other points are all local and global minima.

$$\bar{\mathcal{L}}\left(\begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ (2 - a_1b_1)/a_2 \end{bmatrix}\right) = 0 \quad \text{and} \quad \bar{\mathcal{L}}\left(\begin{bmatrix} a_1 \\ 0 \\ 2/a_1 \\ b_2 \end{bmatrix}\right) = 0.$$

Since $\bar{\mathcal{L}}(\boldsymbol{AB}) \geq 0$, all the critical points other than $\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^\intercal$ are local and global minima. This confirms that the implications of Theorem 2.3 (ii), (iii), and (iv) hold on this example.

Remark 1.3.1. Theorem 2.3 has four requirements: $\boldsymbol{X}\boldsymbol{X}^\intercal$ and $\boldsymbol{X}\boldsymbol{Y}^\intercal$ are full rank, $d_y \leq d_x$, and $\boldsymbol{\Sigma}$ has d_y distinct eigenvalues. The first assumption is commonplace and easy to satisfy. Its failure corresponds to the situation where the predictors are linearly dependent. If we observe this with m large, we can safely drop predictors to make the remaining ones linearly independent without being concerned about losing predictive information. The assumption on $\boldsymbol{X}\boldsymbol{Y}^\intercal$ appears in the proof of Theorem 2.3 (ii). The authors need $\boldsymbol{X}\boldsymbol{Y}^\intercal\boldsymbol{M}_1 = \boldsymbol{X}\boldsymbol{Y}^\intercal\boldsymbol{M}_2$ to imply $\boldsymbol{M}_1 = \boldsymbol{M}_2$ for matrices $\boldsymbol{M}_1, \boldsymbol{M}_2$. Since $\boldsymbol{X}\boldsymbol{Y}^\intercal$ is full rank, it is one-to-one and this follows. However, it is not obvious whether if this assumption fails, the results do not hold (i.e., there exists a counterexample), or whether the authors impose this assumption so that their method of argument can succeed. The assumption on the dimensions d_x and d_y is standard and is satisfied in every practical setting we can think of. Usually we have $d_y < d_x$, (often $d_y << d_x$) and sometimes $d_y = d_x$ (e.g., autoencoders).

The final assumption is perhaps the most interesting. The authors want to establish that $\mathcal{R}(\boldsymbol{C}) = \mathcal{R}(\boldsymbol{U}_{\mathcal{I}_{\bar{p}}})$. They find that $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}$, where $\boldsymbol{P}_{\boldsymbol{A}}$ denotes the matrix of orthogonal projection onto $\mathcal{R}(\boldsymbol{A})$ and $\boldsymbol{\Lambda}$ is the diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}$. From the assumption that the eigenvalues of $\boldsymbol{\Sigma}$ are distinct $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}$ implies $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}$ is a diagonal matrix. Right multiplying a matrix $\boldsymbol{M}$ by a diagonal matrix scales the columns of $\boldsymbol{M}$ by the diagonal entries, and left multiplying $\boldsymbol{M}$ by a diagonal matrix scales the rows of $\boldsymbol{M}$ by the diagonal entries. If these both yield the same matrix, then either the diagonal entries are not distinct or $\boldsymbol{M}$ is itself diagonal. After a few more steps this gives the conclusion the authors want.

But what happens if $\boldsymbol{\Sigma}$ does not have distinct eigenvalues? Suppose that the eigenvalues of $\boldsymbol{\Sigma}$ are $\lambda_1, \ldots, \lambda_r$, with corresponding geometric multiplicities $m_1, \ldots, m_r$. In this case, $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}$ is block-diagonal with blocks $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_r$ of dimensions $m_1, \ldots, m_r$. Since $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}$ is a projection, $\boldsymbol{P}^2_{\boldsymbol{U}^\intercal\boldsymbol{C}} = \boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}$ implying that for each i, $\boldsymbol{P}_i^2 = \boldsymbol{P}_i$. Since $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}}$ is an orthogonal projection, $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}} = \boldsymbol{P}^\intercal_{\boldsymbol{U}^\intercal\boldsymbol{C}}$ implying that $\boldsymbol{P}_i = \boldsymbol{P}_i^\intercal$. Thus, $\boldsymbol{P}_i$ is an orthogonal projector. We can eigendecompose each $\boldsymbol{P}_i = \boldsymbol{V}_i\tilde{\Sigma}\boldsymbol{V}_i^\intercal$. Since the $\boldsymbol{P}_i$'s are orthogonal projectors, their eigenvalues are 0 or 1, and we can write $\boldsymbol{P}_{\boldsymbol{U}^\intercal\boldsymbol{C}} = \boldsymbol{V}_{\mathcal{I}_{\bar{p}}}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}}^\intercal$ for some index set $\mathcal{I}_{\bar{p}}$. So, in this case we have

$$\boldsymbol{P}_C = \boldsymbol{U}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}}^\intercal\boldsymbol{U}^\intercal = \boldsymbol{U}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}}\left((\boldsymbol{U}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}})^\intercal\boldsymbol{U}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}}\right)^{-1}(\boldsymbol{U}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}})^\intercal.$$

Thus, we now get $\mathcal{R}(\boldsymbol{C}) = \mathcal{R}(\boldsymbol{U}\boldsymbol{V}_{\mathcal{I}_{\bar{p}}})$. What has changed? When $\boldsymbol{\Sigma}$'s eigenvalues are assumed distinct, $\mathcal{R}(\boldsymbol{C}) = \mathcal{R}(\boldsymbol{U})$ where $\boldsymbol{U}$ is a matrix of eigenvectors of $\boldsymbol{\Sigma}$ and so does not depend on $\boldsymbol{C}$. Unfortunately, $\boldsymbol{V}_{\mathcal{I}_{\bar{p}}}$ may depend on $\boldsymbol{C}$, so when the eigenvalues are not assumed distinct we no longer have a unique representation of $\mathcal{R}(\boldsymbol{C})$. This complicates the authors' arguments later on, and it is easy to see why they imposed the assumption to avoid those complications. However, it is not obvious that this is, in fact, a necessary condition for Theorem 2.3. It seems possible that a different method of argument could establish the result without the assumption on the eigenvalues of $\boldsymbol{\Sigma}$. For practical purposes, however, the restrictions on $\boldsymbol{\Sigma}$ are not especially problematic. With finite precision and noise, we can slightly disturb $\boldsymbol{X}, \boldsymbol{Y}$ to meet the condition required in this paper [1].

1.4 Limitations

A major limitation of this work is that the result in Theorem 2.3 applies only to linear neural networks. Although analyzing these more limited networks is a natural first step toward analyzing general nonlinear networks, it is not of great practical interest. Linear neural networks, to our knowledge are not used in practice. If we do not concern ourselves with keeping the weight matrices factored and multiply them together, we have an ordinary least-squares regression, for which there is simply a closed form solution.

Furthermore, this paper leaves open questions about how to avoid saddle points when optimizing deep linear neural networks. For very specific networks, we can appeal to Theorem 2.3 (iv). But for general linear networks, we have no theoretical assurance that our optimizers will not get trapped in saddle points. It does us no good that, in the problem setting, all local minima are global if we cannot distinguish the local minima from possibly highly non-optimal saddle points.

2 "Identity Matters in Deep Learning" [4]

2.1 Problem Statement

2.1.1 Challenges of deep networks

A common way to increase the representational capacity of a deep network is to increase its depth. However, effectively optimizing deep networks presents challenges.

First, as we discussed in the preceding review, optimizing the parameters of a neural network is a non-convex problem, and thus, convergence to a global optimum is not guaranteed and could well be intractable. Kawaguchi shows that *shallow linear* neural networks can effectively be optimized, which provides hope that shallow general neural networks can be as well. However, for the case of deep linear neural networks and the possibly analogous case of deep general neural networks, Kawaguchi's contribution leaves us at the mercy of saddle points.

Second, it is common practice to initialize the weights of a neural network by sampling values centered around 0, for instance, using a zero-mean Gaussian. While this is a simple scheme, the matrices could be quite "far" from the identity transformation. This could make converging to the identity transformation during training difficult. In this way, a deep network can result in important features identified by layers being lost in subsequent layers.

2.1.2 Residual networks

Residual layers have emerged as a popular approach in deep learning in recent years. If a typical feed-forward layer with weights $\boldsymbol{W}$ parameterizes a function $h(\boldsymbol{x})$, then a residual layer with the same weights represents the function $h(\boldsymbol{x}) + \boldsymbol{x}$. When the weights of the layer are 0, the layer represents the identity transformation. Using residual layers to create deeper neural networks has shown promise in many empirical settings [5].

2.1.3 This paper's contributions

The paper we review here, which we will refer to subsequently as Hardt and Ma, aims to explain the empirical success of residual layers theoretically. It builds on Kawaguchi's treatment of the optimization of linear neural networks, filling some of the gaps in understanding left by that paper. Kawaguchi's key result is that for linear neural networks of arbitrary depth, all local minima are global minima. This, however, leaves open the possibility of bad saddle points ensnaring optimizers and keeping them from finding the local minima that are also global. Hardt and Ma are able to show that, under some assumptions, all *critical points* of the loss are global minima, a considerably more desirable property than that found in Kawaguchi.

2.2 Main Result

2.2.1 Statement of Theorems 2.1 and 2.2

First, some notation. We aim to learn a linear function $\boldsymbol{R} : \mathbb{R}^d \to \mathbb{R}^d$. We assume access to a dataset $\{\boldsymbol{x}, \boldsymbol{y}\}_{i=1}^n$, where the inputs $\boldsymbol{x}$ are generated from a distribution $\mathcal{D}$, and the responses $\boldsymbol{y}$ are noisy measurements of $\boldsymbol{R}$ applied to $\boldsymbol{x}$: $\boldsymbol{y} = \boldsymbol{R}\boldsymbol{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. The hypothesis class is a linear residual network with l layers. This makes our prediction

$$\hat{y} = (\boldsymbol{I} + \boldsymbol{A}_l)(\boldsymbol{I} + \boldsymbol{A}_{l-1}) \dots (\boldsymbol{I} + \boldsymbol{A}_1)\boldsymbol{x},$$

where $\boldsymbol{A}_1, \dots, \boldsymbol{A}_l \in \mathbb{R}^{d \times d}$ are weight matrices and $\boldsymbol{A} \in \mathbb{R}^{l \times d \times d}$ is a tensor which is formed by stacking the matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_l$. We use the maximum likelihood loss function $l(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$. Sticking with the convention in Hardt and Ma, we overload the notation for f. $f(\boldsymbol{A})$ refers to the population risk whereas $f(\boldsymbol{A}, (\boldsymbol{x}, \boldsymbol{y}))$ is the empirical risk evaluated on $(\boldsymbol{x}, \boldsymbol{y})$. The learning objective is to minimize the population risk. This leads to the following optimization problem.

$$\arg\min_{\boldsymbol{A}} f(\boldsymbol{A}) = \arg\min_{\boldsymbol{A}} \mathbb{E}\left[f(\boldsymbol{A}, (\boldsymbol{x}, \boldsymbol{y}))\right] = \arg\min_{\boldsymbol{A}} \mathbb{E}\left[l(\hat{\boldsymbol{y}}, \boldsymbol{y})\right]. \tag{1}$$

Additionally, we define $|||\boldsymbol{A}||| = \max_{1\leq i\leq l} \|\boldsymbol{A}_i\|$, where $\|\boldsymbol{A}_i\|$ refers to the spectral norm. Finally, let

$$\gamma = \max\{|\log \sigma_{\max}(\boldsymbol{R})|, |\log \sigma_{\min}(\boldsymbol{R})|\},$$

where $\sigma_{\max}(\boldsymbol{R})$ denotes the maximum singular value of $\boldsymbol{R}$ and $\sigma_{\min}(\boldsymbol{R})$ denotes the minimum singular value of $\boldsymbol{R}$. We can now state Theorem 2.1, which establishes an upper bound on the norm of the optimal solution to (1).

Theorem 2.1. *Suppose that $l \geq 3\gamma$ and $det(\boldsymbol{R}) > 0$. Then there exists a globally optimal solution $\boldsymbol{A}^*$ to (1) with*

$$|||\boldsymbol{A}^*||| \leq (4\pi + 3\gamma)/l.$$

Now, define $\mathcal{B}_\tau = \{\boldsymbol{A} : |||\boldsymbol{A}||| \leq \tau\}$, and we can state Theorem 2.2.

Theorem 2.2. *For $\tau < 1$, any critical point $\boldsymbol{A}$ of the objective $f(\cdot)$ restricted to the domain $\mathcal{B}_\tau$ is a global minimum.*

2.2.2 Discussion

Theorem 2.1 establishes that the norm of the optimal solution to the objective in (1) is inversely proportional to the number of layers of the network. This is significant since it allows us to argue that for deep neural networks, Theorem 2.2 becomes relevant. Why? Theorem 2.1 says that we can limit our search for an optimal $\boldsymbol{A^*}$ to $\mathcal{B}_{\tau^*}$ where $\tau^* = (4\pi + 3\gamma)/l$. Moreover, from Theorem 2.1, with $\boldsymbol{R}$ fixed, τ^* is on the order of $1/l$. Thus, for deep networks, τ^* can be made smaller than 1. Then, once $\tau^* < 1$, Theorems 2.1 and 2.2 jointly imply that any critical point of f restricted to the domain $\mathcal{B}_{\tau^*}$ is a global minimum of f on its unrestricted domain. Thus, for sufficiently deep linear residual neural networks, we need only travel toward and converge to a critical point, a task that standard optimization techniques are usually well equipped to accomplish. (This assumes that our optimizer starts in $\mathcal{B}_{\tau^*}$ and stays there.) This is a major improvement over Kawaguchi, where the key result establishes that converging to any local minima is optimal.

2.3 Examples

In this section, we take a closer look at Theorem 2.1 in a simplified setting. We see how the theorem fails when its assumptions are not satisfied on two counterexamples. Then, motivated by our inability to find a counterexample where the theorem suggests there should be one, we prove a substantially strengthened version of Theorem 2.1 for our basic setting.

Before moving to this regime, we establish a fact that will be critical later.

Claim 2.3.1. *If the distribution $\mathcal{D}$ of $\boldsymbol{x}$ is nontrivial, in the sense that it places nonzero probability on the event $\boldsymbol{x} \neq \vec{\boldsymbol{0}}$, $\boldsymbol{A} = \{\boldsymbol{A_1}, \ldots, \boldsymbol{A_l}\}$ is a global minimum of the population risk $f(\cdot)$ if and only if*

$$(\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1}) = \boldsymbol{R}.$$

Proof.

$$\begin{aligned}
f(\boldsymbol{A}) &= \mathbb{E}[f(\boldsymbol{A}, (\boldsymbol{x}, \boldsymbol{y}))] \\
&= \mathbb{E}\left[\|(\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1})\boldsymbol{x} - \boldsymbol{R}\boldsymbol{x} - \boldsymbol{\xi}\|^2\right] \\
&= \mathbb{E}\left[\|((\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1}) - \boldsymbol{R})\,\boldsymbol{x} - \boldsymbol{\xi}\|^2\right] \\
&= \mathbb{E}[\langle((\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1}) - \boldsymbol{R})\,\boldsymbol{x}, ((\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1}) - \boldsymbol{R})\,\boldsymbol{x}\rangle] \\
&\qquad - 2\,\mathbb{E}[\langle((\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1}) - \boldsymbol{R})\,\boldsymbol{x}, \boldsymbol{\xi}\rangle] + \mathbb{E}[\langle\boldsymbol{\xi}, \boldsymbol{\xi}\rangle].
\end{aligned}$$

Since $\langle\cdot,\cdot\rangle$ is an inner product on $\mathbb{R}^d \times \mathbb{R}^d$, there exists a symmetric positive definite $\boldsymbol{M}$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\langle\boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^\intercal \boldsymbol{M}\boldsymbol{y}$. Thus, for some symmetric positive definite $\boldsymbol{M}$,

$$\mathbb{E}[\langle((\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1}) - \boldsymbol{R})\,\boldsymbol{x}, \boldsymbol{\xi}\rangle] = \mathbb{E}[(((\boldsymbol{I} + \boldsymbol{A_l}) \cdots (\boldsymbol{I} + \boldsymbol{A_1}) - \boldsymbol{R})\,\boldsymbol{x})^\intercal\, \boldsymbol{M}\xi] = 0,$$

since $\boldsymbol{x}$ and $\boldsymbol{\xi}$ are independent and $\boldsymbol{\xi}$ is mean zero. Then,

$$f(\boldsymbol{A}) = \mathbb{E}\left[\left\|\left((\boldsymbol{I}+\boldsymbol{A_l})\cdots(\boldsymbol{I}+\boldsymbol{A_1})-\boldsymbol{R}\right)\boldsymbol{x}\right\|^2\right] + \mathbb{E}\left[\|\boldsymbol{\xi}\|^2\right].$$

A norm is non-negative and the expectation of a non-negative random variable is non-negative, so

$$f(\boldsymbol{A}) \geq \mathbb{E}\left[\|\boldsymbol{\xi}\|^2\right].$$

If $(\boldsymbol{I}+\boldsymbol{A_l})\cdots(\boldsymbol{I}+\boldsymbol{A_1}) = \boldsymbol{R}$,

$$f(\boldsymbol{A}) = \mathbb{E}\left[\|\boldsymbol{\xi}\|^2\right], \tag{2}$$

so $\boldsymbol{A}$ is a global minimum of $f(\cdot)$. Now suppose condition 2 prevails. Then, with probability 1

$$(\boldsymbol{I}+\boldsymbol{A_l})\cdots(\boldsymbol{I}+\boldsymbol{A_1})\boldsymbol{x} = \boldsymbol{R}\boldsymbol{x}. \tag{3}$$

Since $\det(\boldsymbol{R}) \neq 0$, $\boldsymbol{R}$ is invertible and we can write condition 3 as $\boldsymbol{R}^{-1}(\boldsymbol{I}+\boldsymbol{A_l})\cdots(\boldsymbol{I}+\boldsymbol{A_1})\boldsymbol{x} = \boldsymbol{x}$. Since $\mathcal{D}$ is nontrivial, this implies $\boldsymbol{R}^{-1}(\boldsymbol{I}+\boldsymbol{A_l})\cdots(\boldsymbol{I}+\boldsymbol{A_1}) = \boldsymbol{I}$, and $(\boldsymbol{I}+\boldsymbol{A_l})\cdots(\boldsymbol{I}+\boldsymbol{A_1})\boldsymbol{R}^{-1} = \boldsymbol{I}$. A matrix inverse is unique, so $(\boldsymbol{I}+\boldsymbol{A_l})\cdots(\boldsymbol{I}+\boldsymbol{A_1}) = \boldsymbol{R}$. □

Now we move to the setting of $l = 1$. Under this regime, Claim 2.3.1 leads to the following simplification of Theorem 2.1.

Theorem 2.1 ($l = 1$)**.** *Suppose $\gamma \leq 1/3$ and $det(\boldsymbol{R}) > 0$. Then, the unique global optimum solution $\boldsymbol{A}^* = \{\boldsymbol{R}-\boldsymbol{I}\}$ of the population risk $f(\cdot)$ has norm*[2]

$$|||\boldsymbol{A}^*||| = \|\boldsymbol{R}-\boldsymbol{I}\| \leq 4\pi + 3\gamma.$$

This formulation will help us see counterexamples to Theorem 2.1.

Counterexample 2.3.1. Consider $\boldsymbol{R} = \begin{bmatrix}-21 & -19\\ -19 & -21\end{bmatrix}$.

$$\boldsymbol{R} = \begin{bmatrix}-\sqrt{2}/2 & -\sqrt{2}/2\\ -\sqrt{2}/2 & \sqrt{2}/2\end{bmatrix}\begin{bmatrix}-40 & 0\\ 0 & -2\end{bmatrix}\begin{bmatrix}-\sqrt{2}/2 & -\sqrt{2}/2\\ -\sqrt{2}/2 & \sqrt{2}/2\end{bmatrix}, \quad \text{implying} \quad \sigma_{\max}(\boldsymbol{R}) = 40, \quad \sigma_{\min}(\boldsymbol{R}) = 2.$$

$$\gamma := \max\{|\log\sigma_{\max}(R)|, |\log\sigma_{\min}(R)|\} = \max\{\log 40, \log 2\} = \log 40 \approx 3.69 > 1/3, \quad \det(\boldsymbol{R}) = 80 > 0.$$

We have γ noncompliant with Theorem 2.1 and $\det(\boldsymbol{R})$ compliant. Moreover $\sigma_{\max}(\boldsymbol{R}-\boldsymbol{I}) = 41$ (for orthogonal $\boldsymbol{U}$, $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal - \boldsymbol{I} = \boldsymbol{U}(\boldsymbol{\Sigma}-\boldsymbol{I})\boldsymbol{U}^\intercal$). Thus,

$$|||\boldsymbol{A}^*||| = \|\boldsymbol{R}-\boldsymbol{I}\| = 41 > 23.63 \approx 4\pi + 3\gamma.$$

Counterexample 2.3.2. Consider $\boldsymbol{R} = \begin{bmatrix}-19 & -21\\ -21 & -19\end{bmatrix}$.

$$\boldsymbol{R} = \begin{bmatrix}-\sqrt{2}/2 & -\sqrt{2}/2\\ -\sqrt{2}/2 & \sqrt{2}/2\end{bmatrix}\begin{bmatrix}-40 & 0\\ 0 & 2\end{bmatrix}\begin{bmatrix}-\sqrt{2}/2 & -\sqrt{2}/2\\ -\sqrt{2}/2 & \sqrt{2}/2\end{bmatrix}, \quad \text{implying} \quad \sigma_{\max}(\boldsymbol{R}) = 40, \quad \sigma_{\min}(\boldsymbol{R}) = 2.$$

$$\gamma := \max\{|\log\sigma_{\max}(R)|, |\log\sigma_{\min}(R)|\} = \max\{\log 40, \log 2\} = \log 40 \approx 3.69 > 1/3, \quad \det(\boldsymbol{R}) = -80 < 0.$$

We have both γ and $\det(\boldsymbol{R})$ noncompliant with Theorem 2.1. Moreover $\sigma_{\max}(\boldsymbol{R}-\boldsymbol{I}) = 41$. Thus,

$$|||\boldsymbol{A}^*||| = \|\boldsymbol{R}-\boldsymbol{I}\| = 41 > 23.63 \approx 4\pi + 3\gamma.$$

[2]All matrix norms are the spectral norm.

We have found counterexamples to Theorem 2.1 in the simplified setting $l = 1$ by choosing γ noncompliant as well as both γ and $\det(\boldsymbol{R})$ noncompliant. What about just $\det(\boldsymbol{R})$ noncompliant? Can we find a counterexample when $\det(\boldsymbol{R}) \leq 0$ and $\gamma \leq 1/3$? The answer is no! It turns out that for $l = 1$, the condition on $\det(\boldsymbol{R})$ is obsolete and we can prove a stronger version of Theorem 2.1. Moreover, our proof will not rely on any results presented in the paper. The only fact that we will need from outside of our review is a well-known singular values inequality. Whereas the proof of Theorem 2.1 in the general case is one of the more technically demanding proofs in the paper, for $l = 1$, it turns out to be entirely straightforward.

Strengthened Theorem 2.1 ($l = 1$)**.** *Suppose $\gamma \leq 1/3$. Then, the unique global optimum solution $\boldsymbol{A}^* = \{\boldsymbol{R} - \boldsymbol{I}\}$ of the population risk $f(\cdot)$ has norm*

$$|||\boldsymbol{A}^*||| = \|\boldsymbol{R} - \boldsymbol{I}\| \leq e^{1/3} + 1.$$

Proof. If $\gamma \leq 1/3$, then

$$\begin{aligned} -1/3 \leq \log \sigma_{\max}(\boldsymbol{R}) \leq 1/3 \\ e^{-1/3} \leq \sigma_{\max}(\boldsymbol{R}) \leq e^{1/3}. \end{aligned}$$

Now by the Weyl inequality,

$$\sigma_{\max}(\boldsymbol{R} - \boldsymbol{I}) \leq \sigma_{\max}(\boldsymbol{R}) + \sigma_{\max}(-\boldsymbol{I}) = \sigma_{\max}(\boldsymbol{R}) + 1 \leq e^{1/3} + 1$$

To establish that this is a (considerably) tighter bound than that in Theorem 2.1, we merely observe $\gamma > 0$, so

$$e^{1/3} + 1 \approx 1.91 < 4\pi < 4\pi + 3\gamma.$$

□

2.4 Limitations

We consider four limitations of Hardt and Ma. First, and most obviously, the setting of the paper is unrealistic. We need the residual neural network to be linear and the underlying function that describes the process being modeled to be linear. In practice, however, if we found ourselves in the linear setting, we simply would not use a residual neural network. These networks thrive in nonlinear and highly nonlinear settings. Second, Hardt and Ma are characterizing the landscape of the population risk. In practice, we must optimize the empirical risk. For sufficiently large samples and appropriate distributions, these optimization problems likely resemble one another. However, it is not inconceivable that in certain settings, perhaps with adversarial distributions, the surface of the empirical loss fails to resemble that of the population loss in important ways that break the optimization guarantees in the paper. Third, our nice result about critical points being local optima holds only on $\mathcal{B}_\tau$ for $\tau < 1$. Even if $\tau^* < 1$, we need our optimizer to stay in $\mathcal{B}_\tau$. Weight matrices are commonly initialized around 0, but for τ sufficiently small, we could initialize $\boldsymbol{A}$ outside of $\mathcal{B}_\tau$. Or, if we initialized inside of $\mathcal{B}_\tau$, iterations could leave this ball. And if we do leave this ball, the landscape could be treacherous and our optimizer could fall victim to a bad saddle point. Finally, the paper's results apply for learning when $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. We find ourselves in this regime with autoencoders. The case of $d_y < d_x$, however, is also of great interest. Extending these results to that setting requires mixing identity and non-identity layers and is not treated by the paper.

References

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

[2] Pierre Baldi and Zhiqin Lu. Complex-Valued Autoencoders. *Neural Networks*, 33:136–147, September 2012. arXiv: 1108.4135.

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

[4] Moritz Hardt and Tengyu Ma. Identity Matters in Deep Learning. *arXiv:1611.04231 [cs, stat]*, July 2018. arXiv: 1611.04231.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[6] Kenji Kawaguchi. Deep Learning without Poor Local Minima. *arXiv:1605.07110 [cs, math, stat]*, December 2016. arXiv: 1605.07110.

[7] Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu. Local minima in training of neural networks, 2017.