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: $W_{H+1} \in \mathbb{R}^{d_y \times d_H}, \dots, W_k \in \mathbb{R}^{d_k \times d_{k-1}}, \dots, W_1 \in \mathbb{R}^{d_1 \times d_x}$. 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 $p = \min(d_H, \dots, d_1)$ 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., $\overline{Y}(W, X) = W_2 W_1 X$), 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 $k \in \{1, \dots, H+1\}$, 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 $d_y, d_H, d_{H-1}, \dots, d_1, d_x \geq 1$ (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 $\operatorname{rank}(W_H \cdots W_2) = p$, 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, $\sigma_k : \mathbb{R}^{d_k \times m} \to \mathbb{R}^{d_k \times m}$ is the element-wise rectified linear function:

$$\sigma_k\left(\begin{bmatrix} b_{11} & \dots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{d_k1} & \cdots & b_{d_km} \end{bmatrix}\right) = \begin{bmatrix} \bar{\sigma}(b_{11}) & \dots & \bar{\sigma}(b_{1m)} \\ \vdots & \ddots & \vdots \\ \bar{\sigma}(b_{d_k1}) & \cdots & \bar{\sigma}(b_{d_km}) \end{bmatrix},$$

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 $\mathcal{D}_{\mathrm{vec}(W_k^T)} f(\cdot) = \frac{\partial f(\cdot)}{\partial_{\mathrm{vec}(W_k^T)}}$ be the partial derivative of f with respect to $\mathrm{vec}(W_k^T)$ in the numerator layout. That is, if $f : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$, we have $\mathcal{D}_{\mathrm{vec}(W_k^T)} f(\cdot) \in \mathbb{R}^{d_{out} \times (d_k d_{k-1})}$. 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 $\bar{p} = \mathrm{rank}(C) \in \{1, \ldots, \min(d_y, p)\}$. 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 $k \in \{1, ..., H + 1\}$,*

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

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*

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

[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 $k \in \{2, ..., H+1\}$,*

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

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

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

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

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

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

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

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 $W_{H+1} \cdots W_1 = U_p U_p^T Y X^T (XX^T)^{-1}$, 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,

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

By using Lemma 4.6 to obtain an expression of C, we deduce that we can have $\text{rank}(W_1) \geq \min(p, d_y)$ 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 $\text{rank}(W_k W_{k-1} \cdots W_1) \geq \min(p, d_x)$ 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 $\text{rank}(C) \geq \text{rank}(W_{H+1} \cdots W_1) \geq \min(p, d_x)$ 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 $\text{rank}(C) \geq \min(p, d_y)$, 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 $\text{rank}(W_H \cdots W_2) = p$, 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.

Deep Learning without Poor Local Minima Appendix

A Proofs of lemmas and corollary in Section 4.2

We complete the proofs of the lemmas and corollary in Section 4.2.

A.1 Proof of Lemma 4.1

Proof Since $\bar{\mathcal{L}}(W) = \frac{1}{2}\|\overline{Y}(W,X) - Y\|_F^2 = \frac{1}{2}\operatorname{vec}(r)^T \operatorname{vec}(r)$,

$$\begin{aligned}
\mathcal{D}_{\operatorname{vec}(W_k^T)}\bar{\mathcal{L}}(W) &= \left(\mathcal{D}_{\operatorname{vec}(r)}\bar{\mathcal{L}}(W)\right)\left(\mathcal{D}_{\operatorname{vec}(W_k^T)}\operatorname{vec}(r)\right)\\
&= \operatorname{vec}(r)^T\left(\mathcal{D}_{\operatorname{vec}(W_k^T)}\operatorname{vec}(X^T I_{d_x} W_1^T\cdots W_{H+1}^T I_{d_y}) - \mathcal{D}_{\operatorname{vec}(W_k^T)}\operatorname{vec}(Y^T)\right)\\
&= \operatorname{vec}(r)^T\left(\mathcal{D}_{\operatorname{vec}(W_k^T)}(W_{H+1}\cdots W_{k+1}\otimes (W_{k-1}\cdots W_1 X)^T)\operatorname{vec}(W_k^T)\right)\\
&= \operatorname{vec}(r)^T\left(W_{H+1}\cdots W_{k+1}\otimes (W_{k-1}\cdots W_1 X)^T\right).
\end{aligned}$$

By setting $\left(\mathcal{D}_{\operatorname{vec}(W_k^T)}\bar{\mathcal{L}}(W)\right)^T = 0$ for all $k \in \{1, ..., H+1\}$, we obtain the statement of Lemma 4.1. For the boundary cases (i.e., $k = H+1$ or $k = 1$), it can be seen from the second to the third lines that we obtain the desired results with the definition, $W_k \cdots W_{k+1} \triangleq I_{d_k}$ (i.e., $W_{H+1}\cdots W_{H+2} \triangleq I_{d_y}$ and $W_0 \cdots W_1 \triangleq I_{d_x}$). □

A.2 Proof of Lemma 4.2

Proof From the critical point condition with respect to W_1 (Lemma 4.1),

$$0 = \left(\mathcal{D}_{\operatorname{vec}(W_k^T)}\bar{\mathcal{L}}(W)\right)^T = \left(W_{H+1}\cdots W_2 \otimes X^T\right)^T \operatorname{vec}(r) = \operatorname{vec}(XrW_{H+1}\cdots W_2),$$

which is true if and only if $XrW_{H+1}\cdots W_2 = 0$. By expanding r, $0 = XX^TW_1^TC^TC - XY^TC$. By solving for W_1,

$$W_1 = (C^TC)^- C^TYX^T(XX^T)^{-1} + (I - (C^TC)^- C^TC)L, \tag{2}$$

for an arbitrary matrix L. Due to the property of any generalized inverse (Zhang, 2006, p. 41), we have that $C(C^TC)^-C^TC = C$. Thus,

$$CW_1 = C(C^TC)^-C^TYX^T(XX^T)^{-1} + (C - C(C^TC)^-C^TC)L = C(C^TC)^-C^TYX^T(XX^T)^{-1}.$$

□

A.3 Proof of Lemma 4.3

Proof For the diagonal blocks: the entries of diagonal blocks are obtained simply using the result of Lemma 4.1 as

$$\mathcal{D}_{\operatorname{vec}(W_k^T)}\left(\mathcal{D}_{\operatorname{vec}(W_k^T)}\bar{\mathcal{L}}(W)\right)^T = \left(W_{H+1}\cdots W_{k+1}\otimes (W_{k-1}\cdots W_1X)^T\right)^T \mathcal{D}_{\operatorname{vec}(W_k^T)}\operatorname{vec}(r).$$

Using the formula of $\mathcal{D}_{\operatorname{vec}(W_k^T)}\operatorname{vec}(r)$ computed in the proof of of Lemma 4.1 yields the desired result.

For the off-diagonal blocks with $k = 2, ..., H$:

$$
\begin{aligned}
&\mathcal{D}_{\mathrm{vec}(W_k^T)}[\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W)]^T \\
&= \big(W_{H+1}\cdots W_2 \otimes X)^T\big)^T \mathcal{D}_{\mathrm{vec}(W_k^T)}\,\mathrm{vec}(r) + \left(\mathcal{D}_{\mathrm{vec}(W_k^T)} W_{H+1}\cdots W_{k+1} \otimes X^T\right)^T \mathrm{vec}(r)
\end{aligned}
$$

The first term above is reduced to the first term of the statement in the same way as the diagonal blocks. For the second term,

$$
\begin{aligned}
&\left(\mathcal{D}_{\mathrm{vec}(W_k^T)} W_{H+1}\cdots W_2 \otimes X^T\right)^T \mathrm{vec}(r) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{d_y} \left(\left(\mathcal{D}_{\mathrm{vec}(W_k^T)} W_{H+1,j} W_H \cdots W_2\right) \otimes X_i^T\right)^T r_{i,j} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{d_y} \left((A_k)_{j,\cdot} \otimes B_k^T \otimes X_i^T\right)^T r_{i,j} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{d_y} \left[(A_k)_{j,1}\left(B_k^T \otimes X_i\right) \quad \dots \quad (A_k)_{j,d_k}\left(B_k^T \otimes X_i\right)\right] r_{i,j} \\
&= \left[\left(B_k^T \otimes \textstyle\sum_{i=1}^{m}\sum_{j=1}^{d_y} r_{i,j}(A_k)_{j,1} X_i\right) \quad \dots \quad \left(B_k^T \otimes \textstyle\sum_{i=1}^{m}\sum_{j=1}^{d_y} r_{i,j}(A_k)_{j,d_k} X_i\right)\right].
\end{aligned}
$$

where $A_k = W_{H+1}\cdots W_{k+1}$ and $B_k = W_{k-1}\cdots W_2$. The third line follows the fact that $(W_{H+1,j}W_H\cdots W_2)^T = \mathrm{vec}(W_2^T\cdots W_H^T W_{H+1,j}^T) = (W_{H+1,j}\cdots W_{k+1} \otimes W_2^T\cdots W_{k-1}^T)\,\mathrm{vec}(W_k^T)$. In the last line, we have the desired result by rewriting $\sum_{i=1}^{m}\sum_{j=1}^{d_y} r_{i,j}(A_k)_{j,t}X_i = X(rW_{H+1}\cdots W_{k+1})_{\cdot,t}$.

For the off-diagonal blocks with $k = H+1$: The first term in the statement is obtained in the same way as above (for the off-diagonal blocks with $k = 2, ..., H$). For the second term, notice that $\mathrm{vec}(W_{H+1}^T) = \left[(W_{H+1})_{1,\cdot}^T \quad \dots \quad (W_{H+1})_{d_y,\cdot}^T\right]^T$ where $(W_{H+1})_{j,\cdot}$ is the j-th row vector of W_{H+1} or the vector corresponding to the j-th output component. That is, it is conveniently organized as the blocks, each of which corresponds to each output component (or rather we chose $\mathrm{vec}(W_k^T)$ instead of $\mathrm{vec}(W_k)$ for this reason, among others). Also,

$$
\begin{aligned}
&\left(\mathcal{D}_{\mathrm{vec}(W_{H+1}^T)} W_{H+1}\cdots W_2 \otimes X^T\right)^T \mathrm{vec}(r) = \\
&= \left[\textstyle\sum_{i=1}^{m}\left(\left(\mathcal{D}_{(W_{H+1})_{1,\cdot}^T} C_{1,\cdot}\right)\otimes X_i^T\right)^T r_{i,1} \quad \dots \quad \textstyle\sum_{i=1}^{m}\left(\left(\mathcal{D}_{(W_{H+1})_{d_y,\cdot}^T} C_{d_y,\cdot}\right)\otimes X_i^T\right)^T r_{i,d_y}\right],
\end{aligned}
$$

where we also used the fact that

$$
\sum_{i=1}^{m}\sum_{j=1}^{d_y}\left(\left(\mathcal{D}_{\mathrm{vec}((W_{H+1})_{t,\cdot}^T)} C_{j,\cdot}\right)\otimes X_i^T\right)^T r_{i,j} = \sum_{i=1}^{m}\left(\left(\mathcal{D}_{\mathrm{vec}((W_{H+1})_{t,\cdot}^T)} C_{t,\cdot}\right)\otimes X_i^T\right)^T r_{i,t}.
$$

For each block entry $t = 1, \dots, d_y$ in the above, similarly to the case of $k = 2, ..., H$,

$$
\sum_{i=1}^{m}\left(\left(\mathcal{D}_{\mathrm{vec}((W_{H+1})_{t,\cdot}^T)} C_{j,\cdot}\right)\otimes X_i^T\right)^T r_{i,t} = \left(B_{H+1}^T \otimes \sum_{i=1}^{m} r_{i,t}(A_{H+1})_{j,t} X_i\right).
$$

Here, we have the desired result by rewriting $\sum_{i=1}^{m} r_{i,t}(A_{H+1})_{j,1}X_i = X(rI_{d_y})_{\cdot,t} = Xr_{\cdot,t}$. □

A.4 Proof of Lemma 4.4

Proof Note that a similarity transformation preserves the eigenvalues of a matrix. For each $k \in \{2, \dots, H+1\}$, we take a similarity transform of $\nabla^2\bar{\mathcal{L}}(W)$ (whose entries are organized as in Lemma 4.3) as

$$
P_k^{-1}\nabla^2\bar{\mathcal{L}}(W)P_k = \begin{bmatrix}
\mathcal{D}_{\mathrm{vec}(W_1^T)}\left(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W)\right)^T & \mathcal{D}_{\mathrm{vec}(W_k^T)}\left(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W)\right)^T & \cdots \\
\mathcal{D}_{\mathrm{vec}(W_1^T)}\left(\mathcal{D}_{\mathrm{vec}(W_k^T)}\bar{\mathcal{L}}(W)\right)^T & \mathcal{D}_{\mathrm{vec}(W_k^T)}\left(\mathcal{D}_{\mathrm{vec}(W_k^T)}\bar{\mathcal{L}}(W)\right)^T & \cdots \\
\vdots & \vdots & \ddots
\end{bmatrix}
$$

Here, $P_k = \begin{bmatrix} \mathbf{e}_{H+1} & \mathbf{e}_k & \tilde{P}_k \end{bmatrix}$ is the permutation matrix where $\mathbf{e}_i$ is the i-th element of the standard basis (i.e., a column vector with 1 in the i-th entry and 0 in every other entries), and $\tilde{P}_k$ is any arbitrarily matrix that makes P_k to be a permutation matrix. Let M_k be the principal submatrix of $P_k^{-1}\nabla^2\bar{\mathcal{L}}(W)P_k$ that consists of the first four blocks appearing in the above equation. Then,

$$\begin{aligned}
&\nabla^2\bar{\mathcal{L}}(W) \succeq 0 \\
&\Rightarrow \forall k \in \{2, \dots, H+1\}, M_k \succeq 0 \\
&\Rightarrow \forall k \in \{2, \dots, H+1\}, \mathcal{R}(\mathcal{D}_{\mathrm{vec}(W_k^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T) \subseteq \mathcal{R}(\mathcal{D}_{\mathrm{vec}(W_1^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T),
\end{aligned}$$

Here, the first implication follows the necessary condition with any principal submatrix and the second implication follows the necessary condition with the Schur complement (Zhang, 2006, theorem 1.20, p. 44).

Note that $\mathcal{R}(M') \subseteq \mathcal{R}(M) \Leftrightarrow (I - MM^-)M' = 0$ (Zhang, 2006, p. 41). Thus, by plugging in the formulas of $\mathcal{D}_{\mathrm{vec}(W_k^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T$ and $\mathcal{D}_{\mathrm{vec}(W_1^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T$ that are derived in Lemma 4.3, $\nabla^2\bar{\mathcal{L}}(W) \succeq 0 \Rightarrow \forall k \in \{2, \dots, H+1\}$,

$$\begin{aligned}
0 = &\left(I - (C^TC \otimes (XX^T))(C^TC \otimes (XX^T))^-\right)(C^TA_k \otimes B_kW_1X) \\
&+ \left(I - (C^TC \otimes (XX^T))(C^TC \otimes (XX^T))^-\right)[B_k^T \otimes X]\left[I_{d_{k-1}} \otimes (rA_k)_{\cdot,1} \;\dots\; I_{d_{k-1}} \otimes (rA_k)_{\cdot,d_k}\right]
\end{aligned}$$

where $A_k = W_{H+1}\cdots W_{k+1}$ and $B_k = W_{k-1}\cdots W_2$. Here, we can replace $(C^TC \otimes (XX^T))^-$ by $((C^TC)^- \otimes (XX^T)^{-1})$ (see Appendix A.7). Thus, $I - (C^TC \otimes (XX^T))(C^TC \otimes (XX^T))^-$can be replaced by $(I_{d_1} \otimes I_{d_y}) - (C^TC(C^TC)^- \otimes I_{d_y}) = (I_{d_1} - C^TC(C^TC)^-) \otimes I_{d_y}$. Accordingly, the first term is reduced to zero as

$$\left((I_{d_1} - C^TC(C^TC)^-) \otimes I_{d_y}\right)\left(C^TA_k \otimes B_kW_1X\right) = ((I_{d_1} - C^TC(C^TC)^-)C^TA_k) \otimes B_kW_1X = 0,$$

since $C^TC(C^TC)^-C^T = C^T$ (Zhang, 2006, p. 41). Thus, with the second term remained, the condition is reduced to

$$\forall k \in \{2, \dots, H+1\}, \forall t \in \{1, \dots, d_y\},\;\; (B_k^T - C^TC(C^TC)^-B_k^T) \otimes X(rA_k)_{\cdot,t} = 0.$$

This implies

$$\forall k \in \{2, \dots, H+1\},\;\; (R(B_k^T) \subseteq \mathcal{R}(C^TC) \;\;\text{or}\;\; XrA_k = 0),$$

which concludes the proof for the positive semidefinite case. For the necessary condition of the negative semidefinite case, we obtain the same condition since

$$\begin{aligned}
&\nabla^2\bar{\mathcal{L}}(W) \preceq\; 0 \\
&\Rightarrow \forall k \in \{2, \dots, H+1\}, M_k \preceq 0 \\
&\Rightarrow \forall k \in \{2, \dots, H+1\}, \mathcal{R}(-\mathcal{D}_{\mathrm{vec}(W_k^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T) \subseteq \mathcal{R}(-\mathcal{D}_{\mathrm{vec}(W_1^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T) \\
&\Rightarrow \forall k \in \{2, \dots, H+1\}, \mathcal{R}(\mathcal{D}_{\mathrm{vec}(W_k^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T) \subseteq \mathcal{R}(\mathcal{D}_{\mathrm{vec}(W_1^T)}(\mathcal{D}_{\mathrm{vec}(W_1^T)}\bar{\mathcal{L}}(W))^T).
\end{aligned}$$

□

A.5 Proof of Corollary 4.5

Proof From the first condition in the statement of Lemma 4.4,

$$\begin{aligned}
&\mathcal{R}(W_2^T \cdots W_{k-1}^T) \subseteq \mathcal{R}(W_2^T \cdots W_{H+1}^T W_{H+1} \cdots W_2) \\
&\Rightarrow \mathrm{rank}(W_k^T \cdots W_{H+1}^T) \geq \mathrm{rank}(W_2^T \cdots W_{k-1}^T) \Rightarrow \mathrm{rank}(W_{H+1} \cdots W_k) \geq \mathrm{rank}(W_{k-1} \cdots W_2).
\end{aligned}$$

The first implication follows the fact that the rank of a product of matrices is at most the minimum of the ranks of the matrices, and the fact that the column space of $W_2^T \cdots W_{H+1}^T$ is subspace of the column space of $W_2^T \cdots W_{k-1}^T$. □

A.6 Proof of Lemma 4.6

Proof For the $(Xr = 0)$ condition: Let M_{H+1} be the principal submatrix as defined in the proof of Lemma 4.4 (the principal submatrix of $P_{H+1}^{-1}\nabla^2\bar{\mathcal{L}}(W)P_{H+1}$ that consists of the first four blocks of it). Let $B_k = W_{k-1}\cdots W_2$. Let $F = B_{H+1}W_1XX^TW_1^TB_{H+1}^T$. Using Lemma 4.3 for the blocks corresponding to W_1 and W_{H+1},

$$M_{H+1} = \begin{bmatrix} C^TC \otimes XX^T & (C^T \otimes XX^T(B_{H+1}W_1)^T) + E \\ (C \otimes B_{H+1}W_1XX^T) + E^T & I_{d_y} \otimes F \end{bmatrix}$$

where $E = \begin{bmatrix} B_{H+1}^T \otimes Xr_{\cdot,1} & \dots & B_{H+1}^T \otimes Xr_{\cdot,d_y} \end{bmatrix}$. Then, by the necessary condition with the Schur complement (Zhang, 2006, theorem 1.20, p. 44), $M_{H+1} \succeq 0$ implies

$$\begin{aligned}
0 &= ((I_{d_y} \otimes I_{d_H}) - (I_{d_y} \otimes F)(I_{d_y} \otimes F)^-)((C \otimes B_{H+1}W_1XX^T) + E^T) \\
\Rightarrow 0 &= (I_{d_y} \otimes I_{d_H} - FF^-)(C \otimes B_{H+1}W_1XX^T) + (I_{d_y} \otimes I_{d_H} - FF^-)E^T \\
&= (I_{d_y} \otimes I_{d_H} - FF^-)E^T \\
&= \begin{bmatrix} I_{d_H} - FF^- \otimes I_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & I_{d_H} - FF^- \otimes I_1 \end{bmatrix} \begin{bmatrix} B_{H+1} \otimes (Xr_{\cdot,1})^T \\ \vdots \\ B_{H+1} \otimes (Xr_{\cdot,d_y})^T \end{bmatrix} \\
&= \begin{bmatrix} (I_{d_H} - FF^-)B_{H+1} \otimes (Xr_{\cdot,1})^T \\ \vdots \\ (I_{d_H} - FF^-)B_{H+1} \otimes (Xr_{\cdot,d_y})^T \end{bmatrix}
\end{aligned}$$

where the second line follows the fact that $(I_{d_y} \otimes F)^-$ can be replaced by $(I_{d_y} \otimes F^-)$ (see Appendix A.7). The third line follows the fact that $(I - FF^-)B_{H+1}W_1X = 0$ because $\mathcal{R}(B_{H+1}W_1X) = \mathcal{R}(B_{H+1}W_1XX^TW_1^TB_{H+1}^T) = \mathcal{R}(F)$. In the fourth line, we expanded E and used the definition of the Kronecker product. It implies

$$FF^-B_{H+1} = B_{H+1} \quad \text{or} \quad Xr = 0.$$

Here, if $Xr = 0$, we have obtained the statement of the lemma. Thus, from now on, we focus on the case where $FF^-B_{H+1} = B_{H+1}$ and $Xr \neq 0$ to obtain the other condition, $C(C^TC)^-C^T = U_{\bar{p}}U_{\bar{p}}$.

For the $(C(C^TC)^-C^T = U_{\bar{p}}U_{\bar{p}})$ condition: By using another necessary condition of a matrix being positive semidefinite with the Schur complement (Zhang, 2006, theorem 1.20, p. 44), $M_{H+1} \succeq 0$ implies that

$$(I_{d_y} \otimes F) - \left(C \otimes B_{H+1}W_1XX^T + E^T\right)(C^TC \otimes XX^T)^-\left(C^T \otimes XX^T(B_{H+1}W_1)^T + E\right) \succeq 0 \quad (3)$$

Since we can replace $(C^TC \otimes XX^T)^-$ by $(C^TC)^- \otimes (XX^T)^{-1}$ (see Appendix A.7), the second term in the left hand side is simplified as

$$\begin{aligned}
&\left(C \otimes B_{H+1}W_1XX^T + E^T\right)(C^TC \otimes XX^T)^-\left(C^T \otimes XX^T(B_{H+1}W_1)^T + E\right) \\
&= \left(\left(C(C^TC)^- \otimes B_{H+1}W_1\right) + E^T\left((C^TC)^- \otimes (XX^T)^{-1}\right)\right)\left(\left(C^T \otimes XX^T(B_{H+1}W_1)^T\right) + E\right) \\
&= \left(C(C^TC)^-C^T \otimes F\right) + E^T\left((C^TC)^- \otimes (XX^T)^{-1}\right)E \\
&= \left(C(C^TC)^-C^T \otimes F\right) + \left(r^TX^T(XX^T)^{-1}Xr \otimes B_{H+1}(C^TC)^-B_{H+1}\right) \qquad (4)
\end{aligned}$$

In the third line, the crossed terms – $\left(C(C^TC)^- \otimes B_{H+1}W_1\right)E$ and its transpose – are vanished to 0 because of the following. From Lemma 4.1, $\left(I_{d_y} \otimes (W_H\cdots W_1X)^T\right)^T \operatorname{vec}(r) = 0 \Leftrightarrow W_H\cdots W_1Xr = B_{H+1}W_1Xr = 0$ at any critical point. Thus, $\left(C(C^TC)^- \otimes B_{H+1}W_1\right)E = \begin{bmatrix} C(C^TC)^-B_{H+1}^T \otimes B_{H+1}W_1Xr_{\cdot,1} & \dots & C(C^TC)^-B_{H+1}^T \otimes B_{H+1}W_1Xr_{\cdot,d_y} \end{bmatrix} = 0$. The forth line follows

$$\begin{aligned}
&E^T \left((C^TC)^- \otimes (XX^T)^{-1}\right) E = \\
&\begin{bmatrix} B_{H+1}(C^TC)^-B_{H+1}^T \otimes (r_{\cdot,1})^T X^T(XX^T)^{-1}Xr_{\cdot,1} & \cdots & B_{H+1}(C^TC)^-B_{H+1}^T \otimes (r_{\cdot,1})^T X^T(XX^T)^{-1}Xr_{\cdot,d_y} \\ \vdots & \ddots & \vdots \\ B_{H+1}(C^TC)^-B_{H+1}^T \otimes (r_{\cdot,d_y})^T X^T(XX^T)^{-1}Xr_{\cdot,1} & \cdots & B_{H+1}(C^TC)^-B_{H+1}^T \otimes (r_{\cdot,d_y})^T X^T(XX^T)^{-1}Xr_{\cdot,d_y} \end{bmatrix} \\
&= r^TX^T(XX^T)^{-1}Xr \otimes B_{H+1}(C^TC)^-B_{H+1},
\end{aligned}$$

where the last line is due to the fact that $\forall t, (r_{\cdot,t})^T X^T(XX^T)^{-1}Xr_{\cdot,t}$ is a scalar and the fact that for any matrix L, $r^TLr = \begin{bmatrix} (r_{\cdot,1})^TLr_{\cdot,1} & \cdots & (r_{\cdot,1})^TLr_{\cdot,d_y} \\ \vdots & \ddots & \vdots \\ (r_{\cdot,d_y})^TLr_{\cdot,1} & \cdots & (r_{\cdot,d_y})^TLr_{\cdot,d_y} \end{bmatrix}$.

From equations 3 and 4, $M_{H+1} \succeq 0 \Rightarrow$

$$((I_{d_y} - C(C^TC)^-C^T) \otimes F) - \left(r^TX^T(XX^T)^{-1}Xr \otimes B_{H+1}(C^TC)^-B_{H+1}\right) \succeq 0. \tag{5}$$

In the following, we simplify equation 5 by first showing that $\mathcal{R}(C) = \mathcal{R}(U_{\mathcal{I}_{\bar{p}}})$ and then simplifying $r^TX^T(XX^T)^{-1}Xr, F$ and $B_{H+1}(C^TC)^-B_{H+1}$.

Showing that $\mathcal{R}(C) = \mathcal{R}(U_{\mathcal{I}_{\bar{p}}})$ (following the proof in Baldi & Hornik, 1989): Let $P_C = C(C^TC)^-C^T$ be the projection operator on $\mathcal{R}(C)$. We first show that $P_C\Sigma P_C = \Sigma P_C = P_C\Sigma$.

$$\begin{aligned}
P_C\Sigma P_C &= W_{H+1} \cdots W_1XX^TW_1^T \cdots W_{H+1}^T \\
&= YX^TW_1^T \cdots W_{H+1}^T \\
&= YX^T(XX^T)^{-1}XY^TP_C \\
&= \Sigma P_C,
\end{aligned}$$

where the first line follows Lemma 4.2, the second line is due to Lemma 4.1 with $k = H+1$ (i.e., $0 = W_H \cdots W_1Xr \Leftrightarrow W_{H+1} \cdots W_1XX^TW_1^T \cdots W_H^T = YX^TW_1^T \cdots W_H^T$), the third line follows Lemma 4.2, and the fourth line uses the definition of Σ. Since $P_C\Sigma P_C$ is symmetric, $\Sigma P_C (= P_C\Sigma P_C)$ is also symmetric and hence $\Sigma P_C = (\Sigma P_C)^T = P_C^T\Sigma^T = P_C\Sigma$. Thus, $P_C\Sigma P_C = \Sigma P_C = P_C\Sigma$. Note that $P_C = UP_{U^TC}U^T$ as $P_{U^TC} = U^TC(C^TUU^TC)^-C^TU = U^TP_CU$. Thus,

$$UP_{U^TC}U^TU\Lambda U^T = P_C\Sigma = \Sigma P_C = U\Lambda U^TUP_{U^TC}U^T,$$

which implies that $P_{U^TC}\Lambda = \Lambda P_{U^TC}$. Since the eigenvalues $(\Lambda_{1,1}, \dots, \Lambda_{d_y,d_y})$ are distinct, this implies that P_{U^TC} is a diagonal matrix (otherwise, $P_{U^TC}\Lambda = \Lambda P_{U^TC}$ implies $\Lambda_{i,i} = \Lambda_{j,j}$ for $i \neq j$, resulting in contradiction). Because P_{U^TC} is the orthogonal projector of rank $\bar{p}$ (as $P_{U^TC} = U^TP_CU$), this implies that P_{U^TC} is a diagonal matrix with its diagonal entries being ones ($\bar{p}$ times) and zeros ($dy - \bar{p}$ times). Thus,

$$C(C^TC)^-C^T = P_C = UP_{U^TC}U^T = U_{\mathcal{I}_{\bar{p}}}U_{\mathcal{I}_{\bar{p}}}^T,$$

for some index set $\mathcal{I}_{\bar{p}}$. This means that $\mathcal{R}(C) = \mathcal{R}(U_{\mathcal{I}_{\bar{p}}})$.

Simplifying $r^TX^T(XX^T)^{-1}Xr$:

$$\begin{aligned}
r^TX^T(XX^T)^{-1}Xr &= (CW_1X - Y)X^T(XX^T)^{-1}X(X^T(CW_1)^T - Y^T) \\
&= CW_1XX^T(CW_1)^T - CW_1XY^T - YX^T(CW_1)^T + \Sigma \\
&= P_C\Sigma P_C - P_C\Sigma - \Sigma P_C + \Sigma \\
&= \Sigma - U_{\bar{p}}\Lambda_{\mathcal{I}_{\bar{p}}}U_{\bar{p}}^T
\end{aligned}$$

where $P_C = C(C^TC)^-C^T = U_{\mathcal{I}_{\bar{p}}}U_{\mathcal{I}_{\bar{p}}}^T$ and the last line follows the facts:

$$P_C\Sigma P_C = U_{\mathcal{I}_{\bar{p}}}U_{\mathcal{I}_{\bar{p}}}^TU\Lambda U^TU_{\mathcal{I}_{\bar{p}}}U_{\mathcal{I}_{\bar{p}}}^T = U_{\mathcal{I}_{\bar{p}}}[I_{\bar{p}} \;\; 0] \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} & 0 \\ 0 & \Lambda_{-\mathcal{I}_{\bar{p}}} \end{bmatrix} \begin{bmatrix} I_{\bar{p}} \\ 0 \end{bmatrix} U_{\mathcal{I}_{\bar{p}}}^T = U_{\mathcal{I}_{\bar{p}}}\Lambda_{\mathcal{I}_{\bar{p}}}U_{\mathcal{I}_{\bar{p}}}^T,$$

$$P_C\Sigma = U_{\mathcal{I}_{\bar{p}}}U_{\mathcal{I}_{\bar{p}}}^TU\Lambda U^T = U_{\mathcal{I}_{\bar{p}}}[I_{\bar{p}} \;\; 0] \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} & 0 \\ 0 & \Lambda_{-\mathcal{I}_{\bar{p}}} \end{bmatrix} \begin{bmatrix} U_{\mathcal{I}_{\bar{p}}}^T \\ U_{-\mathcal{I}_{\bar{p}}} \end{bmatrix} = U_{\mathcal{I}_{\bar{p}}}^T\Lambda_{\mathcal{I}_{\bar{p}}}U_{\mathcal{I}_{\bar{p}}},$$

and similarly, $\Sigma P_C = U_{\mathcal{I}_{\bar{p}}}^T \Lambda_{\mathcal{I}_{\bar{p}}} U_{\mathcal{I}_{\bar{p}}}$.

Simplifying F: In the proof of Lemma 4.2, by using Lemma 4.1 with $k = 1$, we obtained that $W_1 = (C^T C)^- C^T Y X^T (X X^T)^{-1} + (I - (C^T C)^- C^T C)L$. Also, from Lemma 4.4, we have that $Xr = 0$ or $B_{H+1}(C^T C)^- C^T C = (C^T C (C^T C)^- B_{H+1}^T)^T = B_{H+1}$. If $Xr = 0$, we got the statement of the lemma, and so we consider the case of $B_{H+1}(C^T C)^- C^T C = B_{H+1}$. Therefore,

$$B_{H+1} W_1 = B_{H+1}(C^T C)^- C^T Y X^T (X X^T)^{-1}.$$

Since $F = B_{H+1} W_1 X X^T W_1^T B_{H+1}^T$,

$$F = B_{H+1}(C^T C)^- C^T \Sigma C (C^T C)^- B_{H+1}^T.$$

From Lemma 4.4 with $k = H + 1$, $\mathcal{R}(B_{H+1}^T) \subseteq \mathcal{R}(C^T C) = \mathcal{R}(B_{H+1}^T W_{H+1}^T W_{H+1} B_{H+1}) \subseteq \mathcal{R}(B_{H+1}^T)$, which implies that $\mathcal{R}(B_{H+1}^T) = \mathcal{R}(C^T C)$. Then, we have $\mathcal{R}(C(C^T C)^- B_{H+1}^T) = \mathcal{R}(C) = \mathcal{R}(U_{\mathcal{I}_{\bar{p}}})$. Accordingly, we can write it in the form, $C(C^T C)^- B_{H+1}^T = [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_2$, where $\mathbf{0} \in \mathbb{R}^{d_y \times (d_1 - \bar{p})}$ and $G_2 \in GL_{d_1}(\mathbb{R})$ (a $d_1 \times d_1$ invertible matrix). Thus,

$$F = G_2^T \begin{bmatrix} U_{\mathcal{I}_{\bar{p}}}^T \\ \mathbf{0} \end{bmatrix} U \Lambda U^T [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_2 = G_2^T \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \Lambda \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G_2 = G_2^T \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G_2.$$

Simplifying $B_{H+1}(C^T C)^- B_{H+1}$: From Lemma 4.4, $C^T C (C^T C)^- B_{H+1} = B_{H+1}$ (again since we are done if $Xr = 0$). Thus, $B_{H+1}(C^T C)^- B_{H+1} = B_{H+1}(C^T C)^- C^T C (C^T C)^- B_{H+1}^T$. As discussed above, we write $C(C^T C)^- B_{H+1}^T = [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_2$. Thus,

$$B_{H+1}(C^T C)^- B_{H+1} = G_2^T \begin{bmatrix} U_{\mathcal{I}_{\bar{p}}}^T \\ \mathbf{0} \end{bmatrix} [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_2 = G_2^T \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G_2.$$

Putting results together: We use the simplified formulas of $C(C^T C)^- C^T$, $r^T X^T (X X^T)^{-1} X r, F$ and $B_{H+1}(C^T C)^- B_{H+1}$ in equation 5, obtaining

$$((I_{d_y} - U_{\mathcal{I}_{\bar{p}}} U_{\mathcal{I}_{\bar{p}}}^T) \otimes G_2^T \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G_2) - \left((\Sigma - U_{\bar{p}} \Lambda_{\mathcal{I}_{\bar{p}}} U_{\bar{p}}^T) \otimes G_2^T \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G_2 \right) \succeq 0.$$

Due to Sylvester's law of inertia (Zhang, 2006, theorem 1.5, p. 27), with a nonsingular matrix $U \otimes G_2^{-1}$ (it is nonsingular because each of U and G_2^{-1} is nonsingular), the necessary condition is reduced to

$$\begin{aligned}
&\left(U \otimes G_2^{-1}\right)^T \left(\left((I_{d_y} - U_{\mathcal{I}_{\bar{p}}} U_{\mathcal{I}_{\bar{p}}}^T) \otimes G_2^T \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} \mathbf{0} \\ \mathbf{0}\ \mathbf{0} \end{bmatrix} G_2 \right) - \left((\Sigma - U_{\bar{p}} \Lambda_{\mathcal{I}_{\bar{p}}} U_{\bar{p}}^T) \otimes G_2^T \begin{bmatrix} I_{\bar{p}} \mathbf{0} \\ \mathbf{0}\,\mathbf{0} \end{bmatrix} G_2 \right) \right) \left(U \otimes G_2^{-1}\right) \\
&= \left(\left(I_{d_y} - \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \otimes \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) - \left(\left(\Lambda - \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{}p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \otimes \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \\
&= \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{(d_y - \bar{p})} \end{bmatrix} \otimes \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) - \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda_{-\mathcal{I}_{\bar{p}}} \end{bmatrix} \otimes \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \\
&= \left[\begin{array}{c|ccc} 0 & & 0 & \\ \hline & \Lambda_{\mathcal{I}_{\bar{p}}} - (\Lambda_{-\mathcal{I}_{\bar{p}}})_{1,1} I_{\bar{p}} & & 0 \\ 0 & & \ddots & \\ & 0 & & \Lambda_{\mathcal{I}_{\bar{p}}} - (\Lambda_{-\mathcal{I}_{\bar{p}}})_{(d_y - \bar{p}),(d_y - \bar{p})} I_{\bar{p}} \end{array} \right] \succeq 0,
\end{aligned}$$

which implies that for all $(i,j) \in \{(i,j) \mid i \in \{1, \ldots, \bar{p}\},\ j \in \{1, \ldots, (d_y - \bar{p})\}\}$, $(\Lambda_{\mathcal{I}_{\bar{p}}})_{i,i} \geq (\Lambda_{-\mathcal{I}_{\bar{p}}})_{j,j}$. In other words, the index set $\mathcal{I}_{\bar{p}}$ must select the largest $\bar{p}$ eigenvalues whatever $\bar{p}$ is. Since $C(C^T C)^- C^T = U_{\mathcal{I}_{\bar{p}}} U_{\mathcal{I}_{\bar{p}}}^T$ (which is obtained above), we have that $C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}$ in this case.

Summarizing the above case analysis, if $\nabla^2 \bar{\mathcal{L}}(W) \succeq 0$ at a critical point, $C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}$ or $Xr = 0$. □

A.7 Generalized inverse of Kronecker product

$(A^- \otimes B^-)$ is a generalized inverse of $A \otimes B$.

Proof For a matrix M, the definition of a generalized inverse, M^-, is $MM^-M = M$. Setting $M := A \otimes B$, we check if $(A^- \otimes B^-)$ satisfies the definition: $(A \otimes B)(A^- \otimes B^-)(A \otimes B) = (AA^-A \otimes BB^-B) = (A \otimes B)$ as desired. □

Here, we are *not* claiming that $(A^- \otimes B^-)$ is the unique generalized inverse of $A \otimes B$. Notice that the necessary condition that we have in our proof (where we need a generalized inverse of $A \otimes B$) is for any generalized inverse of $A \otimes B$. Thus, replacing it by one of any generalized inverse suffices to obtain a necessary condition. Indeed, choosing Moore−Penrose pseudoinverse suffices here, with which we know $(A \otimes B)^\dagger = (A^\dagger \otimes B^\dagger)$. But, to give a simpler argument later, we keep more generality by choosing $(A^- \otimes B^-)$ as a generalized inverse of $A \otimes B$.

B Proof of Theorem 2.3

We complete the proofs of Theorem 2.3. Since we heavily rely on the necessary conditions of local minima, we remind the reader of the elementary logic: for a point to be a local minimum, it must satisfy all the *necessary* conditions of local minima, but a point satisfying the *necessary* conditions can be a point that is not a local minimum (in contrast, a point satisfying the *sufficient* condition of local minimum is a local minimum).

B.1 Proof of Theorem 2.3 *(ii)*

Proof By case analysis, we show that any point that satisfies the necessary conditions and the definition of a local minimum is a global minimum. When we write a statement in the proof, we often mean that a necessary condition of local minima implies the statement as it should be clear (i.e., we are not claiming that the statement must hold true unless the point is the candidate of local minima.).

Case I: $\operatorname{rank}(W_H \cdots W_2) = p$ and $d_y \leq p$: Assume that $\operatorname{rank}(W_H \cdots W_2) = p$. We first obtain a necessary condition of the Hessian being positive semidefinite at a critical point, $Xr = 0$, and then interpret the condition. If $d_y < p$, Corollary 4.5 with $k = H + 1$ implies the necessary condition that $Xr = 0$. This is because the other condition $p > \operatorname{rank}(W_{H+1}) \geq \operatorname{rank}(W_H \cdots W_2) = p$ is false.

If $d_y = p$, Lemma 4.6 with $k = H + 1$ implies the necessary condition that $Xr = 0$ or $\mathcal{R}(W_H \cdots W_2) \subseteq \mathcal{R}(C^TC)$. Suppose that $\mathcal{R}(W_H \cdots W_2) \subseteq \mathcal{R}(C^TC)$. Then, we have that $p = \operatorname{rank}(W_H \cdots W_2) \leq \operatorname{rank}(C^TC) = \operatorname{rank}(C)$. That is, $\operatorname{rank}(C) \geq p$.

From Corollary 4.5 with $k = 2$ implies the necessary condition that

$$\operatorname{rank}(C) \geq \operatorname{rank}(I_{d_1}) \ \ \textbf{or} \ \ XrW_{H+1} \cdots W_3 = 0.$$

Suppose the latter: $XrW_{H+1} \cdots W_3 = 0$. Since $\operatorname{rank}(W_{H+1} \cdots W_3) \geq \operatorname{rank}(C) \geq p$ and $d_{H+1} = d_y = p$, the left null space of $W_{H+1} \cdots W_3$ contains only zero. Thus,

$$XrW_{H+1} \cdots W_3 = 0 \Rightarrow Xr = 0.$$

Suppose the former: $\operatorname{rank}(C) \geq \operatorname{rank}(I_{d_1})$. Because $d_y = p$, $\operatorname{rank}(C) \geq p$, and $\mathcal{R}(C) \subseteq \mathcal{R}(YX^T)$ as shown in the proof of Lemma 4.6, we have that $\mathcal{R}(C) = \mathcal{R}(YX^T)$.

$$\operatorname{rank}(C) \geq \operatorname{rank}(I_{d_1}) \Rightarrow C^TC \text{ is full rank } \Rightarrow Xr = XY^TC(C^TC)^{-1}C^T - XY^T = 0,$$

where the last equality follows the fact that $(Xr)^T = C(C^TC)^{-1}C^TYX^T - YX^T = 0$ since $\mathcal{R}(C) = \mathcal{R}(YX^T)$ and thereby the projection of YX^T onto the range of C is YX^T. Therefore, we have the condition, $Xr = 0$ when $d_y \leq p$.

To interpret the condition $Xr = 0$, consider a loss function with a linear model without any hidden layer, $f(W') = \|W'X - Y\|_F^2$ where $W' \in \mathbb{R}^{d_y \times d_x}$. Let $r' = (W'X - Y)^T$ be the corresponding error matrix. Then, any point satisfying $Xr' = 0$ is known to be a global minimum of f by its convexity.[5] For any values of $W_{H+1} \cdots W_1$, there exists W' such that $W' = W_{H+1} \cdots W_1$ (the

[5] proof: any point satisfying $Xr' = 0$ is a critical point of f, which directly follows the proof of Lemma 4.1. Also, f is convex since its Hessian is positive semidefinite for all input W_{H+1}, and thus any critical point of f is a global minimum. Combining the pervious two statements results in the desired claim

opposite is also true when $d_y \leq p$ although we don't need it in our proof). That is, image$(\bar{L}) \subseteq$ image(f) and image$(r) \subseteq$ image(r') (as functions of W and W' respectively) (the equality is also true when $d_y \leq p$ although we don't need it in our proof). Summarizing the above, whenever $Xr = 0$, there exists $W' = W_{H+1} \cdots W_1$ such that $Xr = Xr' = 0$, which achieves the global minimum value of f (f^*) and $f^* \leq \bar{\mathcal{L}}^*$ (i.e., the global minimum value of f is at most the global minimum value of $\bar{\mathcal{L}}$ since image$(\bar{\mathcal{L}}) \subseteq$ image(f)). In other words, $W_{H+1} \cdots W_1$ achieving $Xr = 0$ attains a global minimum value of f that is at most the global minimum value of $\bar{\mathcal{L}}$. This means that $W_{H+1} \cdots W_1$ achieving $Xr = 0$ is a global minimum.

Thus, we have proved that when $\operatorname{rank}(W_H \cdots W_2) = p$ and $d_y \leq p$, if $\nabla^2 \bar{\mathcal{L}}(W) \succeq 0$ at a critical point, it is a global minimum.

Case II: $\operatorname{rank}(W_H \cdots W_2) = p$ and $d_y > p$: We first obtain a necessary condition of the Hessian being positive semidefinite at a critical point and then interpret the condition. From Lemma 4.6, we have that $C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}^T$ or $Xr = 0$. If $Xr = 0$, with the exact same proof as in the case of $d_y \leq p$, it is a global minimum. Suppose that $C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}$. Combined with Lemma 4.2, we have a necessary condition:

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

From Lemma 4.4 with $k = H + 1$, $\mathcal{R}(W_2^T \cdots W_H^T) \subseteq \mathcal{R}(C^T C) = \mathcal{R}(C^T)$, which implies that $\bar{p} \triangleq \operatorname{rank}(C) = p$ (since $\operatorname{rank}(W_H \cdots W_2) = p$). Thus, we can rewrite the above equation as $W_{H+1} \cdots W_1 = U_p U_p^T Y X^T (X X^T)^{-1}$, which is the orthogonal projection on to 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 (Baldi & Hornik, 1989; Baldi & Lu, 2012).

Thus, we have proved that when $\operatorname{rank}(W_H \cdots W_2) = p$, if $\nabla^2 \bar{\mathcal{L}}(W) \succeq 0$ at a critical point, it is a global minimum.

Case III: $\operatorname{rank}(W_H \cdots W_2) < p$: Suppose that $\operatorname{rank}(W_H \cdots W_2) < p$. Let $\hat{p} = \min(p, d_y)$. Then, if $\operatorname{rank}(C) \geq \hat{p}$, every local minimum is a global minimum because of the following. If $p \leq d_y$, $\operatorname{rank}(W_H \cdots W_2) \geq \operatorname{rank}(C) \geq \hat{p} = p$ and thereby we have the case of $\operatorname{rank}(W_H \cdots W_2) = p$ (since we have that $p \geq \operatorname{rank}(W_H \cdots W_2) \geq p$ where the first inequality follows the definition of p). For this case, we have already proven the desired statement above. On the other hand, if $p > d_y$, we have $\bar{p} \triangleq \operatorname{rank}(C) \geq d_y$. Thus, $W_{H+1} \cdots W_1 = U_{\bar{p}} U_{\bar{p}}^T Y X^T (X X^T)^{-1} = U U^T Y X^T (X X^T)^{-1}$, which is a global minimum. We can see this in various ways. For example, $Xr = XY^T U U^T - XY^T = 0$, which means that it is a global minimum as discussed above.

Thus, in the following, we consider the remaining case where $\operatorname{rank}(W_H \cdots W_2) < p$ and $\operatorname{rank}(C) < \hat{p}$. In this case, we show that we can have $\operatorname{rank}(C) \geq \hat{p}$ with arbitrarily small perturbations of each entry of $W_{H+1}, \ldots, W_1$, without changing the loss value. In order to show this, by induction on $k = \{1, \ldots, H + 1\}$, we prove that we can have $\operatorname{rank}(W_k \cdots W_1) \geq \hat{p}$ with arbitrarily small perturbation of each entry of $W_k, \ldots, W_1$ without changing the value of $\bar{\mathcal{L}}(W)$.

We start with the base case with $k = 1$. For convenience, we reprint a necessary condition of local minima that is represented by equation 2 in the proof of Lemmas 4.2: for an arbitrary L_1,

$$W_1 = (C^T C)^- C^T Y X^T (X X^T)^{-1} + (I - (C^T C)^- C^T C) L_1 \tag{6}$$

Suppose that $(C^T C) \in \mathbb{R}^{d_1 \times d_1}$ is nonsingular. Then, we have that $\operatorname{rank}(W_H \cdots W_2) \geq \operatorname{rank}(C) = d_1 \geq p$, which is false in the case being analyzed (the case of $\operatorname{rank}(W_H \cdots W_2) < p$). Thus, $C^T C$ is singular.

If $C^T C$ is singular, it is inferred that we can perturb W_1 to have $\operatorname{rank}(W_1) \geq \hat{p}$. To see this in a concrete algebraic way, first note that from Lemma 4.6, $\mathcal{R}(C) = \mathcal{R}(U_{\bar{p}})$ or $Xr = 0$. If $Xr = 0$, with the exact same proof as in the previous case, it is a global minimum. So, we consider the case of $\mathcal{R}(C) = \mathcal{R}(U_{\bar{p}})$. Then, we can write $C = [U_{\bar{p}} \;\; \mathbf{0}] G_1$ for some $G_1 \in GL_{d_1}(\mathbb{R})$ where $\mathbf{0} \in \mathbb{R}^{d_y \times (d_1 - \bar{p})}$. Thus,

$$C^T C = G_1^T \begin{bmatrix} I_{\bar{p}} & 0 \\ 0 & 0 \end{bmatrix} G_1.$$

Again, note that the set of all generalized inverse of $G_1^T \begin{bmatrix} I_{\bar p} & 0 \\ 0 & 0 \end{bmatrix} G_1$ is as follows (Zhang, 2006, p. 41):

$$\left\{ G_1^{-1} \begin{bmatrix} I_{\bar p} & L_1' \\ L_2' & L_3' \end{bmatrix} G_1^{-T} \mid L_1', L_2', L_3' \text{ arbitrary} \right\}.$$

Since equation 6 must necessarily hold for *any generalized inverse* in order for a point to be a local minimum, we choose a generalized inverse with $L_1' = L_2' = L_3' = 0$ to have a weaker yet simpler necessary condition. That is,

$$(C^TC)^- := G_1^{-1} \begin{bmatrix} I_{\bar p} & 0 \\ 0 & 0 \end{bmatrix} G_1^{-T}.$$

By plugging this into equation 6, we obtain the following necessary condition of local minima: for an arbitrary L_1,

$$\begin{aligned} W_1 &= G_1^{-1} \begin{bmatrix} U_{\bar p}^T \\ 0 \end{bmatrix} YX^T(XX^T)^{-1} + (I_{d_1} - G_1^{-1} \begin{bmatrix} I_{\bar p} & 0 \\ 0 & 0 \end{bmatrix} G_1) L_1 \\ &= G_1^{-1} \begin{bmatrix} U_{\bar p}^T YX^T(XX^T)^{-1} \\ 0 \end{bmatrix} + G_1^{-1} \begin{bmatrix} 0 & 0 \\ 0 & I_{(d_1-\bar p)} \end{bmatrix} G_1 L_1 \\ &= G_1^{-1} \begin{bmatrix} U_{\bar p}^T YX^T(XX^T)^{-1} \\ [0 \;\; I_{(d_1-\bar p)}] G_1 L_1 \end{bmatrix}. \end{aligned} \tag{7}$$

Here, $[0 \quad I_{(d_1-\bar p)}]G_1L_1 \in \mathbb{R}^{(d_1-\bar p)\times d_x}$ is the last $(d_1 - \bar p)$ rows of G_1L_1. Since $\operatorname{rank}(YX^T(XX^T)^{-1}) = d_y$ (because the multiplication with the invertible matrix preserves the rank), the submatrix with the first $\bar p$ rows in the above have rank $\bar p$. Thus, W_1 has rank at least $\bar p$, and the possible rank deficiency comes from the last $(d_1 - \bar p)$ rows, $[0 \quad I_{(d_1-\bar p)}]G_1L_1$. Since $W_{H+1} \cdots W_1 = CW_1 = [U_{\bar p} \;\; \mathbf{0}]G_1W_1$,

$$W_{H+1} \cdots W_1 = [U_{\bar p} \;\; \mathbf{0}] \begin{bmatrix} U_{\bar p}^T YX^T(XX^T)^{-1} \\ [0 \;\; I_{(d_1-\bar p)}] G_1 L_1 \end{bmatrix} = U_{\bar p} U_{\bar p}^T YX^T(XX^T)^{-1}.$$

This means that changing the values of the last $(d_1 - \bar p)$ rows of G_1L_1 (i.e., $[0 \;\; I_{(d_1-\bar p)}]G_1L_1$) does not change the value of $\bar{\mathcal{L}}(W)$. Thus, we consider the perturbation of each entry of W_1 as follows:

$$\tilde W_1 := W_1 + \epsilon G_1^{-1} \begin{bmatrix} 0 \\ M_{\text{ptb}} \end{bmatrix} = G_1^{-1} \begin{bmatrix} U_{\bar p}^T YX^T(XX^T)^{-1} \\ [0 \;\; I_{(d_1-\bar p)}] G_1 L_1 + \epsilon M_{\text{ptb}} \end{bmatrix}.$$

Here, with an appropriate choice of M_{ptb}, we can make $\tilde W_1$ to be full rank (see footnote 6 for the proof of the existence of such M_{ptb}).[6]

Thus, we have shown that we can have $\operatorname{rank}(W_1) \geq \min(d_1, d_x) \geq \min(p, d_y) = \hat p$ with arbitrarily small perturbation of each entry of W_1 with the loss value being unchanged. This concludes the proof for the base case of the induction with $k = 1$.

For the inductive step[7] with $k \in \{2, \ldots, H+1\}$, we have the inductive hypothesis that we can have $\operatorname{rank}(W_{k-1} \cdots W_1) \geq \hat p$ with arbitrarily small perturbations of each entry of $W_{k-1}, \ldots W_1$ without changing the loss value. Here, we want to show that if $\operatorname{rank}(W_{k-1} \cdots W_1) \geq \hat p$, we can have

[6] In this footnote, we prove the existence of ϵM_{ptb} that makes W_1 full rank. Although this is trivial since the set of full rank matrices is dense, we show a proof in the following to be complete. Let $\bar p' \geq \bar p$ be the rank of W_1. That is, in $\begin{bmatrix} U_{\bar p}^T YX^T(XX^T)^{-1} \\ [0 \;\; I_{(d_1-\bar p)}] G_1 L_1 \end{bmatrix}$, there exist $\bar p'$ linearly independent row vectors including the first $\bar p$ row vectors, denoted by $b_1, \ldots, b_{\bar p'} \in \mathbb{R}^{1\times d_x}$. Then, we denote the rest of row vectors by $v_1, v_2, \ldots, v_{d_1-\bar p'} \in \mathbb{R}^{1\times d_x}$. Let $c = \min(d_1 - \bar p', d_x - \bar p')$. There exist linearly independent vectors $\bar v_1, \bar v_2, \ldots, \bar v_c$ such that the set, $\{b_1, \ldots, b_{\bar p'}, \bar v_1, \bar v_2, \ldots, \bar v_c\}$, is linearly independent. Setting $v_i := v_i + \epsilon \bar v_i$ for all $i \in \{1, \ldots, c\}$ makes W_1 full rank since $\epsilon \bar v_i$ cannot be expressed as a linear combination of other vectors. Thus, a desired perturbation matrix ϵM_{ptb} can be obtained by setting ϵM_{ptb} to consist of $\epsilon \bar v_1, \epsilon \bar v_2, \ldots, \epsilon \bar v_c$ row vectors for the corresponding rows and 0 row vectors for other rows.

[7] The boundary cases with $k = 2$ and $k = H + 1$ as well pose no problem during the proof for the inductive step: remember our notational definition, $W_k \cdots W_{k'} \triangleq I_{d_k}$ if $k < k'$.

$\operatorname{rank}(W_k \cdots W_1) \geq \hat{p}$ with arbitrarily small perturbation of each entry of W_k without changing the value of $\bar{\mathcal{L}}(W)$. Accordingly, suppose that $\operatorname{rank}(W_{k-1} \cdots W_1) \geq \hat{p}$. From Lemma 4.4, we have the following necessary condition for the Hessian to be (positive or negative) semidefinite at a critical point: for any $k \in \{2, \dots, H+1\}$,

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

where the first condition is shown to imply $\operatorname{rank}(W_{H+1} \cdots W_k) \geq \operatorname{rank}(W_{k-1} \cdots W_2)$ in Corollary 4.5. If the former condition is true, $\operatorname{rank}(C) \geq \operatorname{rank}(W_{k-1} \cdots W_2) \geq \operatorname{rank}(W_{k-1} \cdots W_1) \geq \hat{p}$, which is false in the case being analyzed (i.e., the case where $\operatorname{rank}(C) < \hat{p}$. If this is not the case, we can immediately conclude the desired statement as it has been already proven for the case where $\operatorname{rank}(C) \geq \hat{p}$). Thus, we suppose that the latter condition is true. Let $A_k = W_{H+1} \cdots W_{k+1}$. Then, for an arbitrary L_k,

$$\begin{aligned} 0 &= XrW_{H+1} \cdots W_{k+1} \\ \Rightarrow W_k \cdots W_1 &= \left(A_k^T A_k\right)^{-} A_k^T Y X^T (XX^T)^{-1} + (I - (A_k^T A_k)^{-} A_k^T A_k) L_k \qquad (8) \\ \Rightarrow W_{H+1} \cdots W_1 &= A_k \left(A_k^T A_k\right)^{-} A_k^T Y X^T (XX^T)^{-1} \\ &= C(C^T C)^{-} C^T Y X^T (XX^T)^{-1} = U_{\bar{p}} U_{\bar{p}}^T Y X^T (XX^T)^{-1}, \end{aligned}$$

where the last two equalities follow Lemmas 4.2 and 4.6 (since if $Xr = 0$, we immediately obtain the desired result as discussed above). Taking transpose,

$$(XX^T)^{-1} XY^T A_k \left(A_k^T A_k\right)^{-} A_k^T = (XX^T)^{-1} XY^T U_{\bar{p}} U_{\bar{p}}^T,$$

which implies that

$$XY^T A_k \left(A_k^T A_k\right)^{-} A_k = XY^T U_{\bar{p}} U_{\bar{p}}.$$

Since XY^T is full rank with $d_y \leq d_x$ (i.e., $\operatorname{rank}(XY^T) = d_y$), there exists a left inverse and the solution of the above linear system is unique as $((XY^T)^T XY^T)^{-1}(XY^T)^T XY^T = I$, yielding,

$$A_k \left(A_k^T A_k\right)^{-} A_k = U_{\bar{p}} U_{\bar{p}}^T \; (= U_{\bar{p}} (U_{\bar{p}}^T U_{\bar{p}})^{-1} U_{\bar{p}}^T).$$

In other words, $\mathcal{R}(A_k) = \mathcal{R}(C) = \mathcal{R}(U_{\bar{p}})$.

Suppose that $(A_k^T A_k) \in \mathbb{R}^{d_k \times d_k}$ is nonsingular. Then, since $\mathcal{R}(A_k) = \mathcal{R}(C)$, $\operatorname{rank}(C) = \operatorname{rank}(A_k) = d_k \geq \hat{p} \triangleq \min(p, d_y)$, which is false in the case being analyzed (the case of $\operatorname{rank}(C) < \hat{p}$). Thus, $A_k^T A_k$ is singular. Notice that for the boundary case with $k = H+1$, $A_k^T A_k = I_{d_y}$, which is always nonsingular and thus the proof ends here (i.e., For the case with $k = H+1$, since the latter condition, $XrW_{H+1} \cdots W_{k+1} = 0$, implies a false statement, the former condition, $\operatorname{rank}(C) \geq \hat{p}$, which is the desired statement, must be true).

If $A_k^T A_k$ is singular, it is inferred that we can perturb W_k to have $\operatorname{rank}(W_k \cdots W_1) \geq \min(p, d_x)$. To see this in a concrete algebraic way, first note that since $\mathcal{R}(A_k) = \mathcal{R}(U_{\bar{p}})$, we can write $A_k = [U_{\bar{p}} \;\; \mathbf{0}] G_k$ for some $G_k \in GL_{d_k}(\mathbb{R})$ where $\mathbf{0} \in \mathbb{R}^{d_y \times (d_k - \bar{p})}$. Then, similarly to the base case with $k = 1$, we select a general inverse (we can do this because it remains to be a necessary condition as explained above) to be

$$(A_k^T A_k)^{-} := G_k^{-1} \begin{bmatrix} I_{\bar{p}} & 0 \\ 0 & 0 \end{bmatrix} G_k^{-T},$$

and plugging this into the condition in equation 8: for an arbitrary L_k,

$$W_k \cdots W_1 = G_k^{-1} \begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k \end{bmatrix}. \qquad (9)$$

Here, $[0 \quad I_{(d_k - \bar{p})}] G_k L_k \in \mathbb{R}^{(d_k - \bar{p}) \times d_x}$ is the last $(d_k - \bar{p})$ rows of $G_k L_k$. Since $\operatorname{rank}(Y X^T (XX^T)^{-1}) = d_y$, the first $\bar{p}$ rows in the above have rank $\bar{p}$. Thus, $W_k \cdots W_1$ has rank at least $\bar{p}$ and the possible rank deficiency comes from the last $(d_k - \bar{p})$ rows, $[0 \;\; I_{(d_k - \bar{p})}] G_k L_k$. Since $W_{H+1} \cdots W_1 = A_k W_k \cdots W_1 = [U_{\bar{p}} \;\; \mathbf{0}] G_k W_k \cdots W_1$,

$$W_{H+1} \cdots W_1 = [U_{\bar{p}} \;\; \mathbf{0}] \begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k \end{bmatrix} = U_{\bar{p}} U_{\bar{p}}^T Y X^T (XX^T)^{-1},$$

which means that changing the values of the last $(d_k - \bar{p})$ rows does not change the value of $\bar{\mathcal{L}}(W)$.

We consider the perturbation of each entry of W_k as follows. From equation 9, all the possible solutions of W_k can be written as: for an arbitrary L_{0_k} and L_k,

$$W_k = G_k^{-1} \begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k \end{bmatrix} B_k^\dagger + L_{0_k}^T (I - B_k B_k^\dagger).$$

where $B_k = W_{k-1} \cdots W_1$ and $B_k^\dagger$ is the the Moore–Penrose pseudoinverse of B_k. We perturb W_k as

$$\begin{aligned} \tilde{W}_k &:= W_k + \epsilon G_k^{-1} \begin{bmatrix} 0 \\ M \end{bmatrix} B_k^\dagger \\ &= G_k^{-1} \begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k + \epsilon M \end{bmatrix} B_k^\dagger + L_{0_k}^T (I - B_k B_k^\dagger). \end{aligned}$$

where $M = M_{\text{ptb}} (B_k^T B_k)^\dagger B_k^T B_k$. Then,

$$\begin{aligned} \tilde{W}_k W_{k-1} \cdots W_1 &= \tilde{W}_k B_k \\ &= G_k^{-1} \begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k \end{bmatrix} B_k^\dagger B_k + G_k^{-1} \begin{bmatrix} 0 \\ \epsilon M \end{bmatrix} B_k^\dagger B_k \\ &= G_k^{-1} \begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k \end{bmatrix} + G_k^{-1} \begin{bmatrix} 0 \\ \epsilon M \end{bmatrix} B_k^\dagger B_k \\ &= G_k^{-1} \begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k + \epsilon M_{\text{ptb}} (B_k^T B_k)^\dagger B_k^T B_k \end{bmatrix}, \end{aligned}$$

where the second line follows equation 9 and the third line is due to the fact that $M B_k^\dagger B_k = M_{\text{ptb}} (B_k^T B_k)^\dagger B_k^T (B_k B_k^\dagger B_k) = M_{\text{ptb}} (B_k^T B_k)^\dagger B_k^T B_k$. Here, we can construct M_{ptb} such that $\operatorname{rank}(\tilde{W}_k B_k) \geq \hat{p}$ as follows. Let $\bar{p}' \geq \bar{p}$ be the rank of $\tilde{W}_k B_k$. That is, in $\begin{bmatrix} U_{\bar{p}}^T Y X^T (XX^T)^{-1} \\ [0 \;\; I_{(d_k - \bar{p})}] G_k L_k \end{bmatrix}$, there exist $\bar{p}'$ linearly independent row vectors including the first $\bar{p}$ row vectors, denoted by $b_1, \ldots, b_{\bar{p}'} \in \mathbb{R}^{1 \times d_x}$. Then, we denote the rest of row vectors by $v_1, v_2, \ldots, v_{d_k - \bar{p}'} \in \mathbb{R}^{1 \times d_x}$. Since $\operatorname{rank}(B_k^T B_k) \geq \hat{p}$ (due to the inductive hypothesis), the dimension of $\mathcal{R}(B_k^T B_k)$ is at least $\hat{p}$. Therefore, there exist vectors $\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{(\hat{p} - \bar{p}')}$ such that the set, $\{b_1^T, \ldots, b_{\bar{p}'}^T, \bar{v}_1^T, \bar{v}_2^T, \ldots, \bar{v}_{(\hat{p} - \bar{p}')}^T\}$, is linearly independent and $\bar{v}_1^T, \bar{v}_2^T, \ldots, \bar{v}_{(\hat{p} - \bar{p}')}^T \in \mathcal{R}(B_k^T B_k)$. A desired perturbation matrix M_{ptb} can be obtained by setting M_{ptb} to consist of $\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{(\hat{p} - \bar{p})}$ row vectors for the first $(\hat{p} - \bar{p})$ rows and 0 row vectors for the rest:

$$M_{\text{ptb}}^T := \begin{bmatrix} \bar{v}_1^T & \cdots & \bar{v}_{(\hat{p} - \bar{p})}^T & 0 & \cdots & 0 \end{bmatrix}.$$

Then, $M_{\text{ptb}} (B_k^T B_k)^\dagger B_k^T B_k = (B_k^T B_k (B_k^T B_k)^\dagger M_{\text{ptb}}^T)^T = M_{\text{ptb}}$ (since $\bar{v}_1^T, \bar{v}_2^T, \ldots, \bar{v}_{(\hat{p} - \bar{p})}^T \in \mathcal{R}(B_k^T B_k)$). Thus, as a result of our perturbation, the original row vectors $v_1, v_2, \ldots, v_{(\hat{p} - \bar{p}')}$ are perturbated as $v_i := v_i + \epsilon \bar{v}_i$ for all $i \in \{1, \ldots, \hat{p} - \bar{p}'\}$, which guarantees $\operatorname{rank}(\tilde{W}_k B_k) \geq \hat{p}$ since $\epsilon \bar{v}_i$ cannot be expressed as a linear combination of other row vectors ($b_1, \ldots, b_{\bar{p}'}$ and $\forall j \neq i, \bar{v}_j$) by its construction. Therefore, we have that $\operatorname{rank}(W_k \cdots W_1) \geq \hat{p}$ upon such a perturbation on W_k without changing the loss value.

Thus, we conclude the induction, proving that we can have $\operatorname{rank}(W_{H+1} \cdots W_1) \geq \hat{p}$ with arbitrarily small perturbation of each parameter without changing the value of $\bar{\mathcal{L}}(W)$. Since $\operatorname{rank}(C) \geq \operatorname{rank}(W_{H+1} \cdots W_1) \geq \hat{p}$, upon such a perturbation, we have the case where $\operatorname{rank}(C) \geq \hat{p}$, for which we have already proven that a critical point is not a local minimum unless it is a global minimum. This concludes the proof of the case where $\operatorname{rank}(W_H \cdots W_2) < p$.

Summarizing the above, any point that satisfies the definition (and necessary conditions) of a local minimum is a global minimum, concluding the proof of **Theorem 2.3 *(ii)***. □

B.2 Proof of Theorem 2.3 *(i)*

Proof We can prove the non-convexity and non-concavity from its Hessian (Theorem 2.3 *(i)*). First, consider $\bar{\mathcal{L}}(W)$. For example, from Corollary 4.5 with $k = H + 1$, it is necessary for the Hessian to be positive or negative semidefinite at a critical point that $\operatorname{rank}(W_{H+1}) \geq \operatorname{rank}(W_H \cdots W_2)$ or $Xr = 0$. The instances of W unsatisfying this condition at critical points form some uncountable set. As an example, consider a uncountable set that consists of the points with $W_{H+1} = W_1 = 0$ and with any $W_H, \ldots, W_2$. Then, every point in the set defines a critical point from Lemma 4.1. Also, $Xr = XY^T \neq 0$ as $\operatorname{rank}(XY^T) \geq 1$. So, it does not satisfy the first semidefinite condition. On the other hand, with any instance of $W_H \cdots W_2$ such that $\operatorname{rank}(W_H \cdots W_2) \geq 1$, we have that $0 = \operatorname{rank}(W_{H+1}) \ngeq \operatorname{rank}(W_H \cdots W_2)$. So, it does not satisfy the second semidefinite condition as well. Thus, we have proven that in the domain of the loss function, there exist points, at which the Hessian becomes indefinite. **This implies Theorem 2.3 *(i)*: the functions are non-convex and non-concave.**

□

B.3 Proof of Theorem 2.3 *(iii)*

Proof We now prove Theorem 2.3 *(iii)*: every critical point that is not a global minimum is a saddle point. Here, we want to show that if the Hessian is negative semidefinite at a critical point, then there is a increasing direction so that there is no local maximum. From Lemma 4.3 with $k = 1$,

$$\mathcal{D}_{\mathrm{vec}(W_1^T)} \left(\mathcal{D}_{\mathrm{vec}(W_1^T)} \bar{\mathcal{L}}(W)\right)^T = \left((W_{H+1} \cdots W_2)^T (W_{H+1} \cdots W_2) \otimes XX^T\right) \succeq 0.$$

The positive semidefiniteness follows the fact that $(W_{H+1} \cdots W_2)^T (W_{H+1} \cdots W_2)$ and XX^T are positive semidefinite. Since XX^T is full rank, if $(W_{H+1} \cdots W_2)^T (W_{H+1} \cdots W_2)$ has at least one strictly positive eigenvalue, $(W_{H+1} \cdots W_2)^T (W_{H+1} \cdots W_2) \otimes XX^T$ has at least one strictly positive eigenvalue (by the spectrum property of Kronecker product). Thus, with other variables being fixed, if $W_{H+1} \cdots W_2 \neq 0$, with respect to W_1 at any critical point, there exists some increasing direction that corresponds to the strictly positive eigenvalue. This means that there is no local maximum if $W_{H+1} \cdots W_2 \neq 0$.

If $W_{H+1} \cdots W_2 = 0$, we claim that at a critical point, if the Hessian is negative semidefinite (i.e., a necessary condition of local maxima), we can make $W_{H+1} \cdots W_2 \neq 0$ with arbitrarily small perturbation of each parameter without changing the loss value. We can prove this by using the similar proof procedure to that used for Theorem 2.3 *(ii)* in the case of $\operatorname{rank}(W_H \cdots W_2) < p$. Suppose that $W_{H+1} \cdots W_2 = 0$ and thus $\operatorname{rank}(W_{H+1} \cdots W_2) = 0$. By induction on $k = \{2, \ldots, H + 1\}$, we prove that we can have $W_k \cdots W_2 \neq 0$ with arbitrarily small perturbation of each entry of $W_k, \ldots, W_2$ without changing the loss value.

We start with the base case with $k = 2$. From Lemma 4.4, we have a following necessary condition for the Hessian to be (positive or negative) semidefinite at a critical point: for any $k \in \{2, \ldots, H + 1\}$,

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

where the first condition is shown to imply $\operatorname{rank}(W_{H+1} \cdots W_k) \geq \operatorname{rank}(W_{k-1} \cdots W_2)$ in Corollary 4.5. Let $A_k = W_{H+1} \cdots W_{k+1}$. From the condition with $k = 2$, we have that $\operatorname{rank}(W_{H+1} \cdots W_2) \geq d_1 \geq 1$ or $XrW_{H+1} \cdots W_3 = 0$. The former condition is false since $\operatorname{rank}(W_H \cdots W_2) < 1$. From the latter condition, for an arbitrary L_2,

$$\begin{aligned}
&0 = XrW_{H+1} \cdots W_3 \\
\Rightarrow &W_2 W_1 = \left(A_2^T A_2\right)^- A_2^T Y X^T (XX^T)^{-1} + (I - (A_2^T A_2)^- A_2^T A_2) L_2 \\
\Rightarrow &W_{H+1} \cdots W_1 = A_2 \left(A_2^T A_2\right)^- A_2^T Y X^T (XX^T)^{-1} \\
&\qquad\qquad\quad = C(C^T C)^- C^T Y X^T (XX^T)^{-1}
\end{aligned} \tag{10}$$

where the last follows the critical point condition (Lemma 4.2). Then, similarly to the proof of Theorem 2.3 *(ii)*,

$$A_2 \left(A_2^T A_2\right)^- A_2 = C(C^T C)^- C^T.$$

In other words, $\mathcal{R}(A_2) = \mathcal{R}(C)$.

Suppose that $\mathrm{rank}(A_2^T A_2) \geq 1$. Then, since $\mathcal{R}(A_2) = \mathcal{R}(C)$, we have that $\mathrm{rank}(C) \geq 1$, which is false (or else the desired statement). Thus, $\mathrm{rank}(A_2^T A_2) = 0$, which implies that $A_2 = 0$. Then, since $W_{H+1} \cdots W_1 = A_2 W_2 W_1$ with $A_2 = 0$, we can have $W_2 \neq 0$ without changing the loss value with arbitrarily small perturbation of W_2.

For the inductive step with $k = \{3, \dots, H+1\}$, we have the inductive hypothesis that we can have $W_{k-1} \cdots W_2 \neq 0$ with arbitrarily small perturbation of each parameter without changing the loss value. Accordingly, suppose that $W_{k-1} \cdots W_2 \neq 0$. Again, from Lemma 4.4, for any $k \in \{2, \dots, H+1\}$,

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

If the former is true, $\mathrm{rank}(C) \geq \mathrm{rank}(W_{k-1} \cdots W_2) \geq 1$, which is false (or the desired statement). If the latter is true, for an arbitrary L_1,

$$\begin{aligned}
&0 = XrW_{H+1} \cdots W_{k+1} \\
\Rightarrow &W_k \cdots W_1 = \left(A_k^T A_k\right)^{-} A_k^T Y X^T (XX^T)^{-1} + (I - (A_k^T A_k)^{-} A_k^T A_k) L_1 \\
\Rightarrow &W_{H+1} \cdots W_1 = A_k \left(A_k^T A_k\right)^{-} A_k^T Y X^T (XX^T)^{-1} \\
&\qquad\qquad\qquad = C(C^T C)^{-} C^T Y X^T (XX^T)^{-1} = U_{\bar{p}} U_{\bar{p}}^T Y X^T (XX^T)^{-1},
\end{aligned}$$

where the last follows the critical point condition (Lemma 4.2). Then, similarly to the above,

$$A_k \left(A_k^T A_k\right)^{-} A_k = C(C^T C)^{-} C^T.$$

In other words, $\mathcal{R}(A_k) = \mathcal{R}(C)$.

Suppose that $\mathrm{rank}(A_k^T A_k) \geq 1$. Then, since $\mathcal{R}(A_k) = \mathcal{R}(C)$, we have that $\mathrm{rank}(C) = \mathrm{rank}(A_k) \geq 1$, which is false (or the desired statement). Thus, $\mathrm{rank}(A_k^T A_k) = 0$, which implies that $A_k = 0$. Then, since $W_{H+1} \cdots W_1 = A_k W_k \cdots W_1$ with $A_k = 0$, we can have $W_k \cdots W_1 \neq 0$ without changing the loss value with arbitrarily small perturbation of each parameter.

Thus, we conclude the induction, proving that if $W_{H+1} \cdots W_2 = 0$, with arbitrarily small perturbation of each parameter without changing the value of $\bar{\mathcal{L}}(W)$, we can have $W_{H+1} \cdots W_2 \neq 0$. Thus, at any candidate point for local maximum, the loss function has some strictly increasing direction in an arbitrarily small neighborhood. This means that there is no local maximum. **Thus, we obtained the statement of Theorem 2.3 *(iii)*.**

□

B.4 Proof of Theorem 2.3 *(iv)*

Proof In the proof of Theorem 2.3 *(ii)*, the case analysis with the case, $\mathrm{rank}(W_H \cdots W_2) = p$, revealed that 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. Thus, when $\mathrm{rank}(W_H \cdots W_2) = p$, if W is not a global minimum at a critical point, its Hessian is not positive semidefinite, containing some negative eigenvalue. From Theorem 2.3 *(ii)*, if it is not a global minimum, it is not a local minimum. From Theorem 2.3 *(iii)*, it is a saddle point. Thus, if $\mathrm{rank}(W_H \cdots W_2) = p$, the Hessian at any saddle point has some negative eigenvalue, **which is the statement of Theorem 2.3 *(iv)*.**

□

C Proofs of Corollaries 2.4 and 3.2

We complete the proofs of Corollaries 2.4 and 3.2.

C.1 Proof of Corollary 2.4

Proof If $H = 1$, the condition in Theorem 2.3 *(iv)* reads "if $\mathrm{rank}(W_1 \cdots W_2) = \mathrm{rank}(I_{d_1}) = d_1 = p$", which is always true. This is because p is the smallest width of hidden layers and there is only one hidden layer, the width of which is d_1. Thus, Theorem 2.3 *(iv)* immediately implies the statement of Corollary 2.4. For the statement of Corollary 2.4 with $H \geq 2$, it is suffice to show the existence of a simple set containing saddle points with the Hessian having no negative eigenvalue. Suppose that $W_H = W_{H-1} = \cdots = W_2 = W_1 = 0$. Then, from Lemma 4.1, it defines an uncountable set of critical points, in which W_{H+1} can vary in $\mathbb{R}^{d_y \times d_H}$. Since $r = Y^T \neq 0$ due to $\operatorname{rank}(Y) \geq 1$, it is not a global minimum. To see this, we write

$$\begin{aligned}\bar{\mathcal{L}}(W) &= \frac{1}{2}\|\overline{Y}(W,X) - Y\|_F^2 = \frac{1}{2}\operatorname{tr}(r^T r) \\ &= \frac{1}{2}\operatorname{tr}(YY^T) - \frac{1}{2}\operatorname{tr}(W_{H+1}\cdots W_1 XY^T) - \frac{1}{2}\operatorname{tr}((W_{H+1}\cdots W_1 XY^T)^T) \\ &\quad + \frac{1}{2}\operatorname{tr}(W_{H+1}\cdots W_1 XX^T (W_{H+1}\cdots W_1)^T).\end{aligned}$$

For example, with $W_{H+1}\cdots W_1 = \pm\, U_p U_p^T Y X^T (XX)^{-1}$,

$$\begin{aligned}\bar{\mathcal{L}}(W) &= \frac{1}{2}\left(\operatorname{tr}(YY^T) - \operatorname{tr}(U_p U_p^T \Sigma) - \operatorname{tr}(\Sigma U_p U_p^T) + \operatorname{tr}(U_p U_p^T \Sigma U_p U_p^T)\right) \\ &= \frac{1}{2}\left(\operatorname{tr}(YY^T) - \operatorname{tr}(U_p \Lambda_{1:p} U_p^T)\right) = \frac{1}{2}\left(\operatorname{tr}(YY^T) \pm \sum_{k=1}^{p} \Lambda_{k,k}\right),\end{aligned}$$

where we can see that there exists a strictly lower value of $\bar{\mathcal{L}}(W)$ than the loss value with $r = Y^T$, which is $\frac{1}{2}\operatorname{tr}(YY^T)$ (since $X \neq 0$ and $\operatorname{rank}(\Sigma) \neq 0$).

Thus, these are not global minima, and thereby these are saddle points by Theorem 2.3 *(ii)* and *(iii)*. On the other hand, from the proof of Lemma 4.3, every diagonal and off-diagonal element of the Hessian is zero if $W_H = W_{H-1} = \cdots = W_2 = W_1 = 0$. Thus, the Hessian is simply a zero matrix, which has no negative eigenvalue.

□

C.2 Proof of Corollary 3.2 and discussion of the assumptions used in the previous work

Proof Since $E_Z[\hat{Y}(W,X)] = q\rho\sum_{p=1}^{\Psi}[X_i]_{(j,p)}\prod_{k=1}^{H+1} w_{(j,p)} = \overline{Y}$, $\mathcal{L}(W) = \frac{1}{2}\|E_Z[\hat{Y}(W,X) - Y]\|_F = \frac{1}{2}\|E_Z[\hat{Y}(W,X)] - Y\|_F^2 = \bar{\mathcal{L}}(W)$. □

The previous work also assumes the use of “independent random” loss functions. Consider the hinge loss, $\mathcal{L}_{\text{hinge}}(W)_{j,i} = \max(0,\ 1 - Y_{j,i}\hat{Y}(W,X)_{j,i})$. By modeling the max operator as a Bernoulli random variable ξ, we can then write $\mathcal{L}_{\text{hinge}}(W)_{j,i} = \xi - q\sum_{p=1}^{\Psi} Y_{j,i}[X_i]_{(j,p)}\xi[Z_i]_{(j,p)}\prod_{k=1}^{H+1} w_{(j,p)}^{(k)}$. A1p then assumes that for all i and (j,p), the $\xi[Z_i]_{(j,p)}$ are Bernoulli random variables with equal probabilities of success. Furthermore, A5u assumes that the independence of $\xi[Z_i]_{(j,p)}, Y_{j,i}[X_i]_{(j,p)}$, and $w_{(j,p)}$. Finally, A6u assumes that $Y_{j,i}[X_i]_{(j,p)}$ for all (j,p) and i are independent. In section 3.2, we discuss the effect of all of the seven previous assumptions to see why these are unrealistic.

D Discussion of the 1989 conjecture

The 1989 conjecture is based on the result for a 1-hidden layer network with $p < d_y = d_x$ (e.g., an autoencoder). That is, the previous work considered $\overline{Y} = W_2 W_1 X$ with the same loss function as ours with the additional assumption $p < d_y = d_x$. The previous work denotes $A \triangleq W_2$ and $B \triangleq W_1$.

The conjecture was expressed by Baldi & Hornik (1989) as

> Our results, and in particular the main features of the landscape of E, hold true in the case of linear networks with several hidden layers.

Here, the “main features of the landscape of E” refers to the following features, among other minor technical facts: 1) the function is convex in each matrix A (or B) when fixing other B (or A), and 2) every local minimum is a global minimum. No proof was provided in this work for this conjecture.

In 2012, the proof for the conjecture corresponding to the first feature (convexity in each matrix A (or B) when fixing other B (or A)) was provided in (Baldi & Lu, 2012) for both real-valued and complex-valued cases, while the proof for the conjecture for the second feature (every local minimum being a global minimum) was left for future work.

In (Baldi, 1989), there is an informal discussion regarding the conjecture. Let $i \in \{1, \cdots, H\}$ be an index of a layer with the smallest width p. That is, $d_i = p$. We write

$$A := W_{H+1} \cdots W_{i+1}$$

$$B := W_i \cdots W_1.$$

Then, what A and B can represent is the same as what the original $A := W_2$ and $B := W_1$, respectively, can represent in the 1-hidden layer case, assuming that $p < d_y = d_x$ (i.e., any element in $\mathbb{R}^{d_y \times p}$ and any element in $\mathbb{R}^{p \times d_x}$). Thus, we *would* conclude that all the local minima in the deeper models always correspond to the local minima of the collapsed 1-hidden layer version with $A := W_{H+1} \cdots W_{i+1}$ and $B := W_i \cdots W_1$.

However, the above reasoning turns out to be incomplete. Let us prove the incompleteness of the reasoning by contradiction in a way in which we can clearly see what goes wrong. Suppose that the reasoning is complete (i.e., the following statement is true: if we can collapse the model with the same expressiveness with the same rank restriction, then the local minima of the model correspond to the local minima of the collapsed model). Consider $f(w) = W_3 W_2 W_1 = 2w^2 + w^3$, where $W_1 = [w \;\; w \;\; w]$, $W_2 = [1 \;\; 1 \;\; w]^T$ and $W_3 = w$. Then, let us collapse the model as $a := W_3 W_2 W_1$ and $g(a) = a$. As a result, what $f(w)$ can represent is the same as what $g(a)$ can represent (i.e., any element in $\mathbb{R}$) with the same rank restriction (with a rank of at most one). Thus, with the same reasoning, we can conclude that every local minimum of $f(w)$ corresponds to a local minimum of $g(a)$. However, this is clearly false, as $f(w)$ is a non-convex function with a local minimum at $w = 0$ that is not a global minimum, while $g(a)$ is linear (convex and concave) without any local minima. The convexity for $g(a)$ is preserved after the composition with any norm. Thus, we have a contradiction, proving the incompleteness of the reasoning. What is missed in the reasoning is that even if what a model can represent is the same, the different parameterization creates different local structure in the loss surface, and thus different properties of the critical points (global minima, local minima, saddle points, and local maxima).

Now that we have proved the incompleteness of this reasoning, we discuss where the reasoning actually breaks down in a more concrete example. From Lemmas 4.1 and 4.2, if $H = 1$, we have the following representation at critical points:

$$AB = A(A^T A)^- A^T Y X^T (X X^T)^{-1}.$$

where $A := W_2$ and $B := W_1$. In contrast, from Lemmas 4.1 and 4.2, if H is arbitrary,

$$AB = C(C^T C)^- C^T Y X^T (X X^T)^{-1}.$$

where $A := W_{H+1} \cdots W_{i+1}$ and $B := W_i \cdots W_1$ as discussed above, and $C = W_{H+1} \cdots W_2$. Note that by using other critical point conditions from Lemmas 4.1, we cannot obtain an expression such that $C = A$ in the above expression unless $i = 1$. Therefore, even though what A and B can represent is the same, the critical condition becomes different (and similarly, the conditions from the Hessian). Because the proof in the previous work with $H = 1$ heavily relies on the fact that $AB = A(A^T A)^- A^T Y X^T (X X^T)^{-1}$, the same proof does not apply for deeper models (we may continue providing more evidence as to why the same proof does not work for deeper models, but one such example suffices for the purpose here).

In this respect, we have completed the proof of the conjecture and also provided a complete analytical proof for more general and detailed statements; that is, we did not assume that $p < d_y = d_x$, and we also proved saddle point properties with negative eigenvalue information.