Supplementary Materials of *Deep Neural Networks with Multi-Branch Architectures Are Intrinsically Less Non-Convex*

Hongyang Zhang
Carnegie Mellon University

Junru Shao
Carnegie Mellon University

Ruslan Salakhutdinov
Carnegie Mellon University

A Supplementary Experiments

A.1 Performance of Multi-Branch Architecture

In this section, we test the classification accuracy of the multi-branch architecture on the CIFAR-10 dataset. We use a 9-layer VGG network [4] as our sub-network in each branch, which is memory-efficient for practitioners to fit many branches into GPU memory simultaneously. The detailed network setup of VGG-9 is in Table 1, where the width of VGG-9 is either 16 or 32. We test the performance of varying numbers of branches in the overall architecture from 4 to 32, with cross-entropy loss. Figure 1 presents the test accuracy on CIFAR-10 as the number of branches increases. It shows that the test accuracy improves monotonously with the increasing number of parallel branches/paths.

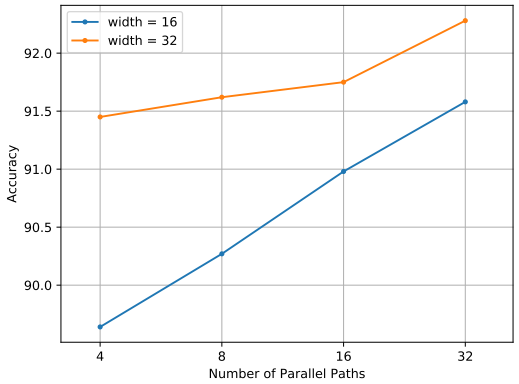

Figure 1: Test accuracy of using VGG-9 as the sub-networks in the multi-branch architecture.

Table 1: Network architecture of VGG-9. Here w is the width of the network, which controls the number of filters in each convolution layer. All convolution layers have a kernel of size 3, and zero padding of size 1. All layers followed by the batch normalization have no bias term. All max pooling layers have a stride of 2.

Layer	**Weight**	**Activation**	**Input size**	**Output size**
Input	N / A	N / A	N / A	$3 \times 32 \times 32$
Conv1	$3 \times 3 \times 3 \times w$	BN + ReLU	$3 \times 32 \times 32$	$w \times 32 \times 32$
Conv2	$3 \times 3 \times w \times w$	BN + ReLU	$w \times 32 \times 32$	$w \times 32 \times 32$
MaxPool	N / A	N / A	$w \times 32 \times 32$	$w \times 16 \times 16$
Conv3	$3 \times 3 \times w \times 2w$	BN + ReLU	$w \times 16 \times 16$	$2w \times 16 \times 16$
Conv4	$3 \times 3 \times 2w \times 2w$	BN + ReLU	$2w \times 16 \times 16$	$2w \times 16 \times 16$
MaxPool	N / A	N / A	$2w \times 16 \times 16$	$2w \times 8 \times 8$
Conv5	$3 \times 3 \times 2w \times 4w$	BN + ReLU	$2w \times 8 \times 8$	$4w \times 8 \times 8$
Conv6	$3 \times 3 \times 4w \times 4w$	BN + ReLU	$4w \times 8 \times 8$	$4w \times 8 \times 8$
Conv7	$3 \times 3 \times 4w \times 4w$	BN + ReLU	$4w \times 8 \times 8$	$4w \times 8 \times 8$
MaxPool	N / A	N / A	$4w \times 8 \times 8$	$4w \times 4 \times 4$
Flatten	N / A	N / A	$4w \times 4 \times 4$	$64w$
FC1	$64w \times 4w$	BN + ReLU	$64w$	$4w$
FC2	$4w \times 10$	Softmax	$4w$	10

A.2 Strong Duality of Deep Linear Neural Networks

We compare the optima of primal problem (4) and dual problem (5) by numerical experiments for three-layer linear neural networks ($H = 3$). The data are generated as follows. We construct the output matrix $\mathbf{Y} \in \mathbb{R}^{100\times 100}$ by drawing the entries of $\mathbf{Y}$ from i.i.d. standard Gaussian distribution and the input matrix $\mathbf{X} \in \mathbb{R}^{100\times 100}$ by the identity matrix. The $d_{\min}$ varies from 5 to 50. Both primal and dual problems are solved by numerical algorithms. Given the non-convex nature of primal problem, we rerun the algorithm by multiple initializations and choose the best solution that we obtain. The results are shown in Figure 2. We can easily see that the optima of primal and dual problems almost match. The small gap is due to the numerical inaccuracy.

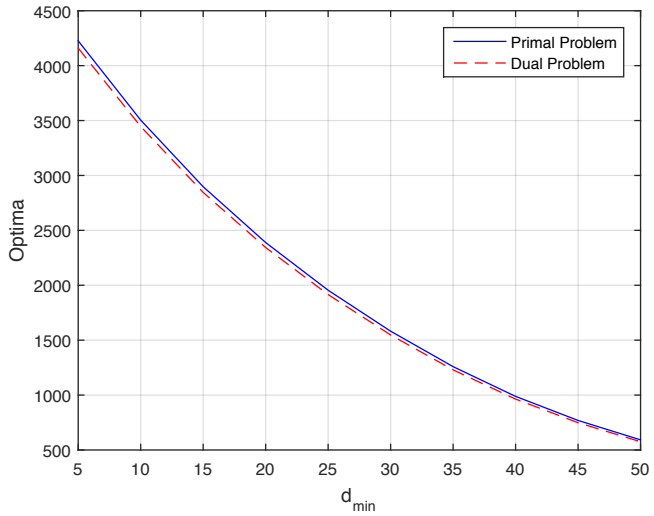

Figure 2: Comparison of optima between primal and dual problems.

We also compare the ℓ_2 distance between the solution $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\dots\mathbf{W}_1^*$ of primal problem and the solution $\mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$ of dual problem in Table 2. We see that the solutions are close to each other.

Table 2: Comparison of the ℓ_2 distance between the solutions of primal and dual problems.

$d_{\min}$	5	10	15	20	25	30	35	40	45	50
ℓ_2 distance ($\times 10^{-10}$)	1.95	1.26	7.89	3.80	3.14	1.92	1.04	3.92	6.53	8.00

B Proofs of Theorem 1: Duality Gap of Multi-Branch Neural Networks

The lower bound $0 \le \frac{\inf(\mathbf{P})-\sup(\mathbf{D})}{\Delta_{worst}}$ is obvious by the weak duality. So we only need to prove the upper bound $\frac{\inf(\mathbf{P})-\sup(\mathbf{D})}{\Delta_{worst}} \le \frac{2}{I}$.

Consider the subset of $\mathbb{R}^2$:

$$\mathcal{Y}_i := \left\{ \mathbf{y}_i \in \mathbb{R}^2 : \mathbf{y}_i = \frac{1}{I}\left[h_i(\mathbf{w}_{(i)}), \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1 - \frac{y\cdot f_i(\mathbf{w}_{(i)};\mathbf{x})}{\tau}\right)\right], \mathbf{w}_{(i)} \in \mathcal{W}_i \right\}, \quad i \in [I].$$

Define the vector summation

$$\mathcal{Y} := \mathcal{Y}_1 + \mathcal{Y}_2 + ... + \mathcal{Y}_I.$$

Since $f_i(\mathbf{w}_{(i)};\mathbf{x})$ and $h_i(\mathbf{w}_{(i)})$ are continuous w.r.t. $\mathbf{w}_{(i)}$ and $\mathcal{W}_i$'s are compact, the set

$$\{(\mathbf{w}_{(i)}, h_i(\mathbf{w}_{(i)}), f_i(\mathbf{w}_{(i)};\mathbf{x})) : \mathbf{w}_{(i)} \in \mathcal{W}_i\}$$

is compact as well. So $\mathcal{Y}$, $\mathsf{conv}(\mathcal{Y})$, $\mathcal{Y}_i$, and $\mathsf{conv}(\mathcal{Y}_i)$, $i \in [I]$ are all compact sets. According to the definition of $\mathcal{Y}$ and the standard duality argument [3], we have

$$\inf(\mathbf{P}) = \min\{w : \text{there exists } (r,w) \in \mathcal{Y} \text{ such that } r \le K\},$$

and

$$\sup(\mathbf{D}) = \min\{w : \text{there exists } (r, w) \in \mathsf{conv}(\mathcal{Y}) \text{ such that } r \le K\}.$$

Technique (a): Shapley-Folkman Lemma. We are going to apply the following Shapley-Folkman lemma.

Lemma 1 (Shapley-Folkman, [5]). *Let $\mathcal{Y}_i, i \in [I]$ be a collection of subsets of $\mathbb{R}^m$. Then for every $\mathbf{y} \in \mathsf{conv}(\sum_{i=1}^I \mathcal{Y}_i)$, there is a subset $\mathcal{I}(\mathbf{y}) \subseteq [I]$ of size at most m such that*

$$\mathbf{y} \in \left[\sum_{i \notin \mathcal{I}(\mathbf{y})} \mathcal{Y}_i + \sum_{i \in \mathcal{I}(\mathbf{y})} \mathsf{conv}(\mathcal{Y}_i)\right].$$

We apply Lemma 1 to prove Theorem 1 with $m = 2$. Let $(\overline{r}, \overline{w}) \in \mathsf{conv}(\mathcal{Y})$ be such that

$$\overline{r} \le K, \quad \text{and} \quad \overline{w} = \sup(\mathbf{D}).$$

Applying the above Shapley-Folkman lemma to the set $\mathcal{Y} = \sum_{i=1}^I \mathcal{Y}_i$, we have that there are a subset $\overline{\mathcal{I}} \subseteq [I]$ of size 2 and vectors

$$(\overline{r}_i, \overline{w}_i) \in \mathsf{conv}(\mathcal{Y}_i), \quad i \in \overline{\mathcal{I}} \qquad \text{and} \qquad \overline{\mathbf{w}}_{(i)} \in \mathcal{W}_i, \quad i \notin \overline{\mathcal{I}},$$

such that

$$\frac{1}{I}\sum_{i \notin \overline{\mathcal{I}}} h_i(\overline{\mathbf{w}}_{(i)}) + \sum_{i \in \overline{\mathcal{I}}} \overline{r}_i = \overline{r} \le K, \tag{1}$$

$$\frac{1}{I}\sum_{i \notin \overline{\mathcal{I}}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1 - \frac{y \cdot f_i(\overline{\mathbf{w}}_{(i)}; \mathbf{x})}{\tau}\right) + \sum_{i \in \overline{\mathcal{I}}} \overline{w}_i = \sup(\mathbf{D}). \tag{2}$$

Representing elements of the convex hull of $\mathcal{Y}_i \subseteq \mathbb{R}^2$ by Carathéodory theorem, we have that for each $i \in \overline{\mathcal{I}}$, there are vectors $\mathbf{w}_{(i)}^1, \mathbf{w}_{(i)}^2, \mathbf{w}_{(i)}^3 \in \mathcal{W}_i$ and scalars $a_i^1, a_i^2, a_i^3 \in \mathbb{R}$ such that

$$\sum_{j=1}^3 a_i^j = 1, \quad a_i^j \ge 0, \ j = 1, 2, 3,$$

$$\overline{r}_i = \frac{1}{I}\sum_{j=1}^3 a_i^j h_i(\mathbf{w}_{(i)}^j), \qquad \overline{w}_i = \frac{1}{I}\sum_{j=1}^3 a_i^j \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}^j; \mathbf{x})}{\tau}\right).$$

Recall that we define

$$\widehat{f}_i(\widetilde{\mathbf{w}}) := \inf_{\mathbf{w}_{(i)} \in \mathcal{W}_i}\left\{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}; \mathbf{x})}{\tau}\right) : h_i(\mathbf{w}_{(i)}) \le h_i(\widetilde{\mathbf{w}})\right\}, \tag{3}$$

$$\widetilde{f}_i(\widetilde{\mathbf{w}}) := \inf_{a^j, \mathbf{w}_{(i)}^j \in \mathcal{W}_i}\left\{\sum_{j=1}^{p_i+2} a^j \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}^j; \mathbf{x})}{\tau}\right) : \widetilde{\mathbf{w}} = \sum_{j=1}^{p_i+2} a^j \mathbf{w}_{(i)}^j, \sum_{j=1}^{p_i+2} a^j = 1, a^j \ge 0\right\},$$

and $\Delta_i := \sup_{\mathbf{w} \in \mathcal{W}_i}\left\{\widehat{f}_i(\mathbf{w}) - \widetilde{f}_i(\mathbf{w})\right\} \ge 0$. We have for $i \in \overline{\mathcal{I}}$,

$$\overline{r}_i \ge \frac{1}{I} h_i\left(\sum_{j=1}^3 a_i^j \mathbf{w}_{(i)}^j\right), \quad \text{(because } h_i(\cdot) \text{ is convex)} \tag{4}$$

and

$$\begin{aligned} \overline{w}_i &\ge \frac{1}{I}\widetilde{f}_i\left(\sum_{j=1}^3 a_i^j \mathbf{w}_{(i)}^j\right) \quad \text{(by the definition of } \widetilde{f}_i(\cdot)) \\ &\ge \frac{1}{I}\widehat{f}_i\left(\sum_{j=1}^3 a_i^j \mathbf{w}_{(i)}^j\right) - \frac{1}{I}\Delta_i. \quad \text{(by the definition of } \Delta_i) \end{aligned} \tag{5}$$

Thus, by Eqns. (1) and (4), we have

$$\frac{1}{I}\sum_{i\notin\overline{\mathcal{I}}} h_i(\overline{\mathbf{w}}_{(i)}) + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}} h_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) \le K, \tag{6}$$

and by Eqns. (2) and (5), we have

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left[\frac{1}{I}\sum_{i\notin\overline{\mathcal{I}}}\left(1-\frac{y\cdot f_i(\overline{\mathbf{w}}_{(i)};\mathbf{x})}{\tau}\right)\right] + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}}\widehat{f}_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) \le \sup(\mathbf{D}) + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}}\Delta_i. \tag{7}$$

Given any $\epsilon > 0$ and $i \in \overline{\mathcal{I}}$, we can find a vector $\overline{\mathbf{w}}_{(i)} \in \mathcal{W}_i$ such that

$$h_i(\overline{\mathbf{w}}_{(i)}) \le h_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) \text{ and } \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1-\frac{y\cdot f_i(\overline{\mathbf{w}}_{(i)};\mathbf{x})}{\tau}\right) \le \widehat{f}_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) + \epsilon, \tag{8}$$

where the first inequality holds because $\mathcal{W}_i$ is convex and the second inequality holds by the definition (3) of $\widehat{f}_i(\cdot)$. Therefore, Eqns. (6) and (8) impliy that

$$\frac{1}{I}\sum_{i=1}^{I} h_i(\overline{\mathbf{w}}_{(i)}) \le K.$$

Namely, $(\overline{\mathbf{w}}_{(1)}, ..., \overline{\mathbf{w}}_{(I)})$ is a feasible solution of problem (2). Also, Eqns. (7) and (8) yield

$$\begin{aligned}
\inf(\mathbf{P}) &\le \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left[\frac{1}{I}\sum_{i=1}^{I}\left(1-\frac{y\cdot f_i(\overline{\mathbf{w}}_{(i)};\mathbf{x})}{\tau}\right)\right] \\
&\le \sup(\mathbf{D}) + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}}(\Delta_i+\epsilon) \\
&\le \sup(\mathbf{D}) + \frac{2}{I}\Delta_{worst} + 2\epsilon,
\end{aligned}$$

where the last inequality holds because $|\overline{\mathcal{I}}| = 2$. Finally, letting $\epsilon \to 0$ leads to the desired result.

C Proofs of Theorem 2: Strong Duality of Deep Linear Neural Networks

Let $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{X}^\dagger\mathbf{X}$. We note that by Pythagorean theorem, for every $\mathbf{Y}$,

$$\frac{1}{2}\|\mathbf{Y}-\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 = \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \underbrace{\frac{1}{2}\|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2}_{\text{independent of } \mathbf{W}_1,\ldots,\mathbf{W}_H}.$$

So we can focus on the following optimization problem instead of problem (4):

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{W}_H...\mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right]. \tag{9}$$

Technique (b): Variational Form. Our work is inspired by a variational form of problem (9) given by the following lemma.

Lemma 2. *If $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ is optimal to problem*

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1,\ldots,\mathbf{W}_H) := \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \gamma\|\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_*, \tag{10}$$

then $(\mathbf{W}_1^{**}, \ldots, \mathbf{W}_H^{**})$ is optimal to problem (9), where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the skinny SVD of $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}$, $\mathbf{W}_i^{**} = [\mathbf{\Sigma}^{1/H}, \mathbf{0}; \mathbf{0}, \mathbf{0}] \in \mathbb{R}^{d_i\times d_{i-1}}$ for $i = 2, 3, ..., H-1$, $\mathbf{W}_H^{**} = [\mathbf{U}\mathbf{\Sigma}^{1/H}, \mathbf{0}] \in \mathbb{R}^{d_H\times d_{H-2}}$ and $\mathbf{W}_1^{**} = [\mathbf{\Sigma}^{1/H}\mathbf{V}^T; \mathbf{0}]\mathbf{X}^\dagger \in \mathbb{R}^{d_1\times d_0}$. Furthermore, problems (9) and (10) have the same optimal objective function value.

Proof of Lemma 2. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the skinny SVD of matrix $\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X} =: \mathbf{Z}$. We notice that

$$\begin{aligned}\|\mathbf{Z}\|_* &= \|\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_* \\ &\le \|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}\prod_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H} \quad \text{(by the generalized Hölder's inequality)} \\ &\le \frac{1}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right]. \quad \text{(by the inequality of mean)}\end{aligned}$$

Hence, on one hand, for every $(\mathbf{W}_1,\ldots,\mathbf{W}_H)$,

$$\begin{aligned}\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1,\ldots,\mathbf{W}_H) &\le \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \gamma\|\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_* \\ &\le \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right],\end{aligned}$$

which yields

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1,\ldots,\mathbf{W}_H) \le \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H}\frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right].$$

On the other hand, suppose $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ is optimal to problem (10), and let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the skinny SVD of matrix $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}$. We choose $(\mathbf{W}_1^{**},\ldots,\mathbf{W}_H^{**})$ such that

$$\mathbf{W}_H^{**} = [\mathbf{U}\mathbf{\Sigma}^{\frac{1}{H}},\mathbf{0}],\ \mathbf{W}_1^{**}\mathbf{X} = [\mathbf{\Sigma}^{\frac{1}{H}}\mathbf{V}^T;\mathbf{0}],\ \mathbf{W}_i^{**} = [\mathbf{\Sigma}^{\frac{1}{H}},\mathbf{0};\mathbf{0},\mathbf{0}],\ i=2,\ldots,H-1.$$

We pad $\mathbf{0}$ around $\mathbf{W}_i^{**}$ so as to adapt to the dimensionality of each $\mathbf{W}_i^{**}$. Notice that

$$\begin{aligned}\|\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}\|_* &= \|\mathbf{W}_H^{**}\mathbf{W}_{H-1}^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}\|_* \\ &= \frac{1}{H}\left[\|\mathbf{W}_1^{**}\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i^{**}\|_{\mathcal{S}_H}^H\right].\end{aligned}$$

Since $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathbf{W}_H^{**}\mathbf{W}_{H-1}^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}$, for every $\widetilde{\mathbf{Y}}$,

$$\|\widetilde{\mathbf{Y}} - \mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}\|_F = \|\widetilde{\mathbf{Y}} - \mathbf{W}_H^{**}\mathbf{W}_{H-1}^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}\|_F.$$

Hence

$$\begin{aligned}&\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1,\ldots,\mathbf{W}_H) = F(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*) = F(\mathbf{W}_1^{**},\ldots,\mathbf{W}_H^{**}) \\ &= \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1^{**}\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i^{**}\|_{\mathcal{S}_H}^H\right] \\ &\ge \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H}\frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right],\end{aligned}$$

which yields the other direction of the inequality and hence completes the proof. □

Technique (c): Reduction to Low-Rank Approximation. We now reduce problem (10) to the classic problem of low-rank approximation of the form $\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H}\frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2$, which has the following nice properties.

Lemma 3. *For any $\widehat{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$, every global minimum $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ of function*

$$f(\mathbf{W}_1,\ldots,\mathbf{W}_H) = \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2$$

obeys $\mathbf{W}_H^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widehat{\mathbf{Y}})$. Here $\widehat{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$ means the row vectors of $\widehat{\mathbf{Y}}$ belongs to the row space of $\mathbf{X}$.

Proof of Lemma 3. Note that the optimal solution to $\min_{\mathbf{W}_H,\ldots,\mathbf{W}_1} \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2$ is equal to the optimal solution to the low-rank approximation problem $\min_{\mathsf{rank}(\mathbf{Z})\le d_{\min}} \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{Z}\|_F^2$ when $\widehat{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$, which has a closed-form solution $\mathsf{svd}_{d_{\min}}(\widehat{\mathbf{Y}})$.[1] □

We now reduce $F(\mathbf{W}_1,\ldots,\mathbf{W}_H)$ to the form of $\frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2$ for some $\widehat{\mathbf{Y}}$ plus an extra additive term that is independent of $(\mathbf{W}_1,\ldots,\mathbf{W}_H)$. To see this, denote by $K(\cdot) = \gamma\|\cdot\|_*$. We have

$$\begin{aligned}
F(\mathbf{W}_1,\ldots,\mathbf{W}_H) &= \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + K^{**}(\mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}) \\
&= \max_{\mathbf{\Lambda}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + \langle \mathbf{\Lambda}, \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\rangle - K^*(\mathbf{\Lambda}) \\
&= \max_{\mathbf{\Lambda}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle \widetilde{\mathbf{Y}}, \mathbf{\Lambda}\rangle \\
&=: \max_{\mathbf{\Lambda}} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}),
\end{aligned}$$

where we define $L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}) := \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle \widetilde{\mathbf{Y}}, \mathbf{\Lambda}\rangle$ as the Lagrangian of problem (10). The first equality holds because $K(\cdot)$ is closed and convex w.r.t. the argument $\mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}$ so $K(\cdot) = K^{**}(\cdot)$, and the second equality is by the definition of conjugate function. One can check that $L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}) = \min_{\mathbf{M}} L'(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{M},\mathbf{\Lambda})$, where $L'(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{M},\mathbf{\Lambda})$ is the Lagrangian of the constraint optimization problem $\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{M}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + K(\mathbf{M})$, s.t. $\mathbf{M} = \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}$. With a little abuse of notation, we call $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ the Lagrangian of the unconstrained problem (10) as well.

The remaining analysis is to choose a proper $\mathbf{\Lambda}^* \in \mathsf{Row}(\mathbf{X})$ such that $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda}^*)$ is a primal-dual saddle point of $L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda})$, so that the problem $\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}^*)$ and problem (10) have the same optimal solution $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$. For this, we introduce the following condition, and later we will show that the condition holds.

Condition 1. *For a solution $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ to optimization problem (10), there exists an*

$$\mathbf{\Lambda}^* \in \partial_{\mathbf{Z}} K(\mathbf{Z})|_{\mathbf{Z}=\mathbf{W}_H^*\cdots\mathbf{W}_1^*\mathbf{X}} \cap \mathsf{Row}(\mathbf{X})$$

such that

$$\begin{aligned}
\mathbf{W}_{i+1}^{*T}\cdots\mathbf{W}_H^{*T}(\mathbf{W}_H^*\cdots\mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}})\mathbf{X}^T\mathbf{W}_1^{*T}\cdots\mathbf{W}_{i-1}^{*T} &= \mathbf{0}, \quad i = 2,\ldots,H-1, \\
\mathbf{W}_2^{*T}\cdots\mathbf{W}_H^{*T}(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}})\mathbf{X}^T &= \mathbf{0}, \\
(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}})\mathbf{X}^T\mathbf{W}_1^{*T}\cdots\mathbf{W}_{H-1}^{*T} &= \mathbf{0}.
\end{aligned} \tag{11}$$

We note that if we set $\mathbf{\Lambda}$ to be the $\mathbf{\Lambda}^*$ in (11), then $\nabla_{\mathbf{W}_i} L(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda}^*) = \mathbf{0}$ for every i. So $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ is either a saddle point, a local minimizer, or a global minimizer of $L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}^*)$ as a function of $(\mathbf{W}_1,\ldots,\mathbf{W}_H)$ for the fixed $\mathbf{\Lambda}^*$. The following lemma states that if it is a global minimizer, then strong duality holds.

Lemma 4. *Let $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ be a global minimizer of $F(\mathbf{W}_1,\ldots,\mathbf{W}_H)$. If there exists a dual certificate $\mathbf{\Lambda}^*$ satisfying Condition 1 and the pair $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ is a global minimizer of $L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}^*)$ for the fixed $\mathbf{\Lambda}^*$, then strong duality holds. Moreover, we have the relation $\mathbf{W}_H^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$.*

Proof of Lemma 4. By the assumption of the lemma, $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ is a global minimizer of

$$L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}^*) = \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^* - \mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + c(\mathbf{\Lambda}^*),$$

where $c(\mathbf{\Lambda}^*)$ is a function of $\mathbf{\Lambda}^*$ that is independent of $\mathbf{W}_i$ for all i's. Namely, $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ globally minimizes $L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda})$ when $\mathbf{\Lambda}$ is fixed to $\mathbf{\Lambda}^*$. Furthermore, $\mathbf{\Lambda}^* \in \partial_{\mathbf{Z}} K(\mathbf{Z})|_{\mathbf{Z}=\mathbf{W}_H^*\cdots\mathbf{W}_1^*\mathbf{X}}$

[1] Note that the low-rank approximation problem might have non-unique solution. However, we will use in this paper the abuse of language $\mathsf{svd}_{d_{\min}}(\widehat{\mathbf{Y}})$ as the non-uniqueness issue does not lead to any issue in our developments.

implies that $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} \in \partial_{\mathbf{\Lambda}} K^*(\mathbf{\Lambda})|_{\mathbf{\Lambda}=\mathbf{\Lambda}^*}$ by the convexity of function $K(\cdot)$, meaning that $\mathbf{0} \in \partial_{\mathbf{\Lambda}} L(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda})$. So $\mathbf{\Lambda}^* = \operatorname{argmax}_{\mathbf{\Lambda}} L(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda})$ due to the concavity of function $L(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda})$ w.r.t. variable $\mathbf{\Lambda}$. Thus $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda}^*)$ is a primal-dual saddle point of $L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda})$.

We now prove the strong duality. By the fact that $F(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*) = \max_{\mathbf{\Lambda}} L(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda})$ and that $\mathbf{\Lambda}^* = \operatorname{argmax}_{\mathbf{\Lambda}} L(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda})$, for every $\mathbf{W}_1,\ldots,\mathbf{W}_H$, we have

$$F(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*) = L(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda}^*) \le L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}^*),$$

where the inequality holds because $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*,\mathbf{\Lambda}^*)$ is a primal-dual saddle point of L. Notice that we also have

$$\begin{aligned}\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H}\max_{\mathbf{\Lambda}} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}) &= F(\mathbf{W}_1^*,\ldots\mathbf{W}_H^*)\\ &\le \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}^*)\\ &\le \max_{\mathbf{\Lambda}}\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}).\end{aligned}$$

On the other hand, by weak duality,

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H}\max_{\mathbf{\Lambda}} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}) \ge \max_{\mathbf{\Lambda}}\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}).$$

Therefore,

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H}\max_{\mathbf{\Lambda}} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}) = \max_{\mathbf{\Lambda}}\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}),$$

i.e., strong duality holds. Hence,

$$\begin{aligned}\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^* &= \operatorname*{argmin}_{\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}^*)\\ &= \operatorname*{argmin}_{\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1} \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}^*-\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}^*\|_F^2 - K^*(\mathbf{\Lambda}^*) + \langle\widetilde{\mathbf{Y}},\mathbf{\Lambda}^*\rangle\\ &= \operatorname*{argmin}_{\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1} \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}^*-\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2\\ &= \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}}-\mathbf{\Lambda}^*).\end{aligned}$$

The proof of Lemma 4 is completed. □

Technique (d): Dual Certificate. We now construct dual certificate $\mathbf{\Lambda}^*$ such that all of conditions in Lemma 4 hold. We note that $\mathbf{\Lambda}^*$ should satisfy the followings by Lemma 4:

$$\begin{aligned}&\text{(a)}\quad \mathbf{\Lambda}^* \in \partial K(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}) \cap \mathsf{Row}(\mathbf{X}); \qquad \text{(by Condition 1)}\\ &\text{(b)}\quad \text{Equations (11)}; \qquad \text{(by Condition 1)}\\ &\text{(c)}\quad \mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}}-\mathbf{\Lambda}^*). \quad \text{(by the global optimality and Lemma 3)}\end{aligned} \tag{12}$$

Before proceeding, we denote by $\widetilde{\mathbf{A}} := \mathbf{W}_H^*\cdots\mathbf{W}_{\min+1}^*$, $\widetilde{\mathbf{B}} := \mathbf{W}_{\min}^*\cdots\mathbf{W}_1^*\mathbf{X}$, where $\mathbf{W}_{\min}^*$ is a matrix among $\{\mathbf{W}_i^*\}_{i=1}^{H-1}$ which has $d_{\min}$ rows, and let

$$\mathcal{T} := \{\widetilde{\mathbf{A}}\mathbf{C}_1^T + \mathbf{C}_2\widetilde{\mathbf{B}} :\ \mathbf{C}_1 \in \mathbb{R}^{n\times d_{\min}},\ \mathbf{C}_2 \in \mathbb{R}^{d_H\times d_{\min}}\}$$

be a matrix space. Denote by $\mathcal{U}$ the left singular space of $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ and $\mathcal{V}$ the right singular space. Then the linear space $\mathcal{T}$ can be equivalently represented as $\mathcal{T} = \mathcal{U}+\mathcal{V}$. Therefore, $\mathcal{T}^\perp = (\mathcal{U}+\mathcal{V})^\perp = \mathcal{U}^\perp\cap\mathcal{V}^\perp$. With this, we note that: (b) $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}+\mathbf{\Lambda}^*-\widetilde{\mathbf{Y}} \in \mathsf{Null}(\widetilde{\mathbf{A}}^T) = \mathsf{Col}(\widetilde{\mathbf{A}})^\perp$ and $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}+\mathbf{\Lambda}^*-\widetilde{\mathbf{Y}} \in \mathsf{Row}(\widetilde{\mathbf{B}})^\perp$ (so $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}+\mathbf{\Lambda}^*-\widetilde{\mathbf{Y}} \in \mathcal{T}^\perp$) imply Equations (11) since either $\mathbf{W}_{i+1}^{*T}\cdots\mathbf{W}_H^{*T}(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}+\mathbf{\Lambda}^*-\widetilde{\mathbf{Y}}) = \mathbf{0}$ or $(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}+\mathbf{\Lambda}^*-\widetilde{\mathbf{Y}})\mathbf{X}^T\mathbf{W}_1^{*T}\cdots\mathbf{W}_{i-1}^{*T} = \mathbf{0}$ for all i's. And (c) for an orthogonal decomposition $\widetilde{\mathbf{Y}}-\mathbf{\Lambda}^* = \mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}+\mathbf{E}$ where $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} \in \mathcal{T}$ and $\mathbf{E} \in \mathcal{T}^\perp$, we have that

$$\|\mathbf{E}\| \le \sigma_{d_{\min}}(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X})$$

and condition (b) together imply $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$ by Lemma 3. Therefore, the dual conditions in (12) are implied by

$$\begin{aligned}
&(1)\quad \mathbf{\Lambda}^* \in \partial K(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}) \cap \mathsf{Row}(\mathbf{X});\\
&(2)\quad \mathcal{P}_{\mathcal{T}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*) = \mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X};\\
&(3)\quad \|\mathcal{P}_{\mathcal{T}^\perp}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)\| \le \sigma_{d_{\min}}(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}).
\end{aligned}$$

It thus suffices to construct a dual certificate $\mathbf{\Lambda}^*$ such that conditions (1), (2) and (3) hold, because conditions (1), (2) and (3) are stronger than conditions (a), (b) and (c). Let $r = \mathsf{rank}(\widetilde{\mathbf{Y}})$ and $\bar{r} = \min\{r, d_{\min}\}$. To proceed, we need the following lemma.

Lemma 5 ([6]). *Suppose $\widetilde{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$. Let $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ be the solution to problem (10) and let $\mathbf{U}\mathsf{diag}(\sigma_1(\widetilde{\mathbf{Y}}), \ldots, \sigma_r(\widetilde{\mathbf{Y}}))\mathbf{V}^T$ denote the skinny SVD of $\widetilde{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$. We have $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathbf{U}\mathsf{diag}((\sigma_1(\widetilde{\mathbf{Y}}) - \gamma)_+, \ldots, (\sigma_{\bar{r}}(\widetilde{\mathbf{Y}}) - \gamma)_+, 0, \ldots, 0)\mathbf{V}^T$.*

Recall that the sub-differential of the nuclear norm of a matrix $\mathbf{Z}$ is

$$\partial_{\mathbf{Z}}\|\mathbf{Z}\|_* = \{\mathbf{U}_{\mathbf{Z}}\mathbf{V}_{\mathbf{Z}}^T + \mathbf{T}_{\mathbf{Z}} : \mathbf{T}_{\mathbf{Z}} \in \mathcal{T}^\perp, \|\mathbf{T}_{\mathbf{Z}}\| \le 1\},$$

where $\mathbf{U}_{\mathbf{Z}}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{V}_{\mathbf{Z}}^T$ is the skinny SVD of the matrix $\mathbf{Z}$. So with Lemma 5, the sub-differential of (scaled) nuclear norm at optimizer $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}$ is given by

$$\partial(\gamma\|\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}\|_*) = \{\gamma\mathbf{U}_{:,1:\bar{r}}\mathbf{V}_{:,1:\bar{r}}^T + \mathbf{T} : \mathbf{T} \in \mathcal{T}^\perp, \|\mathbf{T}\| \le \gamma\}. \tag{13}$$

To construct the dual certificate, we set

$$\mathbf{\Lambda}^* = \underbrace{\gamma\mathbf{U}_{:,1:\bar{r}}\mathbf{V}_{:,1:\bar{r}}^T}_{\text{Component in space } \mathcal{T}} + \underbrace{\mathbf{U}_{:,(\bar{r}+1):r}\mathsf{diag}(\gamma, \ldots, \gamma)\mathbf{V}_{:,(\bar{r}+1):r}^T}_{\text{Component } \mathbf{T} \text{ in space } \mathcal{T}^\perp \text{ with } \|\mathbf{T}\| \le \gamma} \in \mathsf{Row}(\mathbf{X}),$$

where $\mathbf{\Lambda}^* \in \mathsf{Row}(\mathbf{X})$ because $\mathbf{V}^T \in \mathsf{Row}(\mathbf{X})$ (This is because $\mathbf{V}^T$ is the right singular matrix of $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$). So condition (1) is satisfied according to (13). To see condition (2), $\mathcal{P}_{\mathcal{T}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*) = \mathcal{P}_{\mathcal{T}}\widetilde{\mathbf{Y}} - \gamma\mathbf{U}_{:,1:\bar{r}}\mathbf{V}_{:,1:\bar{r}}^T = \mathbf{U}\mathsf{diag}((\sigma_1(\widetilde{\mathbf{Y}}) - \gamma)_+, \ldots, (\sigma_{\bar{r}}(\widetilde{\mathbf{Y}}) - \gamma)_+, 0, 0, \ldots, 0)\mathbf{V}^T = \mathbf{W}_H^*\mathbf{W}_{H-1}^*\ldots\mathbf{W}_1^*\mathbf{X}$, where the last equality is by Lemma 5 and the assumption $\sigma_{\min}(\widetilde{\mathbf{Y}}) > \gamma$. As for condition (3), note that

$$\begin{aligned}
\left\|\mathcal{P}_{\mathcal{T}^\perp}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)\right\| &= \left\|\mathbf{U}_{:,(\bar{r}+1):r}\mathsf{diag}(\sigma_{\bar{r}+1}(\widetilde{\mathbf{Y}}) - \gamma, \ldots, \sigma_r(\widetilde{\mathbf{Y}}) - \gamma)\mathbf{V}_{:,(\bar{r}+1):r}^T\right\|\\
&= \begin{cases} 0, & \text{if } \bar{r} = r,\\ \sigma_{d_{\min}+1}(\widetilde{\mathbf{Y}}) - \gamma, & \text{otherwise.}\end{cases}
\end{aligned}$$

By Lemma 5, $\sigma_{d_{\min}}(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}) \ge \|\mathcal{P}_{\mathcal{T}^\perp}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)\|$. So the proof of strong duality is completed, where the dual problem is given in Section D.

To see the relation between the solutions of primal and dual problems, it is a direct result of Lemmas 2 and 4.

D Dual Problem of Deep Linear Neural Network

In this section, we derive the dual problem of non-convex program (4). Denote by $G(\mathbf{W}_1, \ldots, \mathbf{W}_H)$ the objective function of problem (4). Let $K(\cdot) = \gamma\|\cdot\|_*$, and let $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{X}^\dagger\mathbf{X}$ be the projection of $\mathbf{Y}$ on the row span of $\mathbf{X}$.

We note that

$$\begin{aligned}
&\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} G(\mathbf{W}_1,\ldots,\mathbf{W}_H) - \frac{1}{2}\|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2 \\
=&\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\mathbf{Y}-\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2 + K(\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}) \\
=&\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + K^{**}(\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}) \\
=&\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H}\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \langle\mathbf{\Lambda},\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\rangle - K^*(\mathbf{\Lambda}) \\
=&\min_{\mathbf{W}_1,\cdots,\mathbf{W}_H}\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}-\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle\widetilde{\mathbf{Y}},\mathbf{\Lambda}\rangle,
\end{aligned}$$

where the second equality holds since $K(\cdot)$ is closed and convex w.r.t. the argument $\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}$ and the third equality is by the definition of conjugate function of nuclear norm. Therefore, the dual problem is given by

$$\begin{aligned}
&\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})}\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}-\mathbf{W}_H\ldots\mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle\widetilde{\mathbf{Y}},\mathbf{\Lambda}\rangle + \frac{1}{2}\|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2 \\
=&\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} \frac{1}{2}\sum_{i=d_{\min}+1}^{\min\{d_H,n\}}\sigma_i^2(\widetilde{\mathbf{Y}}-\mathbf{\Lambda}) - \frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \frac{1}{2}\|\mathbf{Y}\|_F^2 \\
=&\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} -\frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}\|_{d_{\min}}^2 - K^*(\mathbf{\Lambda}) + \frac{1}{2}\|\mathbf{Y}\|_F^2,
\end{aligned}$$

where $\|\cdot\|_{d_{\min}}^2 = \sum_{i=1}^{d_{\min}}\sigma_i^2(\cdot)$. We note that

$$K^*(\mathbf{\Lambda}) = \begin{cases} 0, & \|\mathbf{\Lambda}\| \le \gamma; \\ +\infty, & \|\mathbf{\Lambda}\| > \gamma. \end{cases}$$

So the dual problem is given by

$$\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} -\frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}\|_{d_{\min}}^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2, \quad \text{s.t.} \quad \|\mathbf{\Lambda}\| \le \gamma. \tag{14}$$

Problem (14) can be solved efficiently due to their convexity. In particular, Grussler et al. [1] provided a computationally efficient algorithm to compute the proximal operators of functions $\frac{1}{2}\|\cdot\|_r^2$. Hence, the Douglas-Rachford algorithm can find the global minimum up to an ϵ error in function value in time $\mathsf{poly}(1/\epsilon)$ [2].

References

[1] C. Grussler, A. Rantzer, and P. Giselsson. Low-rank optimization with convex constraints. *arXiv:1606.01793*, 2016.

[2] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[3] T. L. Magnanti, J. F. Shapiro, and M. H. Wagner. Generalized linear programming solves the dual. *Management Science*, 22(11):1195–1203, 1976.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] R. M. Starr. Quasi-equilibria in markets with non-convex preferences. *Econometrica: Journal of the Econometric Society*, pages 25–38, 1969.

[6] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.