

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

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

By setting $\left(\mathcal{D}_{\text{vec}(W_k^T)} \bar{\mathcal{L}}(W) \right)^T = 0$ for all $k \in \{1, \dots, H+1\}$, 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 $W_0 \cdots W_1 \triangleq I_{d_x}$). \square

A.2 Proof of Lemma 4.2

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

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

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

$$W_1 = (C^T C)^- C^T Y X^T (X X^T)^{-1} + (I - (C^T C)^- C^T C) L, \quad (2)$$

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

$$CW_1 = C(C^T C)^- C^T Y X^T (X X^T)^{-1} + (C - C(C^T C)^- C^T C) L = C(C^T C)^- C^T Y X^T (X X^T)^{-1}. \quad \square$$

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

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

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

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

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

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,

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

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 $(W_{H+1,j} W_H \cdots W_2)^T = \text{vec}(W_2^T \cdots W_H^T W_{H+1,j}^T) = (W_{H+1,j} \cdots W_{k+1} \otimes W_2^T \cdots W_{k-1}^T) \text{vec}(W_k^T)$. In the last line, we have the desired result by rewriting $\sum_{i=1}^m \sum_{j=1}^{d_y} r_{i,j} (A_k)_{j,t} X_i = X(r W_{H+1} \cdots W_{k+1})_{\cdot,t}$.

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 $k = 2, \dots, H$). For the second term, notice that $\text{vec}(W_{H+1}^T) = [(W_{H+1})_{1,\cdot}^T \quad \dots \quad (W_{H+1})_{d_y,\cdot}^T]^T$ 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 $\text{vec}(W_k^T)$ instead of $\text{vec}(W_k)$ for this reason, among others). Also,

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

where we also used the fact that

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

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

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

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

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

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

Here, $P_k = [\mathbf{e}_{H+1} \quad \mathbf{e}_k \quad \tilde{P}_k]$ 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,

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

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

$$\begin{aligned} 0 &= \left(I - (C^T C \otimes (XX^T))(C^T C \otimes (XX^T))^{-1} \right) (C^T A_k \otimes B_k W_1 X) \\ &\quad + \left(I - (C^T C \otimes (XX^T))(C^T C \otimes (XX^T))^{-1} \right) [B_k^T \otimes X] [I_{d_{k-1}} \otimes (rA_k)_{\cdot,1} \dots I_{d_{k-1}} \otimes (rA_k)_{\cdot,d_k}] \end{aligned}$$

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

$$\left((I_{d_1} - C^T C (C^T C)^{-1}) \otimes I_{d_y} \right) (C^T A_k \otimes B_k W_1 X) = ((I_{d_1} - C^T C (C^T C)^{-1}) C^T A_k) \otimes B_k W_1 X = 0,$$

since $C^T C (C^T C)^{-1} C^T = C^T$ (Zhang, 2006, p. 41). Thus, with the second term remained, the condition is reduced to

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

This implies

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

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

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

□

A.5 Proof of Corollary 4.5

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

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

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 $W_2^T \dots W_{H+1}^T$ is subspace of the column space of $W_2^T \dots W_{k-1}^T$. □

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 $P_{H+1}^{-1} \nabla^2 \tilde{\mathcal{L}}(W) P_{H+1}$ that consists of the first four blocks of it). Let $B_k = W_{k-1} \cdots W_2$. Let $F = B_{H+1} W_1 X X^T W_1^T B_{H+1}^T$. Using Lemma 4.3 for the blocks corresponding to W_1 and W_{H+1} ,

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

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

$$\begin{aligned} 0 &= ((I_{d_y} \otimes I_{d_H}) - (I_{d_y} \otimes F)(I_{d_y} \otimes F)^-)((C \otimes B_{H+1} W_1 X X^T) + E^T) \\ \Rightarrow 0 &= (I_{d_y} \otimes I_{d_H} - F F^-)(C \otimes B_{H+1} W_1 X X^T) + (I_{d_y} \otimes I_{d_H} - F F^-)E^T \\ &= (I_{d_y} \otimes I_{d_H} - F F^-)E^T \\ &= \begin{bmatrix} I_{d_H} - F F^- \otimes I_1 & \mathbf{0} \\ & \ddots \\ \mathbf{0} & I_{d_H} - F F^- \otimes I_1 \end{bmatrix} \begin{bmatrix} B_{H+1} \otimes (Xr_{:,1})^T \\ \vdots \\ B_{H+1} \otimes (Xr_{:,d_y})^T \end{bmatrix} \\ &= \begin{bmatrix} (I_{d_H} - F F^-)B_{H+1} \otimes (Xr_{:,1})^T \\ \vdots \\ (I_{d_H} - F F^-)B_{H+1} \otimes (Xr_{:,d_y})^T \end{bmatrix} \end{aligned}$$

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 $(I - F F^-)B_{H+1} W_1 X = 0$ because $\mathcal{R}(B_{H+1} W_1 X) = \mathcal{R}(B_{H+1} W_1 X X^T W_1^T B_{H+1}^T) = \mathcal{R}(F)$. In the fourth line, we expanded E and used the definition of the Kronecker product. It implies

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

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

For the $(C(C^T C)^- C^T = U_{\bar{p}} U_{\bar{p}}^T)$ 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

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

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

$$\begin{aligned} & \left(C \otimes B_{H+1} W_1 X X^T + E^T \right) (C^T C \otimes X X^T)^- \left(C^T \otimes X X^T (B_{H+1} W_1)^T + E \right) \\ &= \left((C(C^T C)^- \otimes B_{H+1} W_1) + E^T ((C^T C)^- \otimes (X X^T)^-) \right) \left((C^T \otimes X X^T (B_{H+1} W_1)^T) + E \right) \\ &= \left(C(C^T C)^- C^T \otimes F \right) + E^T ((C^T C)^- \otimes (X X^T)^-) E \\ &= \left(C(C^T C)^- C^T \otimes F \right) + \left(r^T X^T (X X^T)^- X r \otimes B_{H+1} (C^T C)^- B_{H+1} \right) \end{aligned} \quad (4)$$

In the third line, the crossed terms $-(C(C^T C)^- \otimes B_{H+1} W_1) E$ and its transpose $-E^T (C(C^T C)^- \otimes B_{H+1} W_1)$ are vanished to 0 because of the following. From Lemma 4.1, $(I_{d_y} \otimes (W_H \cdots W_1 X)^T)^T \text{vec}(r) = 0 \Leftrightarrow W_H \cdots W_1 X r = B_{H+1} W_1 X r = 0$ at any critical point. Thus, $(C(C^T C)^- \otimes B_{H+1} W_1) E = [C(C^T C)^- B_{H+1}^T \otimes B_{H+1} W_1 X r_{:,1} \quad \dots \quad C(C^T C)^- B_{H+1}^T \otimes B_{H+1} W_1 X r_{:,d_y}] = 0$. The forth line

follows

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

where the last line is due to the fact that $\forall t, (r_{\cdot,t})^T X^T (X X^T)^{-1} X r_{\cdot,t}$ is a scalar and the fact that

$$\text{for any matrix } L, r^T L r = \begin{bmatrix} (r_{\cdot,1})^T L r_{\cdot,1} \cdots (r_{\cdot,1})^T L r_{\cdot,dy} \\ \vdots \qquad \qquad \qquad \ddots \qquad \qquad \qquad \vdots \\ (r_{\cdot,dy})^T L r_{\cdot,1} \cdots (r_{\cdot,dy})^T L r_{\cdot,dy} \end{bmatrix}.$$

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

$$((I_{d_y} - C(C^T C)^- C^T) \otimes F) - (r^T X^T (X X^T)^{-1} X r \otimes B_{H+1}(C^T C)^- B_{H+1}) \succeq 0. \quad (5)$$

In the following, we simplify equation 5 by first showing that $\mathcal{R}(C) \subseteq \mathcal{R}(\Sigma)$ and then simplifying $C(C^T C)^- C^T, r^T X^T (X X^T)^{-1} X r, F$ and $B_{H+1}(C^T C)^- B_{H+1}$.

Showing that $\mathcal{R}(C) \subseteq \mathcal{R}(\Sigma)$: Again, using Lemma 4.1 with $k = H + 1$,

$$0 = B_{H+1} W_1 X r \Leftrightarrow F W_{H+1}^T = B_{H+1} W_1 X Y^T \Leftrightarrow W_{H+1}^T = F^- B_{H+1} W_1 X Y^T + (I - F^- F) L,$$

for any arbitrary matrix L . Then,

$$\begin{aligned}
C &= W_{H+1} B_{H+1} \\
&= Y X^T W_1^T B_{H+1}^T F^- B_{H+1} + L^T (I - F F^-) B_{H+1} \\
&= Y X^T W_1^T B_{H+1}^T F^- B_{H+1},
\end{aligned}$$

where the second equality follows the fact that we are conducting the case analysis with the case of $F F^- B_{H+1} = B_{H+1}$ here. Using Lemma 4.1 with $k = 1$,

$$0 = X r W_{H+1} \cdots W_2 \Leftrightarrow W_1 = (C^T C)^- C^T Y X^T (X X^T)^{-1} + (I - (C^T C)^- C^T C) L,$$

for any arbitrary matrix L . Pugging this formula of W_1 into the above,

$$\begin{aligned}
C &= Y X^T ((C^T C)^- C^T Y X^T (X X^T)^{-1} + (I - (C^T C)^- C^T C) L)^T B_{H+1}^T F^- B_{H+1} \\
&= \Sigma C (C^T C)^- B_{H+1}^T F^- B_{H+1}
\end{aligned}$$

where the second line follows Lemma 4.4 with $k = H + 1$ (i.e., $C^T C (C^T C)^- B_{H+1}^T = B_{H+1}^T$). Thus, we have the desired result, $\mathcal{R}(C) \subseteq \mathcal{R}(\Sigma)$.

Simplifying $C(C^T C)^- C^T$: Remember that \bar{p} is the rank of C . To simplify the notation, we rearrange the entries of D and U such that the eigenvalues and eigenvectors selected by the index set $\mathcal{I}_{\bar{p}}$ comes first. That is, $U = [U_{\mathcal{I}_{\bar{p}}} \quad U_{-\mathcal{I}_{\bar{p}}}]$ and $\Lambda = \begin{bmatrix} \Lambda_{\mathcal{I}_{\bar{p}}} & 0 \\ 0 & \Lambda_{-\mathcal{I}_{\bar{p}}} \end{bmatrix}$ where $U_{-\mathcal{I}_{\bar{p}}}$ consists of all the eigenvectors that are not contained in $U_{\mathcal{I}_{\bar{p}}}$, and accordingly $\Lambda_{\mathcal{I}_{\bar{p}}}$ (resp. $\Lambda_{-\mathcal{I}_{\bar{p}}}$) consists of all the eigenvalues that correspond (resp. do not correspond) to the index set $\mathcal{I}_{\bar{p}}$. Since $\mathcal{R}(C) \subseteq \mathcal{R}(\Sigma)$, we can write C in the following form: for some index set $\mathcal{I}_{\bar{p}}$, $C = [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_1$, where $\mathbf{0} \in \mathbb{R}^{d_y \times (d_1 - \bar{p})}$ and $G_1 \in GL_{d_1}(\mathbb{R})$ (a $d_1 \times d_1$ invertible matrix) (notice that $d_1 \geq p \geq \bar{p}$ by their definitions). Then,

$$(C^T C)^- = (G_1^T [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}]^T [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_1)^- = \left(G_1^T \begin{bmatrix} I_{\bar{p}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} G_1 \right)^-.$$

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

$$\left\{ G_1^{-1} \begin{bmatrix} I_{\bar{p}} & L_1 \\ L_2 & L_3 \end{bmatrix} G_1^{-T} \mid L_1, L_2, L_3 \text{ arbitrary} \right\}.$$

Thus, for any arbitrary L_1, L_2 and L_3 ,

$$C(C^T C)^{-1} C^T = C G_1^{-1} \begin{bmatrix} I_{\bar{p}} & L_1 \\ L_2 & L_3 \end{bmatrix} G_1^{-T} C^T = [U_{\mathcal{I}_{\bar{p}}} \quad \mathbf{0}] \begin{bmatrix} I_{\bar{p}} & L_1 \\ L_2 & L_3 \end{bmatrix} \begin{bmatrix} U_{\mathcal{I}_{\bar{p}}}^T \\ \mathbf{0} \end{bmatrix} = U_{\mathcal{I}_{\bar{p}}} U_{\mathcal{I}_{\bar{p}}}^T.$$

Simplifying $r^T X^T (X X^T)^{-1} X r$:

$$\begin{aligned} r^T X^T (X X^T)^{-1} X r &= (C W_1 X - Y) X^T (X X^T)^{-1} X (X^T (C W_1)^T - Y^T) \\ &= C W_1 X X^T (C W_1)^T - C W_1 X Y^T - Y X^T (C W_1)^T + \Sigma \\ &= P_C \Sigma P_C - P_C \Sigma - \Sigma P_C + \Sigma \\ &= \Sigma - U_{\bar{p}} \Lambda_{\mathcal{I}_{\bar{p}}} U_{\bar{p}}^T \end{aligned}$$

where $P_C = C(C^T C)^{-1} C^T = U_{\mathcal{I}_{\bar{p}}} U_{\mathcal{I}_{\bar{p}}}^T$ and the last line follows the facts:

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

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

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 $W_1 = (C^T C)^{-1} C^T Y X^T (X X^T)^{-1} + (I - (C^T C)^{-1} C^T C) L$. Also, from Lemma 4.4, we have that $X r = 0$ or $B_{H+1} (C^T C)^{-1} C^T C = (C^T C (C^T C)^{-1} B_{H+1}^T)^T = B_{H+1}$. If $X r = 0$, we got the statement of the lemma, and so we consider the case of $B_{H+1} (C^T C)^{-1} C^T C = B_{H+1}$. Therefore,

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

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

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

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 $\mathcal{R}(B_{H+1}^T) = \mathcal{R}(C^T C)$. Therefore, $\mathcal{R}(C (C^T C)^{-1} B_{H+1}^T) = \mathcal{R}(C (C^T C)^{-1}) = \mathcal{R}(C) \subseteq \mathcal{R}(\Sigma)$. Accordingly, we can write it in the form, $C (C^T C)^{-1} B_{H+1}^T = [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_2$, where $\mathbf{0} \in \mathbb{R}^{d_y \times (d_1 - \bar{p})}$ and $G_2 \in GL_{d_1}(\mathbb{R})$ (we can write it in the form of $[U_{\mathcal{I}_{\bar{p}}'}, \mathbf{0}] G_2$ for some $\mathcal{I}_{\bar{p}'}$ because of the inclusion $\subseteq \mathcal{R}(\Sigma)$ and $\mathcal{I}_{\bar{p}'} = \mathcal{I}_{\bar{p}}$ because of the equality $= \mathcal{R}(C)$). 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 $B_{H+1} (C^T C)^{-1} B_{H+1}$: From Lemma 4.4, $C^T C (C^T C)^{-1} B_{H+1} = B_{H+1}$ (again since we are done if $X r = 0$). Thus, $B_{H+1} (C^T C)^{-1} B_{H+1} = B_{H+1} (C^T C)^{-1} C^T C (C^T C)^{-1} B_{H+1}^T$. As discussed above, we write $C (C^T C)^{-1} B_{H+1}^T = [U_{\mathcal{I}_{\bar{p}}}, \mathbf{0}] G_2$. Thus,

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

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

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

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

Summarizing the above case analysis, if $\nabla^2 \bar{\mathcal{L}}(W) \succeq 0$ at a critical point, $C(C^T C)^{-1} C^T = U_{\bar{p}} U_{\bar{p}}^T$ or $Xr = 0$. \square

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. \square

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: $\text{rank}(W_H \cdots W_2) = p$ and $d_y \leq p$: Assume that $\text{rank}(W_H \cdots W_2) = p$. 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 $p > \text{rank}(W_{H+1}) \geq \text{rank}(W_H \cdots W_2) = p$ is false.

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

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

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

Suppose the latter: $XrW_{H+1} \cdots W_3 = 0$. Since $\text{rank}(W_{H+1} \cdots W_3) \geq \text{rank}(C) \geq p$ 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: $\text{rank}(C) \geq \text{rank}(I_{d_1})$. Because $d_y = p$, $\text{rank}(C) \geq p$, and $\mathcal{R}(C) \subseteq \mathcal{R}(YX^T)$ as shown in the proof of Lemma 4.6, we have that $\mathcal{R}(C) = \mathcal{R}(YX^T)$.

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

where the last equality follows the fact that $(Xr)^T = C(C^T C)^{-1} C^T YX^T - YX^T = 0$ 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 opposite is also true when $d_y \leq p$ although we don't need it in our proof). That is, $\text{image}(\bar{L}) \subseteq \text{image}(f)$ and $\text{image}(r) \subseteq \text{image}(r')$ (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 $\text{image}(\bar{\mathcal{L}}) \subseteq \text{image}(f)$). 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 $\text{rank}(W_H \cdots W_2) = p$ 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: $\text{rank}(W_H \cdots W_2) = p$ 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 $C(C^T C)^{-1} C^T = U_{\bar{p}} U_{\bar{p}}^T$ 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 $C(C^T C)^{-1} C^T = U_{\bar{p}} U_{\bar{p}}^T$. Combined with Lemma 4.2, we have a necessary condition:

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

From Lemma 4.4 with $k = H + 1$, $\mathcal{R}(W_{H+1}^T \cdots W_1^T) \subseteq \mathcal{R}(C^T C) = \mathcal{R}(C^T)$, which implies that $\bar{p} \triangleq \text{rank}(C) = p$ (since $\text{rank}(W_H \cdots W_2) = p$). Thus, we can rewrite the above equation as $W_{H+1} \cdots W_1 = U_p U_p^T YX^T (XX^T)^{-1}$, 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 $\text{rank}(W_H \cdots W_2) = p$, if $\nabla^2 \bar{\mathcal{L}}(W) \succeq 0$ at a critical point, it is a global minimum.

Case III: $\text{rank}(W_H \cdots W_2) < p$: Suppose that $\text{rank}(W_H \cdots W_2) < p$. Let $\hat{p} = \min(p, d_y)$. Then, if $\text{rank}(C) \geq \hat{p}$, every local minimum is a global minimum because of the following. If $p \leq d_y$, $\text{rank}(W_H \cdots W_2) \geq \text{rank}(C) \geq \hat{p} = p$ and thereby we have the case of $\text{rank}(W_H \cdots W_2) = p$ (since we have that $p \geq \text{rank}(W_H \cdots W_2) \geq p$ where the first inequality follows the definition of p). For this case, we have already proven the desired statement above. On the other hand,

⁵proof: 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 previous two statements results in the desired claim

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

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 ,

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

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

If $C^T C$ is singular, it is inferred that we can perturb W_1 to have $\text{rank}(W_1) \geq \hat{p}$. 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 $C = [U_{\bar{p}} \ 0] G_1$ for some $G_1 \in GL_{d_1}(\mathbb{R})$ where $0 \in \mathbb{R}^{d_y \times (d_1 - \bar{p})}$. 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,

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

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

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

Here, $[0 \ I_{(d_1 - \bar{p})}] G_1 L_1 \in \mathbb{R}^{(d_1 - \bar{p}) \times d_x}$ is the last $(d_1 - \bar{p})$ rows of $G_1 L_1$. Since $\text{rank}(Y X^T (X X^T)^{-1}) = d_y$ (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, $[0 \ I_{(d_1 - \bar{p})}] G_1 L_1$. Since $W_{H+1} \cdots W_1 = C W_1 = [U_{\bar{p}} \ 0] G_1 W_1$,

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

This means that changing the values of the last $(d_1 - \bar{p})$ rows of $G_1 L_1$ (i.e., $[0 \ I_{(d_1 - \bar{p})}] G_1 L_1$) does not change the value of $\mathcal{L}(W)$. Thus, we consider the perturbation of each entry of W_1 as follows:

$$\tilde{W}_1 := W_1 + \epsilon G_1^{-1} \begin{bmatrix} 0 \\ M_{\text{ptb}} \end{bmatrix} = G_1^{-1} \begin{bmatrix} U_{\bar{p}}^T Y X^T (X X^T)^{-1} \\ [0 \ I_{(d_1 - \bar{p})}] G_1 L_1 + \epsilon M_{\text{ptb}} \end{bmatrix}.$$

Here, with an appropriate choice of M_{pth} , we can make \tilde{W}_1 to be full rank (see footnote 6 for the proof of the existence of such M_{pth}).⁶

Thus, we have shown that we can have $\text{rank}(W_1) \geq \min(d_1, d_x) \geq \min(p, d_y) = \hat{p}$ 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 step⁷ with $k \in \{2, \dots, H+1\}$, we have the inductive hypothesis that we can have $\text{rank}(W_{k-1} \cdots W_1) \geq \hat{p}$ with arbitrarily small perturbations of each entry of W_{k-1}, \dots, W_1 without changing the loss value. Here, we want to show that if $\text{rank}(W_{k-1} \cdots W_1) \geq \hat{p}$, we can have $\text{rank}(W_k \cdots W_1) \geq \hat{p}$ with arbitrarily small perturbation of each entry of W_k without changing the value of $\mathcal{L}(W)$. Accordingly, suppose that $\text{rank}(W_{k-1} \cdots W_1) \geq \hat{p}$. 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 $k \in \{2, \dots, H+1\}$,

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

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

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

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,

$$(X X^T)^{-1} X Y^T A_k (A_k^T A_k)^- A_k^T = (X X^T)^{-1} X Y^T U_{\bar{p}} U_{\bar{p}}^T,$$

which implies that

$$X Y^T A_k (A_k^T A_k)^- A_k = X Y^T U_{\bar{p}} U_{\bar{p}}^T.$$

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

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

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 $\mathcal{R}(A_k) = \mathcal{R}(C)$, $\text{rank}(C) = \text{rank}(A_k) = d_k \geq \hat{p} \triangleq \min(p, d_y)$, which is false in the case being analyzed (the case of $\text{rank}(C) < \hat{p}$). 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

⁶In this footnote, we prove the existence of ϵM_{pth} 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 $\begin{bmatrix} U_{\bar{p}}^T Y X^T (X X^T)^{-1} \\ 0 \end{bmatrix} I_{(d_1 - \bar{p})} G_1 L_1$, there exist \bar{p}' linearly independent row vectors including the first \bar{p} row vectors, denoted by $b_1, \dots, b_{\bar{p}'} \in \mathbb{R}^{1 \times d_x}$. Then, we denote the rest of row vectors by $v_1, v_2, \dots, v_{d_1 - \bar{p}'} \in \mathbb{R}^{1 \times d_x}$. Let $c = \min(d_1 - \bar{p}', d_x - \bar{p}')$. There exist linearly independent vectors $\bar{v}_1, \bar{v}_2, \dots, \bar{v}_c$ such that the set, $\{b_1, \dots, b_{\bar{p}'}, \bar{v}_1, \bar{v}_2, \dots, \bar{v}_c\}$, is linearly independent. Setting $v_i := v_i + \epsilon \bar{v}_i$ for all $i \in \{1, \dots, c\}$ 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_{pth} can be obtained by setting ϵM_{pth} to consist of $\epsilon \bar{v}_1, \epsilon \bar{v}_2, \dots, \epsilon \bar{v}_c$ row vectors for the corresponding rows and 0 row vectors for other rows.

⁷The 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 $k < k'$.

$k = H + 1$, since the latter condition, $XrW_{H+1} \cdots W_{k+1} = 0$, implies a false statement, the former condition, $\text{rank}(C) \geq \hat{p}$, 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 $\text{rank}(W_k \cdots W_1) \geq \min(p, d_x)$. To see this in a concrete algebraic way, first note that since $\mathcal{R}(A_k) = \mathcal{R}(U_{\bar{p}})$, we can write $A_k = [U_{\bar{p}} \ 0]G_k$ for some $G_k \in GL_{d_k}(\mathbb{R})$ where $0 \in \mathbb{R}^{d_y \times (d_k - \bar{p})}$. 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 ,

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

Here, $[0 \ I_{(d_k - \bar{p})}]G_k L_k \in \mathbb{R}^{(d_k - \bar{p}) \times d_x}$ is the last $(d_k - \bar{p})$ rows of $G_k L_k$. Since $\text{rank}(Y X^T (X X^T)^{-1}) = d_y$, 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 $W_{H+1} \cdots W_1 = A_k W_k \cdots W_1 = [U_{\bar{p}} \ 0]G_k W_k \cdots W_1$,

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

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

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 ,

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

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

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

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

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

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 $\text{rank}(\tilde{W}_k B_k) \geq \hat{p}$ as follows. Let $\bar{p}' \geq \bar{p}$ be the rank of $\tilde{W}_k B_k$. That is, in $\begin{bmatrix} U_{\bar{p}}^T Y X^T (X X^T)^{-1} \\ 0 \end{bmatrix} \begin{bmatrix} I_{\bar{p}} & 0 \\ 0 & I_{(d_k - \bar{p})} \end{bmatrix} G_k L_k$, there exist \bar{p}' linearly independent row vectors including the first \bar{p} row vectors, denoted by $b_1, \dots, b_{\bar{p}'} \in \mathbb{R}^{1 \times d_x}$. Then, we denote the rest of row vectors by $v_1, v_2, \dots, v_{d_k - \bar{p}'} \in \mathbb{R}^{1 \times d_x}$. Since $\text{rank}(B_k^T B_k) \geq \hat{p}$ (due to the inductive hypothesis), the dimension of $\mathcal{R}(B_k^T B_k)$ is at least \hat{p} . Therefore, there exist vectors $\bar{v}_1, \bar{v}_2, \dots, \bar{v}_{(\bar{p} - \bar{p}')}^T$ such that the set, $\{b_1^T, \dots, b_{\bar{p}'}^T, \bar{v}_1^T, \bar{v}_2^T, \dots, \bar{v}_{(\bar{p} - \bar{p}')}^T\}$, is

linearly independent and $\bar{v}_1^T, \bar{v}_2^T, \dots, \bar{v}_{(\hat{p}-\bar{p})}^T \in \mathcal{R}(B_k^T B_k)$. A desired perturbation matrix M_{ptb} can be obtained by setting M_{ptb} to consist of $\bar{v}_1, \bar{v}_2, \dots, \bar{v}_{(\hat{p}-\bar{p})}$ row vectors for the first $(\hat{p} - \bar{p})$ rows and 0 row vectors for the rest:

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

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

Thus, we conclude the induction, proving that we can have $\text{rank}(W_{H+1} \cdots W_1) \geq \hat{p}$ with arbitrarily small perturbation of each parameter without changing the value of $\tilde{\mathcal{L}}(W)$. Since $\text{rank}(C) \geq \text{rank}(W_{H+1} \cdots W_1) \geq \hat{p}$, upon such a perturbation, we have the case where $\text{rank}(C) \geq \hat{p}$, 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 $\text{rank}(W_H \cdots W_2) < p$.

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)**. \square

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 $\tilde{\mathcal{L}}(W)$. 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 $\text{rank}(W_{H+1}) \geq \text{rank}(W_H \cdots W_2)$ 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, \dots, W_2 . Then, every point in the set defines a critical point from Lemma 4.1. Also, $Xr = XY^T \neq 0$ as $\text{rank}(XY^T) \geq 1$. So, it does not satisfy the first semidefinite condition. On the other hand, with any instance of $W_H \cdots W_2$ such that $\text{rank}(W_H \cdots W_2) \geq 1$, we have that $0 = \text{rank}(W_{H+1}) \not\geq \text{rank}(W_H \cdots W_2)$. 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.**

\square

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 $k = 1$,

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

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 $\text{rank}(W_H \cdots W_2) < p$. Suppose that $W_{H+1} \cdots W_2 = 0$ and thus $\text{rank}(W_{H+1} \cdots W_2) = 0$. By induction on $k = \{2, \dots, H + 1\}$,

we prove that we can have $W_k \cdots W_2 \neq 0$ with arbitrarily small perturbation of each entry of W_k, \dots, 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 $k \in \{2, \dots, H + 1\}$,

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

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

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

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

$$A_2 (A_2^T A_2)^- A_2 = C(C^T C)^- C^T.$$

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

Suppose that $\text{rank}(A_2^T A_2) \geq 1$. Then, since $\mathcal{R}(A_2) = \mathcal{R}(C)$, we have that $\text{rank}(C) \geq 1$, which is false (or else the desired statement). Thus, $\text{rank}(A_2^T A_2) = 0$, 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 $k \in \{2, \dots, H + 1\}$,

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

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

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

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

$$A_k (A_k^T A_k)^- A_k = C(C^T C)^- C^T.$$

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

Suppose that $\text{rank}(A_k^T A_k) \geq 1$. Then, since $\mathcal{R}(A_k) = \mathcal{R}(C)$, we have that $\text{rank}(C) = \text{rank}(A_k) \geq 1$, which is false (or the desired statement). Thus, $\text{rank}(A_k^T A_k) = 0$, 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 $\mathcal{L}(W)$, 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, $\text{rank}(W_H \cdots W_2) = p$, revealed that 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. Thus, when $\text{rank}(W_H \cdots W_2) = p$, 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 $\text{rank}(W_H \cdots W_2) = p$, the Hessian at any saddle point has some negative eigenvalue, which is the statement of Theorem 2.3 (iv). \square

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 $\text{rank}(W_1 \cdots W_2) = \text{rank}(I_{d_1}) = d_1 = p$ ", 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 $\text{rank}(Y) \geq 1$, it is not a global minimum. To see this, we write

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

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

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

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 $\frac{1}{2} \text{tr}(Y Y^T)$ (since $X \neq 0$ and $\text{rank}(\Sigma) \neq 0$).

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. \square

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

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

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 $\bar{Y} = W_2 W_1 X$ 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 $i \in \{1, \dots, H\}$ 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:

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

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

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

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