Appendix

Table of Contents

A. Appendix

Here, we present additional materials and proofs of the main results that are not included in the main paper due to the page limit. We also restate each result before the corresponding proof for the convenience of the reader. We also provide a table for notations below.

Table 2: Notations and variables in this paper.

Notation	Description
$\mathbf{X} \in \mathbb{R}^{n \times d}$	Data matrix
$\mathbf{y} \in \mathbb{R}^n$,$\mathbf{Y} \in \mathbb{R}^{n \times K}$	Label vector and matrix
$\mathbf{W}_{l,j} \in \mathbb{R}^{m_{l-1} \times m_l}$	l^{th} layer weight matrix
$\mathbf{A}_l \in \mathbb{R}^{n \times m_l}$	l^{th} layer activation matrix
$\boldsymbol{\lambda} \in \mathbb{R}^n$,$\boldsymbol{\Lambda} \in \mathbb{R}^{n \times K}$	Dual vector and matrix
$\mathbf{w}^* \in \mathbb{R}^d$,$\mathbf{W}^* \in \mathbb{R}^{d \times K}$	Optimal weight vector and matrix
r	Rank of $\mathbf{X}$
$\mathbf{U}_x \boldsymbol{\Sigma}_x \mathbf{V}_x^T$	Full SVD of $\mathbf{X}$
$\mathbf{e}_j$	j^{th} ordinary basis vector
$\mathcal{L}(\cdot,\cdot)$	Arbitrary convex loss function
$f_{\theta,L}(\mathbf{X})$	Output of an L-layer network

A.1. General loss functions

In this section, we show that our extreme point characterization holds for arbitrary convex loss functions including cross entropy and hinge loss. We first restate the primal training problem after applying the rescaling in Lemma 1.1 as follows

$$\min_{\{\theta_l\}_{l=1}^L, t_j} \mathcal{L}(f_{\theta,L}(\mathbf{X}), \mathbf{y}) + \beta \|\mathbf{w}_L\|_1 + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \text{ s.t. } \mathbf{w}_{L-1,j} \in \mathcal{B}_2, \|\mathbf{W}_{l,j}\|_F \leq t_j,\ \forall l \in [L-2], \forall j \in [m], \quad (21)$$

where $\mathcal{L}(\cdot, \mathbf{y})$ is a convex loss function.

Theorem A.1. *The dual of* (21) *is given by*

$$\min_{t_j} \max_{\boldsymbol{\lambda}} -\mathcal{L}^*(\boldsymbol{\lambda}) + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \textit{ s.t. } \max_{\substack{\mathbf{w}_{L-1,j} \in \mathcal{B}_2 \\ \|\mathbf{W}_{l,j}\|_F \leq t_j}} \|\mathbf{A}_{L-1}^T \boldsymbol{\lambda}\|_\infty \leq \beta\,,$$

where $\mathcal{L}^*$ is the Fenchel conjugate function defined as

$$\mathcal{L}^*(\boldsymbol{\lambda}) = \max_{\mathbf{z}} \mathbf{z}^T\boldsymbol{\lambda} - \mathcal{L}(\mathbf{z}, \mathbf{y})\,.$$

Proof of Theorem A.1. The proof directly follows from the dual derivation in Appendix A.2. □

Theorem A.1 proves that our extreme point characterization applies to arbitrary loss function. Therefore, optimal parameters for (21) are a subset of the same extreme point set, i.e., determined by the input data matrix $\mathbf{X}$, independent of loss function.

Remark A.1. *Since our characterization is generic in the sense that it holds for vector output, deep linear and deep ReLU networks (see the main paper for details), Theorem A.1 is also valid for all of these cases.*

A.2. Derivations for the dual problem in (3)

We first restate the scaled primal problem in Lemma 1.1

$$P^* = \min_{\{\theta_l\}_{l=1}^L, t_j, \hat{\mathbf{y}}} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) + \beta\|\mathbf{w}_L\|_1 + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \;\text{ s.t. }\; \begin{array}{l} \mathbf{w}_{L-1,j} \in \mathcal{B}_2, \|\mathbf{W}_{l,j}\|_F \le t_j,\ \forall l \in [L-2], \forall j \in [m] \\ \hat{\mathbf{y}} = f_{\theta,L}(\mathbf{X}). \end{array} \tag{22}$$

Then, the corresponding Lagrangian is

$$L(\boldsymbol{\lambda}, \hat{\mathbf{y}}, \mathbf{w}_L) = \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) - \boldsymbol{\lambda}^T\hat{\mathbf{y}} + \boldsymbol{\lambda}^T f_{\theta,L}(\mathbf{X}) + \beta\|\mathbf{w}_L\|_1 + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2.$$

Based on the Lagrangian above, we now obtain the dual function as follows

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \min_{\hat{\mathbf{y}}, \mathbf{w}_L} L(\boldsymbol{\lambda}, \hat{\mathbf{y}}, \mathbf{w}_L) \\ &= \min_{\hat{\mathbf{y}}, \mathbf{w}_L} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) - \boldsymbol{\lambda}^T\hat{\mathbf{y}} + \boldsymbol{\lambda}^T f_{\theta,L}(\mathbf{X}) + \beta\|\mathbf{w}_L\|_1 + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \\ &= \min_{\hat{\mathbf{y}}, \mathbf{w}_L} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) - \boldsymbol{\lambda}^T\hat{\mathbf{y}} + \boldsymbol{\lambda}^T \mathbf{A}_{L-1}\mathbf{w}_L + \beta\|\mathbf{w}_L\|_1 + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \\ &= -\mathcal{L}^*(\boldsymbol{\lambda}) + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \text{ s.t. } \|\mathbf{A}_{L-1}^T\boldsymbol{\lambda}\|_\infty \le \beta, \end{aligned}$$

where $\mathcal{L}^*$ is the Fenchel conjugate function defined as (Boyd & Vandenberghe, 2004)

$$\mathcal{L}^*(\boldsymbol{\lambda}) = \max_{\mathbf{z}} \mathbf{z}^T\boldsymbol{\lambda} - \mathcal{L}(\mathbf{z}, \mathbf{y})\,.$$

Thus, taking the dual of (22) in terms of $\mathbf{w}_L$ and $\hat{\mathbf{y}}$ yield

$$\begin{array}{l} P^* = \displaystyle\min_{\{\theta_l\}_{l=1}^{L-1}, t_j} \max_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) \\ \quad \text{s.t. } \mathbf{w}_{L-1,j} \in \mathcal{B}_2, \|\mathbf{W}_{l,j}\|_F \le t_j,\ \forall l, j \end{array} = \begin{array}{l} \displaystyle\min_{\{\theta_l\}_{l=1}^{L-1}, t_j} \max_{\boldsymbol{\lambda}} -\mathcal{L}^*(\boldsymbol{\lambda}) + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \\ \text{s.t. } \mathbf{w}_{L-1,j} \in \mathcal{B}_2, \|\mathbf{W}_{l,j}\|_F \le t_j,\ \forall l, j,\ \|\mathbf{A}_{L-1}^T\boldsymbol{\lambda}\|_\infty \le \beta. \end{array}$$

To achieve the lower bound in the main paper, we now change the order of min (for the layer weights)-max as follows

$$\begin{aligned} P^* \ge D^* &= \min_{t_j} \max_{\boldsymbol{\lambda}} \min_{\substack{\mathbf{w}_{L-1,j} \in \mathcal{B}_2 \\ \|\mathbf{W}_{l,j}\|_F \le t_j}} -\mathcal{L}^*(\boldsymbol{\lambda}) + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \;\text{ s.t. } \|\mathbf{A}_{L-1}^T\boldsymbol{\lambda}\|_\infty \le \beta \\ &= \min_{t_j} \max_{\boldsymbol{\lambda}} -\mathcal{L}^*(\boldsymbol{\lambda}) + \frac{\beta}{2}(L-2)\sum_{j=1}^m t_j^2 \;\text{ s.t. } \max_{\substack{\mathbf{w}_{L-1,j} \in \mathcal{B}_2 \\ \|\mathbf{W}_{l,j}\|_F \le t_j}} \|\mathbf{A}_{L-1}^T\boldsymbol{\lambda}\|_\infty \le \beta, \end{aligned}$$

which completes the derivation.

A.3. Equivalence (Rescaling) lemmas for the non-convex objectives

In this section, we present all the equivalence (scaling transformation) lemmas we used in the main paper and the the proofs are presented in Appendix A.6, A.7, and A.8, two-layer, deep linear, and deep ReLU networks, respectively. We also note that similar scaling techniques were also utilized in (Neyshabur et al., 2014; Savarese et al., 2019; Ergen & Pilanci, 2019; 2020a;b;c).

Lemma 1.1. *The following problems are equivalent :*

$$\begin{aligned} &\min_{\{\theta_l\}_{l=1}^L} \mathcal{L}(f_{\theta,L}(\mathbf{X}),\mathbf{y}) + \frac{\beta}{2}\sum_{j=1}^m\sum_{l=1}^L \|\mathbf{W}_{l,j}\|_F^2 \\ = &\min_{\{\theta_l\}_{l=1}^L,\{t_j\}_{j=1}^m} \mathcal{L}(f_{\theta,L}(\mathbf{X}),\mathbf{y}) + \beta\|\mathbf{w}_L\|_1 \\ &\qquad\qquad + \frac{\beta}{2}(L-2)\sum_{j-1}^m t_j^2. \\ &\text{s.t. } \mathbf{w}_{L-1,j}\in\mathcal{B}_2, \|\mathbf{W}_{l,j}\|_F \le t_j,\ \forall l\in[L-2] \end{aligned}$$

Proof of Lemma 1.1. For any $\theta\in\Theta$, we can rescale the parameters as $\bar{\mathbf{w}}_{L-1,j} = \alpha_j \mathbf{w}_{L-1,j}$ and $\bar{w}_{L,j} = w_{L,j}/\alpha_j$, for any $\alpha_j > 0$. Then, the network output becomes

$$f_{\bar{\theta},L}(\mathbf{X}) = \sum_{j=1}^m \Big((\mathbf{X}\mathbf{W}_{1,j})_+ \dots \bar{\mathbf{w}}_{L-1,j}\Big)_+ \bar{w}_{L,j} = \sum_{j=1}^m \Big((\mathbf{X}\mathbf{W}_{1,j})_+ \dots \mathbf{w}_{L-1,j}\Big)_+ w_{L,j} = f_{\theta,L}(\mathbf{X}),$$

which proves that this scaling does not change the output of the network. In addition to this, we have the following basic inequality

$$\sum_{j=1}^m\sum_{l=1}^L \|\mathbf{W}_{l,j}\|_F^2 \ge \sum_{j=1}^m\sum_{l=1}^{L-2} \|\mathbf{W}_l\|_F^2 + 2\sum_{j=1}^m |w_{L,j}|\,\|\mathbf{w}_{L-1,j}\|_2,$$

where the equality is achieved with the scaling choice $\alpha_j = \big(\frac{|w_{L,j}|}{\|\mathbf{w}_{L-1,j}\|_2}\big)^{\frac{1}{2}}$ is used. Since the scaling operation does not change the right-hand side of the inequality, we can set $\|\mathbf{w}_{L-1,j}\|_2 = 1, \forall j$. Therefore, the right-hand side becomes $\|\mathbf{w}_L\|_1$.

Now, let us consider a modified version of the problem, where the unit norm equality constraint is relaxed as $\|\mathbf{w}_{L-1,j}\|_2 \le 1$. Let us also assume that for a certain index j, we obtain $\|\mathbf{w}_{L-1,j}\|_2 < 1$ with $w_{L,j} \neq 0$ as an optimal solution. This shows that the unit norm inequality constraint is not active for $\mathbf{w}_{L-1,j}$, and hence removing the constraint for $\mathbf{w}_{L-1,j}$ will not change the optimal solution. However, when we remove the constraint, $\|\mathbf{w}_{L-1,j}\|_2 \to \infty$ reduces the objective value since it yields $w_{L,j} = 0$. Therefore, we have a contradiction, which proves that all the constraints that correspond to a nonzero $w_{L,j}$ must be active for an optimal solution. This also shows that replacing $\|\mathbf{w}_{L-1,j}\|_2 = 1$ with $\|\mathbf{w}_{L-1,j}\|_2 \le 1$ does not change the solution to the problem.

Then, we use the epigraph form for the sum of the norm of the first $L-2$ layers to achieve the equivalence, i.e., we introduce $\|\mathbf{W}_{l,j}\|_F \le t_j$ constraint and replace $\sum_{l=1}^{L-2}\|\mathbf{W}_{l,j}\|_F^2$ with $(L-2)t_j^2$ in the objective. We also note that since the optimal layer weights have the same Frobenius norm as proven in Proposition 3.1, we can replace the Frobenius norm of each layer weight matrix with the same variable t without loss of generality. □

Lemma A.1. *The following two problems are equivalent:*

$$\begin{aligned} &\min_{\theta\in\Theta} \|\mathbf{W}_1\|_F^2 + \|\mathbf{w}_2\|_2^2 \\ &\text{s.t. } f_{\theta,2}(\mathbf{X}) = \mathbf{y} \end{aligned} \quad = \quad \begin{aligned} &\min_{\theta\in\Theta} \|\mathbf{w}_2\|_1 \\ &\text{s.t. } f_{\theta,2}(\mathbf{X}) = \mathbf{y}, \mathbf{w}_{1,j}\in\mathcal{B}_2 \end{aligned}.$$

Proof of Lemma A.1. For any $\theta \in \Theta$, we can rescale the parameters as $\bar{\mathbf{w}}_{1,j} = \alpha_j \mathbf{w}_{1,j}$ and $\bar{w}_{2,j} = w_{2,j}/\alpha_j$, for any $\alpha_j > 0$. Then, the network output becomes

$$f_{\bar{\theta},2}(\mathbf{X}) = \sum_{j=1}^{m} \bar{w}_{2,j} \mathbf{X} \bar{\mathbf{w}}_{1,j} = \sum_{j=1}^{m} \frac{w_{2,j}}{\alpha_j} \alpha_j \mathbf{X} \mathbf{w}_{1,j} = \sum_{j=1}^{m} w_{2,j} \mathbf{X} \mathbf{w}_{1,j},$$

which proves $f_{\theta,2}(\mathbf{X}) = f_{\bar{\theta},2}(\mathbf{X})$. In addition to this, we have the following basic inequality

$$\frac{1}{2} \sum_{j=1}^{m} (w_{2,j}^2 + \|\mathbf{w}_{1,j}\|_2^2) \geq \sum_{j=1}^{m} (|w_{2,j}| \, \|\mathbf{w}_{1,j}\|_2),$$

where the equality is achieved with the scaling choice $\alpha_j = \big(\frac{|w_{2,j}|}{\|\mathbf{w}_{1,j}\|_2}\big)^{\frac{1}{2}}$ is used. Since the scaling operation does not change the right-hand side of the inequality, we can set $\|\mathbf{w}_{1,j}\|_2 = 1, \forall j$. Therefore, the right-hand side becomes $\|\mathbf{w}_2\|_1$.

Now, let us consider a modified version of the problem, where the unit norm equality constraint is relaxed as $\|\mathbf{w}_{1,j}\|_2 \leq 1$. Let us also assume that for a certain index j, we obtain $\|\mathbf{w}_{1,j}\|_2 < 1$ with $w_{2,j} \neq 0$ as an optimal solution. This shows that the unit norm inequality constraint is not active for $\mathbf{w}_{1,j}$, and hence removing the constraint for $\mathbf{w}_{1,j}$ will not change the optimal solution. However, when we remove the constraint, $\|\mathbf{w}_{1,j}\|_2 \to \infty$ reduces the objective value since it yields $w_{2,j} = 0$. Therefore, we have a contradiction, which proves that all the constraints that correspond to a nonzero $w_{2,j}$ must be active for an optimal solution. This also shows that replacing $\|\mathbf{w}_{1,j}\|_2 = 1$ with $\|\mathbf{w}_{1,j}\|_2 \leq 1$ does not change the solution to the problem. □

Lemma A.2. *The following problems are equivalent:*

$$\begin{array}{lcl} \min\limits_{\theta \in \Theta} \|\mathbf{W}_1\|_F^2 + \|\mathbf{W}_2\|_F^2 & & \min\limits_{\theta \in \Theta} \sum\limits_{j=1}^{m} \|\mathbf{w}_{2,j}\|_2 \\ & = & \\ \textit{s.t. } f_{\theta,2}(\mathbf{X}) = \mathbf{Y} & & \textit{s.t. } f_{\theta,2}(\mathbf{X}) = \mathbf{Y}, \mathbf{w}_{1,j} \in \mathcal{B}_2, \forall j \end{array}.$$

Proof of Lemma A.2. The proof directly follows from Proof of Lemma A.1 using the following inequality

$$\frac{1}{2} \sum_{j=1}^{m} (\|\mathbf{w}_{2,j}\|_2^2 + \|\mathbf{w}_{1,j}\|_2^2) \geq \sum_{j=1}^{m} (\|\mathbf{w}_{2,j}\|_2 \, \|\mathbf{w}_{1,j}\|_2).$$

Then, if we set $\|\mathbf{w}_{1,j}\|_2 = 1, \forall j$, the right-hand side becomes $\sum_{j=1}^{m} \|\mathbf{w}_{2,j}\|_2$. □

Lemma A.3. *The following problems are equivalent:*

$$\begin{array}{lcl} \min\limits_{\{\theta_l\}_{l=1}^{L}} \frac{1}{2} \sum\limits_{j=1}^{m} \sum\limits_{l=1}^{L} \|\mathbf{W}_{l,j}\|_F^2 & & \min\limits_{\{\theta_l\}_{l=1}^{L}, \{t_j\}_{j=1}^{m}} \|\mathbf{w}_L\|_1 + \frac{1}{2}(L-2) \sum\limits_{j=1}^{m} t_j^2 \\ & = & \\ \textit{s.t. } f_{\theta,L}(\mathbf{X}) = \mathbf{y} & & \textit{s.t. } f_{\theta,L}(\mathbf{X}) = \mathbf{y}, \; \mathbf{w}_{L-1,j} \in \mathcal{B}_2, \|\mathbf{W}_{l,j}\|_F \leq t_j, \; \forall l \in [L-2], \forall j \in [m] \end{array}.$$

Proof of Lemma A.3. Applying the rescaling in Lemma A.1 to the last two layers of the L-layer network in (15) gives

$$\begin{array}{l} \min\limits_{\{\theta_l\}_{l=1}^{L}} \|\mathbf{w}_L\|_1 + \frac{1}{2} \sum\limits_{j=1}^{m} \sum\limits_{l=1}^{L-2} \|\mathbf{W}_{l,j}\|_F^2 \\ \text{s.t. } \|\mathbf{w}_{L-1,j}\|_2 \leq 1, \forall j \in [m], \; \sum\limits_{j=1}^{m} \mathbf{X}\mathbf{W}_{1,j} \ldots \mathbf{w}_{L-1,j} w_{L,j} = \mathbf{y} \end{array}.$$

Then, we use the epigraph form for the norm of the first $L-2$ to achieve the equivalence. □

Lemma A.4. *The following problems are equivalent:*

$$\begin{array}{lcl} \displaystyle\min_{\{\theta_l\}_{l=1}^L} \frac{1}{2}\sum_{j=1}^m \sum_{l=1}^L \|\mathbf{W}_{l,j}\|_F^2 & = & \displaystyle\min_{\{\theta_l\}_{l=1}^L, \{t_l\}_{l=1}^{L-2}} \sum_{j=1}^m \|\mathbf{w}_{L,j}\|_2 + \frac{1}{2}(L-2)\sum_{j=1}^m t_j^2 \\ \textit{s.t. } f_{\theta,L}(\mathbf{X}) = \mathbf{Y} & & \textit{s.t. } f_{\theta,L}(\mathbf{X}) = \mathbf{Y},\ \mathbf{w}_{L-1,j} \in \mathcal{B}_2, \|\mathbf{W}_{l,j}\|_F \leq t_j,\ \forall l \in [L-2],\ \forall j \in [m] \end{array}.$$

Proof of Lemma A.4. Applying the rescaling in Lemma A.1 to the last two layer of the L-layer network in (17) gives

$$\begin{aligned} &\min_{\{\theta_l\}_{l=1}^L} \sum_{j=1}^m \|\mathbf{w}_{L,j}\|_2 + \frac{1}{2}\sum_{j=1}^m \sum_{l=1}^{L-2} \|\mathbf{W}_l\|_F^2 \\ &\text{s.t. } \|\mathbf{w}_{L-1,j}\|_2 \leq 1, \forall j \in [m],\ \sum_{j=1}^m \mathbf{X}\mathbf{W}_{1,j}\ldots\mathbf{w}_{L-1,j}\mathbf{w}_{L,j}^T = \mathbf{Y} \end{aligned}.$$

Then, we use the epigraph form for the norm of the first $L-2$ to achieve the equivalence. □

A.4. Regularization in Theorem 4.4

In this section, we prove that regularizing the all the parameters do not alter the claims in Theorem 4.4. We first state the primal problem, where all the parameters are regularized, as follows

$$P_r^* = \min_{\theta \in \Theta} \frac{1}{2}\left\|\sum_{j=1}^m \left(\text{BN}_{\gamma,\alpha}\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)\right)_+ \mathbf{w}_{L,j}^T - \mathbf{Y}\right\|_F^2 + \frac{\beta}{2}\sum_{j=1}^m\sum_{l=1}^L \left(\left\|\boldsymbol{\gamma}_j^{(l)}\right\|_2^2 + \left\|\boldsymbol{\alpha}_j^{(l)}\right\|_2^2 + \|\mathbf{W}_{l,j}\|_F^2\right), \tag{23}$$

where we use $\boldsymbol{\gamma}^{(L)} = \boldsymbol{\alpha}^{(L)} = \mathbf{0}$ as dummy variables for notational simplicity. Now, we rewrite (23) as

$$\begin{aligned} P_r^* = \min_{t \geq 0}\min_{\theta \in \Theta} &\frac{1}{2}\left\|\sum_{j=1}^m \left(\text{BN}_{\gamma,\alpha}\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)\right)_+ \mathbf{w}_{L,j}^T - \mathbf{Y}\right\|_F^2 + \frac{\beta}{2}\sum_{j=1}^m \left(\gamma_j^{(L-1)^2} + \alpha_j^{(L-1)^2} + \|\mathbf{w}_{L,j}\|_2^2\right) + \frac{\beta}{2}t \\ \text{s.t. } &\sum_{j=1}^m \sum_{l=1}^{L-2} \left(\left\|\boldsymbol{\gamma}_j^{(l)}\right\|_2^2 + \left\|\boldsymbol{\alpha}_j^{(l)}\right\|_2^2 + \|\mathbf{W}_{l,j}\|_F^2\right) + \|\mathbf{W}_{L-1}\|_F^2 \leq t. \end{aligned}$$

After applying the scaling between $\mathbf{W}_L$ and $(\boldsymbol{\gamma}^{(L-1)}, \boldsymbol{\alpha}^{(L-1)})$ as in Lemma A.4, we take the dual with respect to $\mathbf{W}_L$ to obtain the following problem

$$\begin{aligned} P_r^* \geq D_r^* = &\max_{t \geq 0}\max_{\boldsymbol{\Lambda}} -\frac{1}{2}\|\boldsymbol{\Lambda} - \mathbf{Y}\|_F^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2 + \frac{\beta}{2}t \\ &\text{s.t. } \max_{\theta \in \Theta_r}\left\|\boldsymbol{\Lambda}^T\left(\text{BN}_{\gamma,\alpha}\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)\right)_+\right\|_2 \leq \beta, \end{aligned} \tag{24}$$

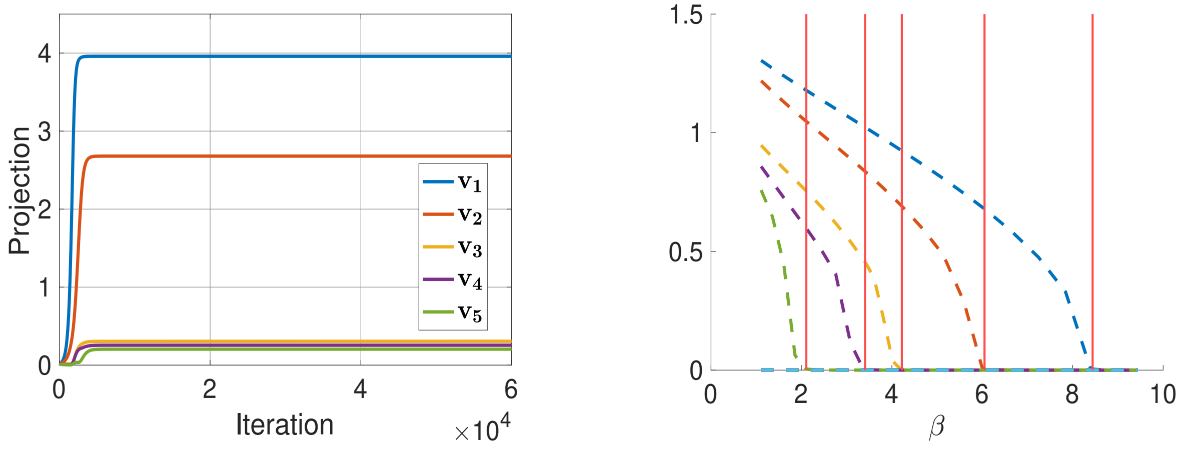
where $\Theta_r = \{\theta \in \Theta : \gamma_j^{(L-1)^2} + \alpha_j^{(L-1)^2} = 1,\ \forall j \in [m],\ \sum_{l=1}^{L-2}\left(\left\|\boldsymbol{\gamma}^{(l)}\right\|_2^2 + \left\|\boldsymbol{\alpha}^{(l)}\right\|_2^2 + \|\mathbf{W}_{l,j}\|_F^2\right) + \|\mathbf{W}_{L-1}\|_F^2 \leq t\}$. Since

$$\text{BN}_{\gamma,\alpha}\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right) = \underbrace{\frac{(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n})\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}}{\|(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n})\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\|_2}}_{\mathbf{h}(\theta')}\gamma_j^{(L-1)} + \frac{\mathbf{1}_n}{\sqrt{n}}\alpha_j^{(L-1)},$$

where θ' denotes all the parameters except $\boldsymbol{\gamma}^{(L-1)}, \boldsymbol{\alpha}^{(L-1)}, \mathbf{W}_L$. Then, independent of the value t, $\mathbf{h}(\theta')$ is always a unit norm vector. Therefore, the maximization constraint in (24) is independent of the norms of the parameters in θ', which also proves that regularizing the weights in θ' does not affect the dual characterization in (24).

A.5. Additional numerical results

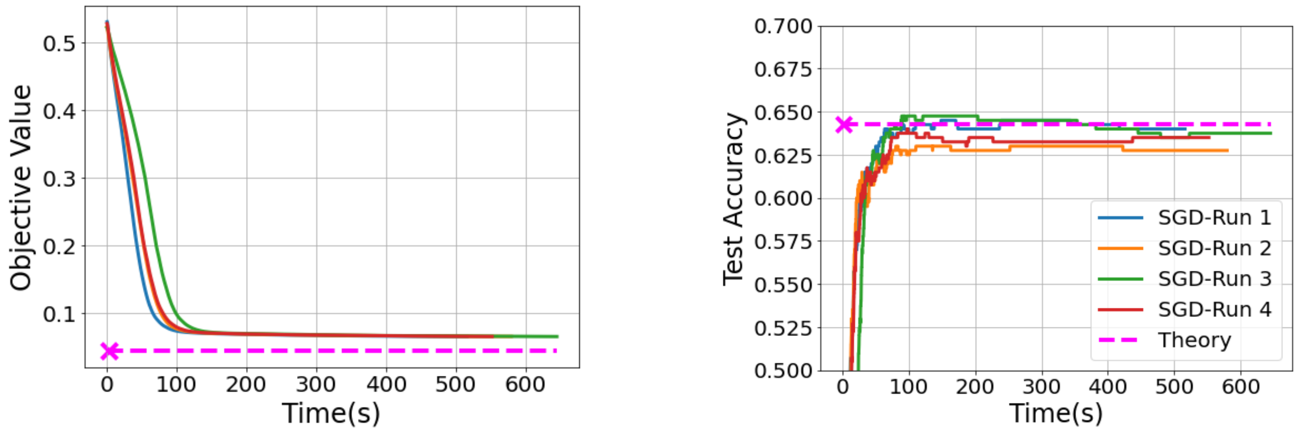
Here, we present additional numerical results that are not included in the main paper. In Figure 5a, we perform an experiment to check whether the hidden neurons of a two-layer linear network align with the proposed right singular vectors. For this experiment, we select a certain β such that $\mathbf{W}_1$ becomes rank-two. After training, we first normalize each neuron to have unit norm, i.e., $\|\mathbf{w}_{1,j}\|_2 = 1, \forall j$, and then compute the sum of the projections of each neuron onto each right singular vector, i.e., denoted as $\mathbf{v}_i$. Since we choose β such that $\mathbf{W}_1$ is a rank-two matrix, most of the neurons align with the first two right singular vectors as expected. Therefore, this experiment verifies our analysis and claims in Remark 2.1. Furthermore, as an alternative to Figure 2a, we plot the singular values of $\mathbf{W}_1$ with respect to the regularization parameter β in Figure 5b.



(a)

(b)

Figure 5: (a) Projection of the hidden neurons to the right singular vectors claimed in Remark 2.1 and (b) singular values of $\mathbf{W}_1$ with respect to β.



(a) CIFAR-100-Training objective

(b) CIFAR-100-Test accuracy

Figure 6: Training and test performance of full batch SGD (4 initializations) on the CIFAR-100 datasets for a four class classification tasks, where $(n, d) = (2000, 3072)$, $K = 4$, $L = 2$ with 100 neurons and we use squared loss with one hot encoding. For Theory, we use the layer weights in Theorem 4.4, which achieves the optimal performance as guaranteed by Theorem 4.4. We also use a marker to denote the time required to compute the closed-form solution.

We also conduct an experiment on CIFAR-100 (Krizhevsky et al., 2014) datasets, for which we consider a four class classification task. In order to verify our results in Theorem 4.4, we train a two-layer regularized ReLU networks with batch normalization using four different initializations and then plot the results with respect to wall-clock time. As demonstrated in Figure 6, our closed form solution, i.e., denoted as Theory, achieves lower objective value as proven in Theorem 4.4 and higher test accuracy.

A.6. Proofs for two-layer networks

Theorem 2.1. *The dual of the problem in* (5) *is given by*

$$P^* \geq D^* = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \boldsymbol{\lambda}^T \mathbf{y} \;\; \textit{s.t.} \;\; \max_{\mathbf{w}_1 \in \mathcal{B}_2} \left| \boldsymbol{\lambda}^T \mathbf{X} \mathbf{w}_1 \right| \leq 1 \,. \tag{6}$$

For (5)*,* $\exists m^* \leq n+1$ *such that strong duality holds, i.e.,* $P^* = D^*$*,* $\forall m \geq m^*$ *and* $\mathbf{W}_1^*$ *satisfies* $\|(\mathbf{X}\mathbf{W}_1^*)^T \boldsymbol{\lambda}^*\|_\infty = 1$ *,* *where* $\boldsymbol{\lambda}^*$ *is the dual optimal parameter.*

Corollary 2.1. *By Theorem 2.1, the optimal neurons are extreme points which solve* $\operatorname{argmax}_{\mathbf{w}_1 \in \mathcal{B}_2} |{\boldsymbol{\lambda}^*}^T \mathbf{X} \mathbf{w}_1 |$.

Proof of Theorem 2.1 and Corollary 2.1. We first note that the dual of (5) with respect to $\mathbf{w}_2$ is

$$\min_{\theta \in \Theta \backslash \{\mathbf{w}_2\}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \text{ s.t. } \|(\mathbf{X}\mathbf{W}_1)_+^T \boldsymbol{\lambda}\|_\infty \leq 1, \; \|\mathbf{w}_{1,j}\|_2 \leq 1, \forall j.$$

Then, we can reformulate the problem as follows

$$P^* = \min_{\theta \in \Theta \backslash \{\mathbf{w}_2\}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} + \mathcal{I}(\|(\mathbf{X}\mathbf{W}_1)_+^T \boldsymbol{\lambda}\|_\infty \leq 1), \text{ s.t. } \|\mathbf{w}_{1,j}\|_2 \leq 1, \forall j.$$

where $\mathcal{I}(\|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1)$ is the characteristic function of the set $\|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1$, which is defined as

$$\mathcal{I}(\|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1) = \begin{cases} 0 & \text{if } \|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1 \\ -\infty & \text{otherwise} \end{cases} .$$

Since the set $\|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1$ is closed, the function $\Phi(\boldsymbol{\lambda}, \mathbf{W}_1) = \boldsymbol{\lambda}^T \mathbf{y} + \mathcal{I}(\|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1)$ is the sum of a linear function and an upper-semicontinuous indicator function and therefore upper-semicontinuous. The constraint on $\mathbf{W}_1$ is convex and compact. We use P^* to denote the value of the above min-max program. Exchanging the order of min-max we obtain the dual problem given in (6), which establishes a lower bound D^* for the above problem:

$$\begin{aligned} P^* \geq D^* &= \max_{\boldsymbol{\lambda}} \min_{\theta \in \Theta \backslash \{\mathbf{w}_2\}} \boldsymbol{\lambda}^T \mathbf{y} + \mathcal{I}(\|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1), \text{ s.t. } \|\mathbf{w}_{1,j}\|_2 \leq 1, \forall j, \\ &= \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y}, \text{ s.t. } \|(\mathbf{X}\mathbf{W}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1 \, \forall \mathbf{w}_{1,j} : \|\mathbf{w}_{1,j}\|_2 \leq 1, \forall j, \\ &= \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y}, \text{ s.t. } \|(\mathbf{X}\mathbf{w}_1)^T \boldsymbol{\lambda}\|_\infty \leq 1 \, \forall \mathbf{w}_1 : \|\mathbf{w}_1\|_2 \leq 1, \end{aligned}$$

We now show that strong duality holds for infinite size NNs. The dual of the semi-infinite program in (6) is given by (see Section 2.2 of (Goberna & López-Cerdá, 1998) and also (Bach, 2017))

$$\begin{aligned} &\min \|\boldsymbol{\mu}\|_{TV} \\ &\text{s.t.} \int_{\mathbf{w}_1 \in \mathcal{B}_2} \mathbf{X}\mathbf{w}_1 d\boldsymbol{\mu}(\mathbf{w}_1) = \mathbf{y} \,, \end{aligned}$$

where TV is the total variation norm of the Radon measure $\boldsymbol{\mu}$. This expression coincides with the infinite-size NN as given in (Bach, 2017), and therefore strong duality holds. We also note that although the above formulation involves an infinite dimensional integral form, by Caratheodory's theorem, the integral can be represented as a finite summation of at most $n+1$ Dirac delta functions (Rosset et al., 2007). Next we invoke the semi-infinite optimality conditions for the dual problem in (6), in particular we apply Theorem 7.2 of (Goberna & López-Cerdá, 1998). We first define the set

$$\mathbf{K} = \mathbf{cone} \left\{ \begin{pmatrix} s\mathbf{X}\mathbf{w}_1 \\ 1 \end{pmatrix}, \mathbf{w}_1 \in \mathcal{B}_2, s \in \{-1, +1\}; \begin{pmatrix} \mathbf{0}_n \\ -1 \end{pmatrix} \right\} .$$

Note that $\mathbf{K}$ is the union of finitely many convex closed sets, since the function $\mathbf{X}\mathbf{w}_1$ can be expressed as the union of finitely many convex closed sets. Therefore the set $\mathbf{K}$ is closed. By Theorem 5.3 (Goberna & López-Cerdá, 1998), this implies that the set of constraints in (6) forms a Farkas-Minkowski system. By Theorem 8.4 of (Goberna & López-Cerdá, 1998), primal and dual values are equal, given that the system is consistent. Moreover, the system is discretizable, i.e., there exists a sequence of problems with finitely many constraints whose optimal values approach to the optimal value of (6). The optimality conditions in Theorem 7.2 (Goberna & López-Cerdá, 1998) implies that $\mathbf{y} = \mathbf{X}\mathbf{W}_1^*\mathbf{w}_2^*$ for some vector $\mathbf{w}_2^*$. Since the primal and dual values are equal, we have ${\boldsymbol{\lambda}^*}^T \mathbf{y} = {\boldsymbol{\lambda}^*}^T \mathbf{X}\mathbf{W}_1^*\mathbf{w}_2^* = \|\mathbf{w}_2^*\|_1$, which shows that the primal-dual pair $(\{\mathbf{w}_2^*, \mathbf{W}_1^*\}, \boldsymbol{\lambda}^*)$ is optimal. Thus, the optimal neuron weights $\mathbf{W}_1^*$ satisfy $\|(\mathbf{X}\mathbf{W}_1^*)^T \boldsymbol{\lambda}^*\|_\infty = 1$. □

Proposition 2.1. *[(Du & Hu, 2019)] Given* $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2$, *we have*

$$\operatorname*{argmin}_{\mathbf{W}_1,\mathbf{w}_2} \|\mathbf{X}\mathbf{W}_1\mathbf{w}_2 - \mathbf{X}\mathbf{w}^*\|_2^2 = \operatorname*{argmin}_{\mathbf{W}_1,\mathbf{w}_2} \|\mathbf{X}\mathbf{W}_1\mathbf{w}_2 - \mathbf{y}\|_2^2.$$

Proof of Proposition 2.1. Let us first define a variable $\mathbf{w}^*$ that minimizes the following problem

$$\mathbf{w}^* = \min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2.$$

Thus, the following relation holds

$$\mathbf{X}^T(\mathbf{X}\mathbf{w}^* - \mathbf{y}) = \mathbf{0}_d.$$

Then, for any $\mathbf{w} \in \mathbb{R}^d$, we have

$$\begin{aligned} f(\mathbf{w}) &= \|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}^* + \mathbf{X}\mathbf{w}^* - \mathbf{y}\|_2^2 \\ &= \|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}^*\|_2^2 + 2(\mathbf{w} - \mathbf{w}^*)^T \underbrace{\mathbf{X}^T(\mathbf{X}\mathbf{w}^* - \mathbf{y})}_{=\mathbf{0}_d} + \|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|_2^2 \\ &= \|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}^*\|_2^2 + \|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|_2^2. \end{aligned}$$

Notice that $\|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|_2^2$ does not depend on $\mathbf{w}$, thus, the relation above proves that minimizing $f(\mathbf{w})$ is equivalent to minimizing $\|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}^*\|_2^2$, where $\mathbf{w}^*$ is the planted model parameter. Therefore, the planted model assumption does not change solution to the linear network training problem in (5). □

Theorem 2.2. *Let* $\{\mathbf{X}, \mathbf{y}\}$ *be feasible for* (5)*, then strong duality holds for finite width networks.*

Proof of Theorem 2.2. Since there exists a single extreme point, we can construct a weight vector $\mathbf{w}_e \in \mathbb{R}^d$ that is the extreme point. Then, the dual of (5) with $\mathbf{W}_1 = \mathbf{w}_e$ is

$$D_e^* = \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\mathbf{y} \text{ s.t. } \|(\mathbf{X}\mathbf{w}_e)^T\boldsymbol{\lambda}\|_\infty \leq 1. \tag{25}$$

Then, we have

$$\begin{aligned} P^* = \min_{\theta\in\Theta\backslash\{\mathbf{w}_2\}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\mathbf{y} \quad &\geq \quad \max_{\boldsymbol{\lambda}} \min_{\theta\in\Theta\backslash\{\mathbf{w}_2\}} \boldsymbol{\lambda}^T\mathbf{y} \\ \text{s.t } \|(\mathbf{X}\mathbf{W}_1)^T\boldsymbol{\lambda}\|_\infty \leq 1,\ \|\mathbf{w}_{1,j}\|_2 \leq 1, \forall j \quad & \quad\ \text{s.t. } \|(\mathbf{X}\mathbf{W}_1)^T\boldsymbol{\lambda}\|_\infty \leq 1,\ \|\mathbf{w}_{1,j}\|_2 \leq 1, \forall j \\ &= \quad \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\mathbf{y} \\ & \quad\ \text{s.t. } \|(\mathbf{X}\mathbf{w}_e)^T\boldsymbol{\lambda}\|_\infty \leq 1 \\ &= \quad D_e^* = D^* \end{aligned} \tag{26}$$

where the first inequality follows from changing order of min-max to obtain a lower bound and the equality in the second line follows from Corollary 2.1.

From the fact that an infinite width NN can always find a solution with the objective value lower than or equal to the objective value of a finite width NN, we have

$$\begin{aligned} P_e^* = \min_{\theta\in\Theta\backslash\{\mathbf{W}_1, m\}} |w_2| \quad &\geq \quad P^* = \min_{\theta\in\Theta} \|\mathbf{w}_2\|_1 \\ \text{s.t. } \mathbf{X}\mathbf{w}_e w_2 = \mathbf{y} \quad & \quad\ \text{s.t. } \mathbf{X}\mathbf{W}_1\mathbf{w}_2 = \mathbf{y},\ \|\mathbf{w}_{1,j}\|_2 \leq 1, \forall j, \end{aligned} \tag{27}$$

where P^* is the optimal value of the original problem with infinitely many neurons. Now, notice that the optimization problem on the left hand side of (27) is convex since it is an ℓ_1-norm minimization problem with linear equality constraints. Therefore, strong duality holds for this problem, i.e., $P_e^* = D_e^*$. Using this result along with (26), we prove that strong duality holds for a finite width NN, i.e., $P_e^* = P^* = D^* = D_e^*$.

□

Theorem 2.3. *Strong duality holds for* (10) *with finite width.*

Proof of Theorem 2.3. Since there exists a single extreme point, we can construct a weight vector $\mathbf{w}_e \in \mathbb{R}^d$ that is the extreme point. Then, the dual of (10) with $\mathbf{W}_1 = \mathbf{w}_e$

$$D_e^* = \max_{\boldsymbol{\lambda}} -\frac{1}{2}\|\boldsymbol{\lambda} - \mathbf{y}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2^2 \text{ s.t. } |\boldsymbol{\lambda}^T \mathbf{X}\mathbf{w}_e| \leq \beta.$$

Then the rest of the proof directly follows Proof of Theorem 2.2. □

Theorem 2.4. *Let* $\{\mathbf{X}, \mathbf{Y}\}$ *be feasible for* (12)*, then strong duality holds for finite width networks.*

Proof of Theorem 2.4. Since there exist r_w possible extreme points, we can construct a weight matrix $\mathbf{W}_e \in \mathbb{R}^{d \times r_w}$ that consists of all the possible extreme points. Then, the dual of (12) with $\mathbf{W}_1 = \mathbf{W}_e$

$$D_e^* = \max_{\boldsymbol{\Lambda}} \mathbf{tr}(\boldsymbol{\Lambda}^T \mathbf{Y}) \text{ s.t. } \|\boldsymbol{\Lambda}^T \mathbf{X}\mathbf{w}_{e,j}\|_2 \leq 1, \forall j \in [r_w].$$

Then the rest of the proof directly follows Proof of Theorem 2.2. □

A.7. Proofs for deep linear networks

Proposition 3.1. *First* $L-2$ *layer weight matrices in* (15) *have the same operator and Frobenius norms, i.e.,* $t_j = \|\mathbf{W}_{l,j}\|_F = \|\mathbf{W}_{l,j}\|_2, \forall l \in [L-2]$, $\forall j \in [m]$.

Proof of Proposition 3.1. Let us first rescale the first $L-2$ layer weights as $\bar{\mathbf{W}}_{l,j} = \frac{t_{l,j}}{\|\mathbf{W}_{l,j}\|_F}\mathbf{W}_{l,j}$, where $t_{l,j} > 0$. Defining $t_j^{L-2} = \prod_{l=1}^{L-2} \|\mathbf{W}_{l,j}\|_F$, if $t_{l,j}$'s are chosen such that $\prod_{l=1}^{L-2} t_{l,j} = t_j^{L-2}$, then the rescaling does not alter the output of the network, i.e., $f_{\theta,L}(\mathbf{X}) = f_{\bar{\theta},L}(\mathbf{X})$. Therefore, we can optimize $\{t_{l,j}\}_{l=1}^{L-2}$ as follows

$$\min_{\{t_{l,j}\}_{l=1}^{L-2}} \frac{1}{2}\sum_{l=1}^{L-2} t_{l,j}^2 \text{ s.t. } \prod_{l=1}^{L-2} t_{l,j} = t_j^{L-2},$$

for each $j \in [m]$. Apparently, the optimal scaling parameters obey $t_{1,j} = t_{2,j} = \ldots = t_{L-2,j} = t_j$. We also note that the optimal layer weights satisfy $t_j = \|\mathbf{W}_{l,j}\|_2 = \|\mathbf{W}_{l,j}\|_F$, $\forall l \in [L-2]$, since the upper-bound is achieved when the matrices are rank-one (see (34)).

□

Theorem 3.1. *Optimal layer weights for* (15) *are*

$$\mathbf{W}_{l,j}^* = \begin{cases} t_j^* \frac{\mathbf{V}_x \tilde{\mathbf{w}}_r^*}{\|\tilde{\mathbf{w}}_r^*\|_2}\boldsymbol{\rho}_{1,j}^T & \text{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j}\boldsymbol{\rho}_{l,j}^T & \text{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \text{if } l = L-1 \end{cases},$$

where $\boldsymbol{\rho}_{l,j} \in \mathbb{R}^{m_l}$ *such that* $\|\boldsymbol{\rho}_{l,j}\|_2 = 1$, $\forall l \in [L-2]$, $\forall j \in [m]$ *and* $\tilde{\mathbf{w}}_r^*$ *is defined in* (9).

Proof of Theorem 3.1. Using Lemma A.3 and Proposition 3.1, we have the following dual problem for (15)

$$P^* = \min_{\{\theta_l\}_{l=1}^{L-1}, \{t_j\}_{j=1}^m} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} + \frac{1}{2}(L-2)\sum_{j=1}^m t_j^2 \text{ s.t. } \begin{array}{l} |(\mathbf{X}\mathbf{W}_{1,j}\ldots\mathbf{w}_{L-1,j})^T\boldsymbol{\lambda}| \leq 1, \ \mathbf{w}_{L-1,j} \in \mathcal{B}_2 \\ \|\mathbf{W}_{l,j}\|_F \leq t_j, \ \forall l \in [L-2], \ \forall j \in [m]. \end{array} \tag{28}$$

Now, let us assume that the optimal Frobenius norm for each layer l is t_j^* [9]. Then, if we define $\Theta_{L-1} = \{\theta_1, \ldots, \theta_{L-1} | \|\mathbf{w}_{L-1,j}\|_2 \leq 1, \ \|\mathbf{W}_{l,j}\|_F \leq t_j^*, \ \forall l \in [L-2], \forall j \in [m]\}$, (28) reduces to the following problem

$$P^* \geq D^* = \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \text{ s.t. } |(\mathbf{X}\mathbf{W}_{1,j}\ldots\mathbf{w}_{L-1,j})^T\boldsymbol{\lambda}| \leq 1, \ \forall \theta_l \in \Theta_{L-1}, \ \forall l, \tag{29}$$

[9] With this assumption, $(L-2)\sum_{j=1}^m t_j^2$ becomes constant so we ignore this term for the rest of our derivations.

where we change the order of min-max to obtain a lower bound for (28). The dual of the semi-infinite problem in (29) is given by

$$\min \|\boldsymbol{\mu}\|_{TV} \text{ s.t.} \int_{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}} \mathbf{X}\mathbf{W}_1 \dots \mathbf{w}_{L-1} d\boldsymbol{\mu}(\theta_1, \dots, \theta_{L-1}) = \mathbf{y}\,, \tag{30}$$

where $\boldsymbol{\mu}$ is a signed Radon measure and $\|\cdot\|_{TV}$ is the total variation norm. We emphasize that (30) has infinite width in each layer, however, an application of Caratheodory's theorem shows that the measure $\boldsymbol{\mu}$ in the integral can be represented by finitely many (at most $n+1$) Dirac delta functions (Rosset et al., 2007). Such selection of $\boldsymbol{\mu}$ yields the following problem

$$P_m^* = \min_{\{\theta_l\}_{l=1}^{L}} \|\mathbf{w}_L\|_1 \text{ s.t.} \sum_{j=1}^{m^*} \mathbf{X}\mathbf{W}_{1,j} \dots \mathbf{w}_{L-1,j} w_{L,j} = \mathbf{y},\ \theta_l \in \Theta_{L-1},\ \forall l \tag{31}$$

We first note that since the model in (31) has the same expressive power with the network in (15) as long as $m \geq m^*$, we have $P^* = P_m^*$. Since the dual of (15) and (31) are the same, we also have $D_m^* = D^*$, where D_m^* is the optimal dual value for (31).

We now apply the variable change in (8) to (29) as follows

$$\max_{\boldsymbol{\lambda}} \tilde{\boldsymbol{\lambda}}^T \boldsymbol{\Sigma}_x \tilde{\mathbf{w}}_r^* \text{ s.t. } \|\mathbf{W}_{L-2,j}^T \dots \mathbf{W}_{1,j}^T \mathbf{V}_x \boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}}\|_2 \leq 1,\ \forall \theta_l \in \Theta_{L-1},\ \forall l. \tag{32}$$

We note that an upper-bound for the constraint in (32) can be achieved as follows

$$\|\mathbf{W}_{L-2,j}^T \dots \mathbf{W}_{1,j}^T \mathbf{V}_x \boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}}\|_2 \leq \|\mathbf{W}_{L-2,j}\|_2 \dots \|\mathbf{W}_{1,j}\|_2 \|\mathbf{V}_x \boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}}\|_2 \leq {t_j^*}^{L-2} \|\boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}}\|_2,$$

where the last inequality follows from the constraint on each layer weight's norm, i.e., $\|\mathbf{W}_{l,j}\|_F \leq t_j^*$. The equality can be reached when the layer weights are

$$\mathbf{W}_l = t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T\ \forall l \in [L-2],$$

where $\{\boldsymbol{\rho}_{l,j}\}_{l=1}^{L-2}$ is a set of arbitrary unit norm vectors and $\boldsymbol{\rho}_0 = \mathbf{V}_x \boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}} / \|\mathbf{V}_x \boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}}\|_2$. Hence, we can rewrite (32) as

$$\max_{\boldsymbol{\lambda}} \tilde{\boldsymbol{\lambda}}^T \boldsymbol{\Sigma}_x \tilde{\mathbf{w}}_r^* \text{ s.t. } {t_j^*}^{L-2} \|\boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}}\|_2 \leq 1, \forall j \in [m]. \tag{33}$$

Therefore, the maximum objective value is achieved when $\boldsymbol{\Sigma}_x^T \tilde{\boldsymbol{\lambda}} = c_1 \tilde{\mathbf{w}}_r^*$ for some $c_1 > 0$, which yields the following set of optimal layer weight matrices

$$\mathbf{W}_{l,j}^* = \begin{cases} t_j^* \frac{\mathbf{V}_x \tilde{\mathbf{w}}_r^*}{\|\tilde{\mathbf{w}}_r^*\|_2} \boldsymbol{\rho}_{1,j}^T & \text{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T & \text{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \text{if } l = L-1 \end{cases}, \tag{34}$$

where $\boldsymbol{\rho}_{l,j} \in \mathbb{R}^{m_l}$ such that $\|\boldsymbol{\rho}_{l,j}\|_2 = 1,\ \forall l \in [L-2], \forall j \in [m]$. This shows that the weight matrices are rank-one and align with each other. Therefore, an arbitrary set of unit norm vectors, i.e., $\{\boldsymbol{\rho}_{l,j}\}_{l=1}^{L-2}$ can be chosen to achieve the maximum dual objective.

□

Theorem 3.2. *Let* $\{\mathbf{X}, \mathbf{y}\}$ *be feasible for* (15)*, then strong duality holds for finite width networks.*

Proof of Theorem 3.2. We first select a set of unit norm vectors, i.e., $\{\boldsymbol{\rho}_{l,j}\}_{l=1}^{L-2}$, to construct weight matrices $\{\mathbf{W}_{l,j}^e\}_{l=1}^{L-1}$ that satisfies (34). Then, the dual of (15) can be written as

$$\begin{aligned} D_e^* &= \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \\ &\text{s.t. } |(\mathbf{X}\mathbf{W}_{1,j}^e \dots \mathbf{w}_{L-1,j}^e)^T \boldsymbol{\lambda}| \leq 1,\ \forall j \in [m] \end{aligned}.$$

Then, we have

$$\begin{aligned} P^* = \min_{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \quad &\geq \quad \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \\ \text{s.t. } |(\mathbf{X}\mathbf{W}_1 \ldots \mathbf{w}_{L-1})^T \boldsymbol{\lambda}| \leq 1 \quad & \quad\quad \text{s.t. } |(\mathbf{X}\mathbf{W}_1 \ldots \mathbf{w}_{L-1})^T \boldsymbol{\lambda}| \leq 1, \; \forall \theta_l \in \Theta_{L-1} \\ &= \quad \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \\ & \quad\quad \text{s.t. } |(\mathbf{X}\mathbf{W}^e_{1,j} \ldots \mathbf{w}^e_{L-1,j})^T \boldsymbol{\lambda}| \leq 1, \; \forall j \\ &= \quad D_e^* = D^* = D_m^*, \end{aligned} \tag{35}$$

where the first inequality follows from changing the order of min-max to obtain a lower bound and the first equality follows from the fact that $\{\mathbf{W}^e_{l,j}\}_{l=1}^{L-1}$ maximizes the dual problem. Furthermore, we have the following relation between the primal problems

$$\begin{aligned} P_e^* = \min_{\mathbf{w}_L} \|\mathbf{w}_L\|_1 \quad &\geq \quad P^* = \min_{\{\theta_l\}_{l=1}^{L} \in \Theta_{L-1}} \|\mathbf{w}_L\|_1 \\ \text{s.t. } \sum_{j=1}^m \mathbf{X}\mathbf{W}^e_{1,j} \ldots \mathbf{w}^e_{L-1.j} w_{L,j} = \mathbf{y} \quad & \quad\quad \text{s.t. } \sum_{j=1}^m \mathbf{X}\mathbf{W}_{1,j} \ldots \mathbf{w}_{L-1,j} w_{L,j} = \mathbf{y}, \end{aligned} \tag{36}$$

where the inequality follows from the fact that the original problem has infinite width in each layer. Now, notice that the optimization problem on the left hand side of (36) is convex since it is an ℓ_1-norm minimization problem with linear equality constraints. Therefore, strong duality holds for this problem, i.e., $P_e^* = D_e^*$ and we have $P_e^* \geq P^* = P_m^* \geq D_e^* = D^* = D_m^*$. Using this result along with (35), we prove that strong duality holds, i.e., $P_e^* = P^* = P_m^* = D_e^* = D^* = D_m^*$.

□

Corollary 3.1. *Theorem 3.1 implies that deep linear networks can obtain a scaled version of* $\mathbf{y}$ *using only the first layer, i.e.,* $\mathbf{X}\mathbf{W}_1\boldsymbol{\rho}_1 = c\mathbf{y}$*, where* $c > 0$*. Therefore, the remaining layers do not contribute to the expressive power of the network.*

Proof of Corollary 3.1. The proof directly follows from (34). □

Theorem 3.3. *Optimal layer weights for* (16) *are*

$$\mathbf{W}^*_{l,j} = \begin{cases} t_j^* \frac{\mathbf{X}^T \mathcal{P}_{\mathbf{X},\beta}(\mathbf{y})}{\|\mathbf{X}^T \mathcal{P}_{\mathbf{X},\beta}(\mathbf{y})\|_2} \boldsymbol{\rho}_{1,j}^T & \textit{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T & \textit{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \textit{if } l = L-1 \end{cases},$$

where $\mathcal{P}_{\mathbf{X},\beta}(\cdot)$ *projects to* $\left\{\mathbf{u} \in \mathbb{R}^n \mid \|\mathbf{X}^T\mathbf{u}\|_2 \leq \beta t_j^{*^{2-L}}\right\}$.

Proof of Theorem 3.3. Using Lemma A.3 and Proposition 3.1, we have the following dual for (16)

$$\max_{\boldsymbol{\lambda}} -\frac{1}{2}\|\boldsymbol{\lambda} - \mathbf{y}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2 \text{ s.t. } \|(\mathbf{X}\mathbf{W}_{1,j} \ldots \mathbf{W}_{L-2,j})^T \boldsymbol{\lambda}\|_2 \leq \beta, \; \forall \theta_l \in \Theta_{L-1}, \; \forall l, j,$$

where $\Theta_{L-1} = \{\theta_1, \ldots, \theta_{L-1} | \|\mathbf{w}_{L-1,j}\|_2 \leq 1, \; \|\mathbf{W}_{l,j}\|_F \leq t_j^*, \; \forall l \in [L-2], \forall j \in [m]\}$. Then, the weight matrices that maximize the value of the constraint can be described as

$$\mathbf{W}^*_{l,j} = \begin{cases} t_j^* \frac{\mathbf{X}^T \mathcal{P}_{\mathbf{X},\beta}(\mathbf{y})}{\|\mathbf{X}^T \mathcal{P}_{\mathbf{X},\beta}(\mathbf{y})\|_2} \boldsymbol{\rho}_{1,j}^T & \text{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T & \text{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \text{if } l = L-1 \end{cases}.$$

where $\mathcal{P}_{\mathbf{X},\beta}(\cdot)$ projects its input to $\left\{\mathbf{u} \in \mathbb{R}^n \mid \|\mathbf{X}^T\mathbf{u}\|_2 \leq \beta t_j^{*^{2-L}}\right\}$. □

Corollary 3.2. *Theorem 3.2 also shows that strong duality holds for the training problem in* (16).

Proof of Corollary 3.2. The proof directly follows from the analysis in this section and Theorem 3.2. □

Theorem 3.4. *Optimal layer weight for* (17) *are*

$$\mathbf{W}_{l,j}^* = \begin{cases} t_j^* \tilde{\mathbf{v}}_{w,j} \boldsymbol{\rho}_{1,j}^T & \textit{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T & \textit{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \textit{if } l = L-1 \end{cases},$$

where $j \in [K]$, $\tilde{\mathbf{v}}_{w,j}$ is the j^{th} maximal right singular vector of $\boldsymbol{\Lambda}^{*^T}\mathbf{X}$ and $\{\boldsymbol{\rho}_{l,j}\}_{l=1}^{L-2}$ are arbitrary unit norm vectors such that $\boldsymbol{\rho}_{l,j}^T \boldsymbol{\rho}_{l,k} = 0,\ \forall j \neq k$.

Proof of Theorem 3.4. Using Proposition 3.1 and Lemma A.4, we obtain the following dual problem

$$\begin{aligned} D &= \max_{\boldsymbol{\Lambda}} \mathbf{tr}(\boldsymbol{\Lambda}^T \mathbf{Y}) \text{ s.t. } \|\boldsymbol{\Lambda}^T \mathbf{X} \mathbf{W}_{1,j} \dots \mathbf{W}_{L-2,j} \mathbf{w}_{L-1,j}\|_2 \leq 1,\ \forall \theta_l \in \Theta_{L-1},\ \forall j \in [m] \\ &= \max_{\boldsymbol{\Lambda}} \mathbf{tr}(\boldsymbol{\Lambda}^T \mathbf{Y}) \text{ s.t. } \sigma_{max}(\boldsymbol{\Lambda}^T \mathbf{X} \mathbf{W}_{1,j} \dots \mathbf{W}_{L-2,j}) \leq 1,\ \forall \theta_l \in \Theta_{L-1},\ \forall j \in [m], \end{aligned} \tag{37}$$

where $\Theta_{L-1} = \{\theta_1, \dots, \theta_{L-1} | \|\mathbf{w}_{L-1,j}\|_2 \leq 1,\ \|\mathbf{W}_{l,j}\|_F \leq t_j^*,\ \forall l \in [L-2], \forall j \in [m]\}$.

It is straightforward to show that the optimal layer weights are the extreme points of the constraint in (37), which achieves the following upper-bound

$$\max_{\{\theta_l\}_{l=1}^{L-2} \in \Theta_{L-1}} \sigma_{max}(\boldsymbol{\Lambda}^T \mathbf{X} \mathbf{W}_{1,j} \dots \mathbf{W}_{L-2,j}) \leq \sigma_{max}(\boldsymbol{\Lambda}^T \mathbf{X}) t_j^{*^{L-2}}.$$

This upper-bound is achieved when the first $L-2$ layer weights are rank-one with the singular value t_j^* by Proposition 3.1. Additionally, the left singular vector of $\mathbf{W}_{1,j}$ needs to align with one of the maximum right singular vectors of $\boldsymbol{\Lambda}^T \mathbf{X}$. Since the upper-bound for the objective is achievable for any $\boldsymbol{\Lambda}$, we can maximize the objective value, as in (14), by choosing a matrix $\boldsymbol{\Lambda}$ such that

$$\boldsymbol{\Lambda}^T \mathbf{U}_x \boldsymbol{\Sigma}_x = \mathbf{V}_w \begin{bmatrix} t_j^{*^{2-L}} \mathbf{I}_{r_w} & \mathbf{0}_{r_x \times d - r_w} \\ \mathbf{0}_{k - r_w \times r_x} & \mathbf{0}_{k - r_w \times d - r_w} \end{bmatrix} \mathbf{U}_w^T,$$

where $\tilde{\mathbf{W}}_r^* = \mathbf{V}_x^T \mathbf{W}_r^* = \mathbf{U}_w \boldsymbol{\Sigma}_w \mathbf{V}_w^T$. Thus, a set of optimal layer weights can be formulated as follows

$$\mathbf{W}_{l,j} = \begin{cases} t_j^* \tilde{\mathbf{v}}_{w,j} \boldsymbol{\rho}_{1,j}^T & \text{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T & \text{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \text{if } l = L-1 \end{cases}, \tag{38}$$

where $\tilde{\mathbf{v}}_{w,j}$ is the j^{th} maximal right singular vector of $\boldsymbol{\Lambda}^{T^*}\mathbf{X}$ and we select a set of unit norm vectors $\{\boldsymbol{\rho}_{l,j}\}_{l=1}^{L-2}$ such that $\boldsymbol{\rho}_{l,j}^T \boldsymbol{\rho}_{l,k} = 0,\ \forall j \neq k$. We now note that since there exist at most K singular vectors of $\boldsymbol{\Lambda}^T \mathbf{X}$ with non-zeros singular values, we can replace m with K without loss of generality.

□

Theorem 3.5. *Let $\{\mathbf{X}, \mathbf{Y}\}$ be feasible for* (17)*, then strong duality holds for finite width networks.*

Proof of Theorem 3.5. We first select a set of unit norm vectors, i.e., $\{\boldsymbol{\rho}_{l,j}\}_{l=1}^{L-2}$, to construct weight matrices $\{\mathbf{W}_{l,j}^e\}_{l=1}^{L-1}$ that satisfies (38). Then, we have

$$\begin{array}{rlcl} P^* = & \displaystyle\min_{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}} \max_{\boldsymbol{\Lambda}} \mathbf{tr}(\boldsymbol{\Lambda}^T \mathbf{Y}) & \geq & \displaystyle\max_{\boldsymbol{\Lambda}} \mathbf{tr}(\boldsymbol{\Lambda}^T \mathbf{Y}) \\ & \text{s.t. } \sigma_{max}(\boldsymbol{\Lambda}^T \mathbf{X} \mathbf{W}_{1,j} \dots \mathbf{W}_{L-2,j}) \leq 1, \forall j & & \text{s.t. } \sigma_{max}(\boldsymbol{\Lambda}^T \mathbf{X} \mathbf{W}_{1,j} \dots \mathbf{W}_{L-2,j}) \leq 1,\ \forall j, \forall \theta_l \in \Theta_{L-1} \\ & & = & \displaystyle\max_{\boldsymbol{\Lambda}} \mathbf{tr}(\boldsymbol{\Lambda}^T \mathbf{Y}) \\ & & & \text{s.t. } \sigma_{max}(\boldsymbol{\Lambda}^T \mathbf{X} \mathbf{W}_{1,j}^e \dots \mathbf{W}_{L-2,j}^e) \leq 1,\ \forall j \\ & & = & D_e^* = D^* = D_m^*, \end{array} \tag{39}$$

where the first inequality follows from changing the order of min-max to obtain a lower bound and the first equality follows from the fact that $\{\mathbf{W}_{l,j}^e\}_{l=1}^{L-1}$ maximizes the dual problem. Furthermore, we have the following relation between the primal problems

$$\begin{aligned} P_e^* = \min_{\mathbf{W}_L} \sum_{j=1}^m \|\mathbf{w}_{L,j}\|_2 \quad & \geq \quad P^* = \min_{\{\theta_l\}_{l=1}^L \in \Theta_{L-1}} \sum_{j=1}^m \|\mathbf{w}_{L,j}\|_2 \\ \text{s.t. } \sum_{j=1}^m \mathbf{X}\mathbf{W}_{1,j}^e \ldots \mathbf{W}_{L-1,j}^e \mathbf{w}_{L,j}^T = \mathbf{Y} \quad & \quad\quad \text{s.t. } \sum_{j=1}^m \mathbf{X}\mathbf{W}_{1,j} \ldots \mathbf{w}_{L-1,j}\mathbf{w}_{L,j}^T = \mathbf{Y}, \end{aligned} \tag{40}$$

where the inequality follows from the fact that the original problem has infinite width in each layer. Now, notice that the optimization problem on the left hand side of (40) is convex since it is an ℓ_2-norm minimization problem with linear equality constraints. Therefore, strong duality holds for this problem, i.e., $P_e^* = D_e^*$ and we have $P_e^* \geq P^* = P_m^* \geq D_e^* = D^* = D_m^*$. Using this result along with (39), we prove that strong duality holds, i.e., $P_e^* = P^* = P_m^* = D_e^* = D^* = D_m^*$. □

Theorem 3.6. *Optimal layer weights for* (18) *are*

$$\mathbf{W}_{l,j}^* = \begin{cases} t_j^* \tilde{\mathbf{v}}_{x,j} \boldsymbol{\rho}_{1,j}^T & \textit{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T & \textit{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \textit{if } l = L-1 \end{cases},$$

where $j \in [K]$, $\tilde{\mathbf{v}}_{x,j}$ is a maximal right singular vector of $\mathcal{P}_{\mathbf{X},\beta}(\mathbf{Y})^T\mathbf{X}$ and $\mathcal{P}_{\mathbf{X},\beta}(\cdot)$ projects to $\{\mathbf{U} \in \mathbb{R}^{n\times k} \mid \sigma_{max}(\mathbf{U}^T\mathbf{X}) \leq \beta {t_j^*}^{2-L}\}$. Additionally, $\boldsymbol{\rho}_{l,j}$'s is an orthonormal set. Therefore, the rank of each hidden layer is determined by β as in Remark 2.1.

Proof of Theorem 3.6. Using Lemma A.4 and Proposition 3.1, we have the following dual for (18)

$$\max_{\mathbf{\Lambda}} -\frac{1}{2}\|\mathbf{\Lambda} - \mathbf{Y}\|_F^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2 \text{ s.t. } \sigma_{max}(\mathbf{\Lambda}^T\mathbf{X}\mathbf{W}_{1,j}\ldots\mathbf{W}_{L-2,j}) \leq \beta,\ \forall \theta_l \in \Theta_{L-1},\ \forall j \in [m],$$

where we define $\Theta_{L-1} = \{\theta_1, \ldots, \theta_{L-1} | \|\mathbf{w}_{L-1,j}\|_2 \leq 1,\ \|\mathbf{W}_{l,j}\|_F \leq t_j^*,\ \forall l \in [L-2], \forall j \in [m]\}$. Then, as in (38), a set of optimal layer weights is

$$\mathbf{W}_{l,j}^* = \begin{cases} t_j^* \tilde{\mathbf{v}}_{x,j} \boldsymbol{\rho}_{1,j}^T & \text{if } l = 1 \\ t_j^* \boldsymbol{\rho}_{l-1,j} \boldsymbol{\rho}_{l,j}^T & \text{if } 1 < l \leq L-2 \\ \boldsymbol{\rho}_{L-2,j} & \text{if } l = L-1 \end{cases},$$

where $\tilde{\mathbf{v}}_{x,j}$ is a maximal right singular vector of $\mathcal{P}_{\mathbf{X},\beta}(\mathbf{Y})^T\mathbf{X}$ and $\mathcal{P}_{\mathbf{X},\beta}(\cdot)$ projects its input to the set $\{\mathbf{U} \in \mathbb{R}^{n\times k} \mid \sigma_{max}(\mathbf{U}^T\mathbf{X}) \leq \beta {t_j^*}^{2-L}\}$. Additionally, $\boldsymbol{\rho}_{l,j}$'s is an orthonormal set. □

A.8. Proofs for deep ReLU networks

Theorem 4.1. *Let $\mathbf{X}$ be a rank-one matrix such that $\mathbf{X} = \mathbf{c}\mathbf{a}_0^T$, where $\mathbf{c} \in \mathbb{R}_+^n$ and $\mathbf{a}_0 \in \mathbb{R}^d$, then strong duality holds and the optimal weights are*

$$\mathbf{W}_{l,j} = \frac{\boldsymbol{\phi}_{l-1,j}}{\|\boldsymbol{\phi}_{l-1,j}\|_2}\boldsymbol{\phi}_{l,j}^T,\ \forall l \in [L-2],\ \mathbf{w}_{L-1,j} = \frac{\boldsymbol{\phi}_{L-2,j}}{\|\boldsymbol{\phi}_{L-2,j}\|_2},$$

where $\boldsymbol{\phi}_{0,j} = \mathbf{a}_0$ and $\{\boldsymbol{\phi}_{l,j}\}_{l=1}^{L-2}$ is a set of vectors such that $\boldsymbol{\phi}_{l,j} \in \mathbb{R}_+^{m_l}$ and $\|\boldsymbol{\phi}_{l,j}\|_2 = t_j^*,\ \forall l \in [L-2], \forall j \in [m]$.

Proposition 1. *First $L-2$ hidden layer weight matrices in* (19) *have the same operator and Frobenius norms.*

Proof of Proposition 1. Let us first denote the sum of the norms for the first $L-2$ layer as t_j, i.e., $t_j = \sum_{l=1}^{L-2} t_{l,j}$, where $t_{l,j} = \|\mathbf{W}_{l,j}\|_2 = \|\mathbf{W}_{l,j}\|_F$ since the upper-bound is achieved when the matrices are rank-one. Then, to find the extreme points (see the details in Proof of Theorem 4.1), we need to solve the following problem

$$\operatorname*{argmax}_{\{\theta_l\}_{l=1}^{L-2}} |{\boldsymbol{\lambda}^*}^T \mathbf{c}| \, \|\mathbf{a}_{L-2}\|_2 = \operatorname*{argmax}_{\{\theta_l\}_{l=1}^{L-2} \in \Theta_{L-1}} |{\boldsymbol{\lambda}^*}^T \mathbf{c}| \, \|(\mathbf{a}_{L-3,j}^T \mathbf{W}_{L-2,j})_+\|_2$$

where we use $\mathbf{a}_{L-2,j}^T = (\mathbf{a}_{L-3,j}^T \mathbf{W}_{L-2,j})_+$. Since $\|\mathbf{W}_{L-2,j}\|_F = t_{L-2,j} = t_j - \sum_{l=1}^{L-3} t_{l,j}$, the objective value above becomes $|{\boldsymbol{\lambda}^*}^T \mathbf{c}| \, \|(\mathbf{a}_{L-3,j}\|_2 \left(t_j - \sum_{l=1}^{L-3} t_{l,j}\right)$. Applying this step to all the remaining layer weights gives the following problem

$$\operatorname*{argmax}_{\{t_{l,j}\}_{l=1}^{L-3}} |{\boldsymbol{\lambda}^*}^T \mathbf{c}| \, \|\mathbf{a}_0\|_2 \left(t_j - \sum_{l=1}^{L-3} t_{l,j}\right) \prod_{j=1}^{L-3} t_{l,j} \text{ s.t. } \sum_{l=1}^{L-3} t_{l,j} \leq t_j, \; t_{l,j} \geq 0.$$

Then, the proof directly follows from Proof of Proposition 3.1. □

Proof of Theorem 4.1. Using Lemma A.3 and Proposition 1, this problem can be equivalently stated as

$$\begin{aligned} \min_{\{\theta_l\}_{l=1}^{L} \in \Theta_{L-1}} \|\mathbf{w}_L\|_1 \text{ s.t. } & \mathbf{A}_{l,j} = (\mathbf{A}_{l-1,j}\mathbf{W}_{l,j})_+, \; \forall l \in [L-1], \forall j \in [m] \\ & \mathbf{A}_{L-1}\mathbf{w}_L = \mathbf{y} \end{aligned}, \tag{41}$$

which also has the following dual form

$$\begin{aligned} P^* = \min_{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}} & \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \\ & \text{s.t. } \|\mathbf{A}_{L-1}^T \boldsymbol{\lambda}\|_\infty \leq 1 \end{aligned}. \tag{42}$$

Notice that we remove the recursive constraint in (42) for notational simplicity, however, $\mathbf{A}_{L-1}$ is still a function of all the layer weights except $\mathbf{w}_L$. Changing the order of min-max in (42) gives

$$P^* \geq D^* = \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y} \text{ s.t. } \|\mathbf{A}_{L-1}^T \boldsymbol{\lambda}\|_\infty \leq 1, \; \forall \theta_l \in \Theta_{L-1}, \; \forall l \in [L-1]. \tag{43}$$

The dual of the semi-infinite problem in (43) is given by

$$\begin{aligned} & \min \|\boldsymbol{\mu}\|_{TV} \\ & \text{s.t. } \int_{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}} \left(\mathbf{A}_{L-2}\mathbf{w}_{L-1}\right)_+ d\boldsymbol{\mu}(\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_{L-1}}) = \mathbf{y} \, , \end{aligned} \tag{44}$$

where $\boldsymbol{\mu}$ is a signed Radon measure and $\|\cdot\|_{TV}$ is the total variation norm. We emphasize that (44) has infinite width in each layer, however, an application of Caratheodory's theorem shows that the measure $\boldsymbol{\mu}$ in the integral can be represented by finitely many (at most $n+1$) Dirac delta functions (Rosset et al., 2007). Thus, we choose

$$\boldsymbol{\mu} = \sum_{j=1}^{m} \delta(\mathbf{W}_1 - \mathbf{W}_{1,j}, \ldots, \mathbf{w}_{L-1} - \mathbf{w}_{L-1,j}) w_{L,j},$$

where $\delta(\cdot)$ is the Dirac delta function and the superscript indicates a particular choice for the corresponding layer weight. This selection of $\boldsymbol{\mu}$ yields the following problem

$$\begin{aligned} P_m^* = & \min_{\{\theta_l\}_{l=1}^{L}} \|\mathbf{w}_L\|_1 \\ & \text{s.t.} \sum_{j=1}^{m} \left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)_+ w_{L,j} = \mathbf{y}, \; \theta_l \in \Theta_{L-1}, \; \forall l \in [L-1] \end{aligned}. \tag{45}$$

Here, we note that the model in (45) has the same expressive power with ReLU networks, thus, we have $P^* = P_m^*$.

As a consequence of (43), we can characterize the optimal layer weights for (45) as the extreme points that solve

$$\underset{\{\theta_l\}_{l=1}^{L-1}\in\Theta_{L-1}}{\operatorname{argmax}} \quad |{\boldsymbol{\lambda}^*}^T(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j})_+| \tag{46}$$

where $\boldsymbol{\lambda}^*$ is the optimal dual parameter. Since we assume that $\mathbf{X} = \mathbf{c}\mathbf{a}_0^T$ with $\mathbf{c} \in \mathbb{R}_+^n$, we have $\mathbf{A}_{L-2,j} = \mathbf{c}\mathbf{a}_{L-2,j}^T$, where $\mathbf{a}_{l,j}^T = (\mathbf{a}_{l-1,j}^T\mathbf{W}_{l,j})_+$, $\mathbf{a}_{l,j} \in \mathbb{R}_+^{m_l}$ and $\forall l \in [L-1]$, $\forall j \in [m]$. Based on this observation, we have $\mathbf{w}_{L-1,j} = \mathbf{a}_{L-2,j}/\|\mathbf{a}_{L-2,j}\|_2$, which reduces (46) to the following

$$\underset{\{\theta_l\}_{l=1}^{L-2}\in\Theta_{L-1}}{\operatorname{argmax}} \quad |{\boldsymbol{\lambda}^*}^T\mathbf{c}|\,\|\mathbf{a}_{L-2,j}\|_2 \tag{47}$$

We then apply the same approach to all the remaining layer weights. However, notice that each neuron for the first $L-2$ layers must have bounded Frobenius norms due to the norm constraint. If we denote the optimal ℓ_2 norms vector for the neuron in the l^{th} layer as $\boldsymbol{\phi}_{l,j} \in \mathbb{R}_+^{m_l}$, then we have the following formulation for the layer weights that solve (46)

$$\mathbf{W}_{l,j} = \frac{\boldsymbol{\phi}_{l-1,j}}{\|\boldsymbol{\phi}_{l-1,j}\|_2}\boldsymbol{\phi}_{l,j}^T,\ \forall l \in [L-2],\ \mathbf{w}_{L-1,j} = \frac{\boldsymbol{\phi}_{L-2,j}}{\|\boldsymbol{\phi}_{L-2}\|_2}, \tag{48}$$

where $\boldsymbol{\phi}_{0,j} = \mathbf{a}_0$, $\{\boldsymbol{\phi}_{l,j}\}_{l=1}^{L-2}$ is a set of nonnegative vectors satisfying $\|\boldsymbol{\phi}_{l,j}\|_2 = t_j^*,\ \forall l \in [L-2]$. Therefore, the set of weights in (48) are optimal for (19). Moreover, as a direct consequence of Theorem 3.2, strong duality holds for this case as well.

□

Theorem 4.2. *Let* $\mathbf{X}$ *be a matrix such that* $\mathbf{X} = \mathbf{c}\mathbf{a}_0^T$*, where* $\mathbf{c} \in \mathbb{R}^n$ *and* $\mathbf{a}_0 \in \mathbb{R}^d$*. Then, when* $L = 2$*, a set of optimal solutions to* (19) *is* $\{(\mathbf{w}_i, b_i)\}_{i=1}^m$*, where* $\mathbf{w}_i = s_i\frac{\mathbf{a}_0}{\|\mathbf{a}_0\|_2}, b_i = -s_i c_i\|\mathbf{a}_0\|_2$ *with* $s_i = \pm 1, \forall i \in [m]$*.*

Proof of Theorem 4.2. Given $\mathbf{X} = \mathbf{c}\mathbf{a}_0^T$, all possible extreme points can be characterized as follows

$$\begin{aligned}\underset{b,\mathbf{w}:\|\mathbf{w}\|_2=1}{\operatorname{argmax}} |\boldsymbol{\lambda}^T(\mathbf{X}\mathbf{w}+b\mathbf{1})_+| &= \underset{b,\mathbf{w}:\|\mathbf{w}\|_2=1}{\operatorname{argmax}} |\boldsymbol{\lambda}^T\left(\mathbf{c}\mathbf{a}_0^T\mathbf{w}+b\mathbf{1}\right)_+| \\ &= \underset{b,\mathbf{w}:\|\mathbf{w}\|_2=1}{\operatorname{argmax}} \left|\sum_{i=1}^n \lambda_i\left(c_i\mathbf{a}_0^T\mathbf{w}+b\right)_+\right|\end{aligned}$$

which can be equivalently stated as

$$\underset{b,\mathbf{w}:\|\mathbf{w}\|_2=1}{\operatorname{argmax}} \sum_{i\in\mathcal{S}}\lambda_i c_i\mathbf{a}_0^T\mathbf{w} + \sum_{i\in\mathcal{S}}\lambda_i b \text{ s.t. } \begin{cases} c_i\mathbf{a}_0^T\mathbf{w}+b \geq 0, \forall i \in \mathcal{S} \\ c_j\mathbf{a}_0^T\mathbf{w}+b \leq 0, \forall j \in \mathcal{S}^c \end{cases},$$

which shows that $\mathbf{w}$ must be either positively or negatively aligned with $\mathbf{a}_0$, i.e., $\mathbf{w} = s\frac{\mathbf{a}_0}{\|\mathbf{a}_0\|_2}$, where $s = \pm 1$. Thus, b must be in the range of $[\max_{i\in\mathcal{S}}(-sc_i\|\mathbf{a}_0\|_2),\ \min_{k\in\mathcal{S}^c}(-sc_k\|\mathbf{a}_0\|_2)]$ Using these observations, extreme points can be formulated as follows

$$\mathbf{w}_\lambda = \begin{cases} \frac{\mathbf{a}_0}{\|\mathbf{a}_0\|_2} & \text{if } \sum_{i\in\mathcal{S}}\lambda_i c_i \geq 0 \\ \frac{-\mathbf{a}_0}{\|\mathbf{a}_0\|_2} & \text{otherwise} \end{cases} \text{ and } b_\lambda = \begin{cases} \min_{k\in\mathcal{S}^c}(-s_\lambda c_k\|\mathbf{a}_0\|_2) & \text{if } \sum_{i\in\mathcal{S}}\lambda_i \geq 0 \\ \max_{i\in\mathcal{S}}(-s_\lambda c_i\|\mathbf{a}_0\|_2) & \text{otherwise} \end{cases},$$

where $s_\lambda = \operatorname{sign}(\sum_{i\in\mathcal{S}}\lambda_i c_i)$. □

Proposition 4.1. *Theorem 4.1 still holds when we add a bias term to the last hidden layer, i.e.,* $\sum_j\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j} + \mathbf{1}_n b_j\right)_+ w_{L,j} = \mathbf{y}$.

Proof of Proposition 4.1. Here, we add biases to the neurons in the last hidden layer of (19). For this case, all the equations in (41)-(43) hold except notational changes due to the bias term. Thus, (46) changes as

$$\begin{aligned}\underset{\{\theta_l\}_{l=1}^{L-1}\in\Theta_{L-1},b_j}{\operatorname{argmax}} |{\boldsymbol{\lambda}^*}^T(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}+b_j\mathbf{1}_n)_+| &= \underset{\{\theta_l\}_{l=1}^{L-1}\in\Theta_{L-1},b_j}{\operatorname{argmax}} |{\boldsymbol{\lambda}^*}^T\left(\mathbf{c}\mathbf{a}_{L-2,j}^T\mathbf{w}_{L-1,j}+b_j\mathbf{1}_n\right)_+| \\ &= \underset{\{\theta_l\}_{l=1}^{L-2}\in\Theta_{L-1},b_j}{\operatorname{argmax}} \left|\sum_{i=1}^n \lambda_i^*\left(c_i\mathbf{a}_{L-2,j}^T\mathbf{w}_{L-1,j}+b_j\right)_+\right| \end{aligned} \tag{49}$$

which can also be written as

$$\underset{\{\theta_l\}_{l=1}^{L-1}\in\Theta_{L-1},b_j}{\operatorname{argmax}} \sum_{i\in\mathcal{S}}\lambda_i^* c_i \mathbf{a}_{L-2,j}^T \mathbf{w}_{L-1,j} + \sum_{i\in\mathcal{S}}\lambda_i^* b_j \text{ s.t. } \begin{cases} c_i \mathbf{a}_{L-2,j}^T \mathbf{w}_{L-1} + b_j \geq 0, \forall i \in \mathcal{S} \\ c_j \mathbf{a}_{L-2,j}^T \mathbf{w}_{L-1,j} + b_j \leq 0, \forall j \in \mathcal{S}^c \end{cases},$$

where $\mathcal{S}$ and $\mathcal{S}^c$ are the indices for which ReLU is active and inactive, respectively. This shows that $\mathbf{w}_{L-1,j}$ must be $\mathbf{w}_{L-1,j} = \pm 1 \frac{\mathbf{a}_{L-2,j}}{\|\mathbf{a}_{L-2,j}\|_2}$ and $b_j \in [\max_{i\in\mathcal{S}}(-c_i\|\mathbf{a}_{L-2,j}\|_2),\ \min_{k\in\mathcal{S}^c}(-c_k\|\mathbf{a}_{L-2,j}\|_2)]$. Then, we obtain the following

$$\mathbf{w}_{L-1,j}^* = \begin{cases} \frac{\mathbf{a}_{L-2,j}}{\|\mathbf{a}_{L-2,j}\|_2} & \text{if } \sum_{i\in\mathcal{S}}\lambda_i^* c_i \geq 0 \\ \frac{-\mathbf{a}_{L-2,j}}{\|\mathbf{a}_{L-2,j}\|_2} & \text{otherwise} \end{cases} \text{ and } b_j^* = \begin{cases} \min_{k\in\mathcal{S}^c}(-s_{\lambda^*} c_k \|\mathbf{a}_{L-2,j}\|_2) & \text{if } \sum_{i\in\mathcal{S}}\lambda_i^* \geq 0 \\ \max_{i\in\mathcal{S}}(-s_{\lambda^*} c_i \|\mathbf{a}_{L-2,j}\|_2) & \text{otherwise} \end{cases}, \tag{50}$$

where $s_{\lambda^*} = \operatorname{sign}(\sum_{i\in\mathcal{S}}\lambda_i^* c_i)$. This result reduces (49) to the following problem

$$\underset{\{\theta_l\}_1^{L-2}\in\Theta_{L-1}}{\operatorname{argmax}} \quad |C(\boldsymbol{\lambda}^*, \mathbf{c})| \, \|\mathbf{a}_{L-2,j}\|_2,$$

where $C(\boldsymbol{\lambda}^*, \mathbf{c})$ is constant scalar independent of $\{\mathbf{W}_{l,j}\}_{l=1}^{L-2}$. Hence, this problem and its solutions are the same with (47) and (48), respectively.

□

Corollary 4.1. *As a result of Theorem 4.2, when we have one dimensional data, i.e.,* $\mathbf{x} \in \mathbb{R}^n$*, an optimal solution to* (19) *can be formulated as* $\{(w_i, b_i)\}_{i=1}^m$*, where* $w_i = s_i$*,* $b_i = -s_i x_i$ *with* $s_i = \pm 1, \forall i \in [m]$*. Therefore, the optimal network output has kinks only at the input data points, i.e., the output function is in the following form:* $f_{\theta,2}(\hat{x}) = \sum_i (\hat{x} - x_i)_+$*. Hence, the network output becomes a linear spline interpolation.*

Corollary 4.2. *As a result of Theorem 4.2 and Proposition 4.1, for one dimensional data, i.e.,* $\mathbf{x} \in \mathbb{R}^n$*, the optimal network output has kinks only at the input data points, i.e., the output function is in the following form:* $f_{\theta,L}(\hat{x}) = \sum_i (\hat{x} - x_i)_+$*. Therefore, the optimal network output is a linear spline interpolation.*

Proof of Corollary 4.1 and 4.2. Let us particularly consider the input sample $\mathbf{a}_0$. Then, the activations of the network defined by (48) and (50) are

$$\begin{aligned}
\mathbf{a}_{1,j}^T &= (\mathbf{a}_0^T \mathbf{W}_1)_+ = \left(\mathbf{a}_0^T \frac{\mathbf{a}_0}{\|\mathbf{a}_0\|_2} \boldsymbol{\phi}_1^T\right)_+ = \|\mathbf{a}_0\|_2 \boldsymbol{\phi}_{1,j}^T \\
\mathbf{a}_{2,j}^T &= (\mathbf{a}_{1,j}^T \mathbf{W}_2)_+ = \left(\mathbf{a}_{1,j}^T \frac{\mathbf{a}_{1,j}}{\|\mathbf{a}_{1,j}\|_2} \boldsymbol{\phi}_{2,j}^T\right)_+ = \|\mathbf{a}_0\|_2 \|\boldsymbol{\phi}_{1,j}^T\|_2 \boldsymbol{\phi}_{2,j}^T \\
&\vdots \\
\mathbf{a}_{L-2,j}^T &= (\mathbf{a}_{L-3,j}^T \mathbf{W}_{L-2,j})_+ = \left(\mathbf{a}_{L-3,j}^T \frac{\mathbf{a}_{L-3,j}}{\|\mathbf{a}_{L-3}\|_2} \boldsymbol{\phi}_{L-2}^T\right)_+ = \|\mathbf{a}_0\|_2 \|\boldsymbol{\phi}_{1,j}^T\|_2 \dots \|\boldsymbol{\phi}_{L-3,j}^T\|_2 \boldsymbol{\phi}_{L-2,j}^T \\
a_{L-1,j} &= (\mathbf{a}_{L-2,j}^T \mathbf{w}_{L-1,j} + b)_+ = (\|\mathbf{a}_{L-2,j}\|_2 - \|\mathbf{a}_{L-2,j}\|_2)_+ = 0.
\end{aligned}$$

Thus, if we feed $c_i \mathbf{a}_0$ to the network, we get $a_{L-1,j} = (c_i\|\mathbf{a}_{L-2,j}\|_2 - c_i\|\mathbf{a}_{L-2,j}\|_2)_+ = 0$, where we use the fact that optimal biases are in the form of $b_j = -c_i\|\mathbf{a}_{L-2,j}\|_2$ as proved in (50). This analysis proves that the kink of each ReLU activation occurs exactly at one of the data points. □

Proposition 4.2. *Theorem 4.1 extends to deep ReLU networks with vector outputs, therefore, the optimal layer weights can be formulated as in Theorem 4.1.*

Proof of Proposition 4.2. For vector outputs, we have the following training problem

$$\min_{\{\theta_l\}_{l=1}^L} \frac{1}{2}\|f_{\theta,L}(\mathbf{X}) - \mathbf{Y}\|_F^2 + \frac{\beta}{2}\sum_{j=1}^m \sum_{l=1}^L \|\mathbf{W}_{l,j}\|_F^2.$$

After a suitable rescaling as in the previous case, the above problem has the following dual

$$P^* \geq D^* = \max_{\boldsymbol{\Lambda}} -\frac{1}{2}\|\boldsymbol{\Lambda} - \mathbf{Y}\|_F^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2 \text{ s.t. } \|\boldsymbol{\Lambda}^T \left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)_+\|_2 \leq \beta,\ \forall \theta_l \in \Theta_{L-1},\ \forall l \in [L-1], \forall j \in [m]. \tag{51}$$

Using (51), we can characterize the optimal layer weights as the extreme points that solve

$$\underset{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}}{\operatorname{argmax}} \quad \|\boldsymbol{\Lambda}^{*^T}(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j})_+\|_2, \tag{52}$$

where $\boldsymbol{\Lambda}^*$ is the optimal dual parameter. Since we assume that $\mathbf{X} = \mathbf{c}\mathbf{a}_0^T$ with $\mathbf{c} \in \mathbb{R}_+^n$, we have $\mathbf{A}_{L-2,j} = \mathbf{c}\mathbf{a}_{L-2,j}^T$, where $\mathbf{a}_{l,j}^T = (\mathbf{a}_{l-1,j}^T \mathbf{W}_{l,j})_+$, $\mathbf{a}_{l,j} \in \mathbb{R}_+^{m_l}$ and $\forall l \in [L-1]$. Based on this observation, we have $\mathbf{w}_{L-1,j} = \mathbf{a}_{L-2,j}/\|\mathbf{a}_{L-2,j}\|_2$, which reduces (52) to the following

$$\underset{\{\theta_l\}_{l=1}^{L-2} \in \Theta_{L-1}}{\operatorname{argmax}} \quad \|\boldsymbol{\Lambda}^{*^T}\mathbf{c}\|_2 \, \|\mathbf{a}_{L-2,j}\|_2.$$

Then, the rest of steps directly follow Theorem 4.1 yielding the following weight matrices

$$\mathbf{W}_{l,j} = \frac{\boldsymbol{\phi}_{l-1,j}}{\|\boldsymbol{\phi}_{l-1,j}\|_2}\boldsymbol{\phi}_{l,j}^T,\ \forall l \in [L-2],\ \mathbf{w}_{L-1,j} = \frac{\boldsymbol{\phi}_{L-2,j}}{\|\boldsymbol{\phi}_{L-2,j}\|_2},$$

where $\boldsymbol{\phi}_{0,j} = \mathbf{a}_0$, $\{\boldsymbol{\phi}_{l,j}\}_{l=1}^{L-2}$ is a set of nonnegative vectors satisfying $\|\boldsymbol{\phi}_{l,j}\|_2 = t_j^*,\ \forall l \in [L-2], \forall j \in [m]$. □

Theorem 4.3. *Let $\{\mathbf{X}, \mathbf{Y}\}$ be a dataset such that $\mathbf{X}\mathbf{X}^T = \mathbf{I}_n$ and $\mathbf{Y}$ is one-hot encoded, then a set of optimal solutions for the following regularized training problem*

$$\min_{\theta \in \Theta} \frac{1}{2}\|f_{\theta,L}(\mathbf{X}) - \mathbf{Y}\|_F^2 + \frac{\beta}{2}\sum_{j=1}^{m}\sum_{l=1}^{L}\|\mathbf{W}_{l,j}\|_F^2 \tag{20}$$

can be formulated as follows

$$\mathbf{W}_{l,j} = \begin{cases} \frac{\boldsymbol{\phi}_{l-1,j}}{\|\boldsymbol{\phi}_{l-1,j}\|_2}\boldsymbol{\phi}_{l,j}^T, & \textit{if } l \in [L-1] \\ \left(\|\boldsymbol{\phi}_{0,j}\|_2 - \beta\right)_+ \boldsymbol{\phi}_{l-1,j}\mathbf{e}_r^T & \textit{if } l = L \end{cases},$$

where $\boldsymbol{\phi}_{0,j} = \mathbf{X}^T\mathbf{y}_j$, $\{\boldsymbol{\phi}_{l,j}\}_{l=1}^{L-2}$ are vectors such that $\boldsymbol{\phi}_{l,j} \in \mathbb{R}_+^{m_l}$, $\|\boldsymbol{\phi}_{l,j}\|_2 = t_j^*$, and $\boldsymbol{\phi}_{l,i}^T\boldsymbol{\phi}_{l,j} = 0,\ \forall i \neq j$, Moreover, $\boldsymbol{\phi}_{L-1,j} = \mathbf{e}_j$ is the j^{th} ordinary basis vector.

Proof of Theorem 4.3. For vector outputs, we have the following training problem

$$P^* = \min_{\theta \in \Theta} \frac{1}{2}\|f_{\theta,L}(\mathbf{X}) - \mathbf{Y}\|_F^2 + \frac{\beta}{2}\sum_{j=1}^{m}\sum_{l=1}^{L}\|\mathbf{W}_{l,j}\|_F^2 \tag{53}$$

After a suitable rescaling as in the previous case, the above problem has the following dual

$$P^* \geq D^* = \max_{\boldsymbol{\lambda}} -\frac{1}{2}\left\|\boldsymbol{\Lambda} - \mathbf{Y}\right\|_F + \frac{1}{2}\left\|\mathbf{Y}\right\|_F \text{ s.t. } \|\boldsymbol{\Lambda}^T \left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)_+\|_2 \leq \beta,\ \forall \theta_l \in \Theta_{L-1},\ \forall l \in [L-1],\ \forall j \in [m]. \tag{54}$$

Using (54), we can characterize the optimal layer weights as the extreme points that solve

$$\underset{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}}{\operatorname{argmax}} \quad \|\boldsymbol{\Lambda}^{*^T}(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j})_+\|_2, \tag{55}$$

where $\mathbf{\Lambda}^*$ is the optimal dual parameter. We first note that since $\mathbf{X}$ is whitened such that $\mathbf{X}\mathbf{X}^T = \mathbf{I}_n$ and labels are one-hot encoded, the dual problem has a closed-form solution as follows

$$\boldsymbol{\lambda}_k^* = \begin{cases} \beta {t_j^*}^{2-L} \frac{\mathbf{y}_k}{\|\mathbf{y}_k\|_2} & \text{if } \beta \leq \|\mathbf{y}_k\|_2 \\ \mathbf{y}_k & \text{otherwise} \end{cases}, \quad \forall k \in [K]. \tag{56}$$

We now note that since $\mathbf{Y}$ has orthogonal one-hot encoded columns, the dual constraint can be decomposed into k maximization problems each of which can be maximized independently to find a set of extreme points. In particular, the j^{th} problem can be formulated as follows

$$\underset{\{\theta_l\}_{l=1}^{L-1} \in \Theta_{L-1}}{\operatorname{argmax}} \quad |\mathbf{y}_k^T(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j})_+| \leq \max\left\{\|\left(\mathbf{y}_k\right)_+\|_2, \|\left(-\mathbf{y}_k\right)_+\|_2\right\}.$$

Then, noting the whitened data assumption, the rest of steps directly follow Theorem 4.1 yielding the following weight matrices

$$\mathbf{W}_{l,j} = \frac{\boldsymbol{\phi}_{l-1,j}}{\|\boldsymbol{\phi}_{l-1,j}\|_2}\boldsymbol{\phi}_{l,j}^T, \ \forall l \in [L-2], \ \mathbf{w}_{L-1,j} = \frac{\boldsymbol{\phi}_{L-2,j}}{\|\boldsymbol{\phi}_{L-2,j}\|_2}, \tag{57}$$

where $\boldsymbol{\phi}_{0,j} = \mathbf{X}^T\mathbf{y}_k$ and $\{\boldsymbol{\phi}_{l,j}\}_{l=1}^{L-2}$ is a set of nonnegative vectors satisfying $\|\boldsymbol{\phi}_{l,j}\|_2 = t_j^*$, $\forall l$ and $\boldsymbol{\phi}_{l,i}^T\boldsymbol{\phi}_{l,j} = 0 \ \forall i \neq j$.

We now note that given the hidden layer weight in (57), the primal problem in (53) is convex and differentiable with respect to the output layer weight $\mathbf{W}_L$. Thus, we can find the optimal output layer weights by simply taking derivative and equating it to zero. Applying these steps yields the following output layer weights

$$\begin{aligned} \mathbf{W}_{L-1} &= \begin{bmatrix} \frac{\boldsymbol{\phi}_{L-2,1}}{\|\boldsymbol{\phi}_{L-2,1}\|_2} & \dots & \frac{\boldsymbol{\phi}_{L-2,K}}{\|\boldsymbol{\phi}_{L-2,K}\|_2} \end{bmatrix} = \sum_{r=1}^{K} \frac{\boldsymbol{\phi}_{L-2,r}}{\|\boldsymbol{\phi}_{L-2,r}\|_2}\boldsymbol{\phi}_{L-1,r}^T \\ \mathbf{W}_L &= \sum_{r=1}^{K}\left(\|\boldsymbol{\phi}_{0,r}\|_2 - \beta\right)_+ \boldsymbol{\phi}_{L-1,r}\mathbf{e}_r^T, \end{aligned} \tag{58}$$

where $\boldsymbol{\phi}_{L-1,r} = \mathbf{e}_r$ is the r^{th} ordinary basis vector.

Let us now assume that $t_j^* = 1$ for notational simplicity and then show that strong duality holds, i.e., $P^* = D^*$. We first denote the set of indices that yield an extreme point as $\mathcal{E} = \{j : \beta \leq \|\mathbf{y}_j\|_2, j \in [o]\}$. Then we compute the objective values for the dual problem (54) using (56)

$$\begin{aligned} D &= -\frac{1}{2}\|\mathbf{\Lambda}^* - \mathbf{Y}\|_F^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2 \\ &= -\frac{1}{2}\sum_{j\in\mathcal{E}}(\beta - \|\mathbf{y}_j\|_2)^2 + \frac{1}{2}\sum_{j=1}^{o}\|\mathbf{y}_j\|_2^2 \\ &= -\frac{1}{2}\beta^2|\mathcal{E}| + \beta\sum_{j\in\mathcal{E}}\|\mathbf{y}_j\|_2 + \frac{1}{2}\sum_{j\notin\mathcal{E}}\|\mathbf{y}_j\|_2^2. \end{aligned} \tag{59}$$

We next compute the objective value for the primal problem (53) (after applying the rescaling in Lemma A.4) using the weights in (57) and (58) as follows

$$\begin{aligned} P &= \frac{1}{2}\|f_{\theta,L}(\mathbf{X}) - \mathbf{Y}\|_F^2 + \frac{\beta}{2}\sum_{j=1}^{m}\|\mathbf{w}_{L,j}\|_2 \\ &= \frac{1}{2}\left\|\sum_{j\in\mathcal{E}}\left(\|\mathbf{y}_j\|_2 - \beta\right)\frac{\mathbf{y}_j}{\|\mathbf{y}_j\|_2}\mathbf{e}_j^T - \mathbf{Y}\right\|_F^2 + \beta\sum_{j\in\mathcal{E}}(\|\mathbf{y}_j\|_2 - \beta) \\ &= \frac{1}{2}\sum_{j\in\mathcal{E}}\left\|\beta\frac{\mathbf{y}_j}{\|\mathbf{y}_j\|_2}\mathbf{e}_j^T\right\|_F^2 + \frac{1}{2}\sum_{j\notin\mathcal{E}}\|\mathbf{y}_j\mathbf{e}_j^T\|_F^2 + \beta\sum_{j\in\mathcal{E}}\|\mathbf{y}_j\|_2 - \beta^2|\mathcal{E}| \\ &= -\frac{1}{2}\beta^2|\mathcal{E}| + \frac{1}{2}\sum_{j\notin\mathcal{E}}\|\mathbf{y}_j\|_2^2 + \beta\sum_{j\in\mathcal{E}}\|\mathbf{y}_j\|_2, \end{aligned}$$

which has the same value with (59). Therefore, strong duality holds, i.e., $P^* = D^*$, and the set of weights proposed in (57) and (58) is optimal. □

Theorem 4.4. *Suppose $\mathbf{Y}$ is one hot encoded and the network is overparameterized such that the range of $\mathbf{A}_{L-2,j}$ is $\mathbb{R}^n$, then an optimal solution to the following problem*[10]

$$\min_{\theta\in\Theta} \frac{1}{2}\left\|\sum_{j=1}^{m}\left(\mathrm{BN}_{\gamma,\alpha}\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)\right)_+ {\mathbf{w}_{L,j}}^T - \mathbf{Y}\right\|_F^2$$
$$+\frac{\beta}{2}\sum_{j=1}^{m}\left({\gamma_j^{(L-1)}}^2 + {\alpha_j^{(L-1)}}^2 + \|\mathbf{w}_{L,j}\|_2^2\right),$$

can be formulated in closed-form as follows

$$\left(\mathbf{w}_{L-1,j}^*, \mathbf{w}_{L,j}^*\right) = \left(\mathbf{A}_{L-2,j}^\dagger \mathbf{y}_j, \left(\|\mathbf{y}_j\|_2 - \beta\right)_+ \mathbf{e}_j\right)$$
$$\begin{bmatrix}\gamma_j^{(L-1)^*}\\ \alpha_j^{(L-1)^*}\end{bmatrix} = \frac{1}{\|\mathbf{y}_j\|_2}\begin{bmatrix}\|\mathbf{y}_j - \frac{1}{n}\mathbf{1}_{n\times n}\mathbf{y}_j\|_2\\ \frac{1}{\sqrt{n}}\mathbf{1}_n^T\mathbf{y}_j\end{bmatrix}$$

$\forall j \in [K]$, *where $\mathbf{e}_j$ is the j^{th} ordinary basis vector.*

Proof of Theorem 4.4. We first state the primal problem after applying the scaling between $\mathbf{w}_L$ and $(\boldsymbol{\gamma}^{(L-1)}, \boldsymbol{\alpha}^{(L-1)})$ as in Lemma A.4

$$P^* = \min_{\theta\in\Theta_s} \frac{1}{2}\left\|\sum_{j=1}^{m}\left(\mathrm{BN}_{\gamma,\alpha}\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)\right)_+ {\mathbf{w}_{L,j}}^T - \mathbf{Y}\right\|_F^2 + \beta\sum_{j=1}^{m}\|\mathbf{w}_{L,j}\|_2, \tag{60}$$

where $\Theta_s = \{\theta\in\Theta : {\gamma_j^{(L-1)}}^2 + {\alpha_j^{(L-1)}}^2 = 1, \forall j\in[m]\}$ and the corresponding dual is

$$P^* \geq D^* = \max_{\mathbf{\Lambda}} -\frac{1}{2}\|\mathbf{\Lambda} - \mathbf{Y}\|_F^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2 \text{ s.t. } \max_{\theta\in\Theta_s}\left|\mathbf{\Lambda}^T\left(\mathrm{BN}_{\gamma,\alpha}\left(\mathbf{A}_{L-2,j}\mathbf{w}_{L-1,j}\right)\right)_+\right| \leq \beta. \tag{61}$$

We now show that the following set of solutions for the primal and dual problem achieves strong duality, i.e., $P^* = D^*$, therefore, optimal.

$$\left(\mathbf{w}_{L-1,j}^*, \mathbf{w}_{L,j}^*\right) = \begin{cases}\left(\mathbf{A}_{L-2,j}^\dagger \mathbf{y}_j, \left(\|\mathbf{y}_j\|_2 - \beta\right)\mathbf{e}_j\right) & \text{if } \beta \leq \|\mathbf{y}_j\|_2\\ - & \text{otherwise}\end{cases}$$
$$\begin{bmatrix}\gamma_j^{(L-1)^*}\\ \alpha_j^{(L-1)^*}\end{bmatrix} = \frac{1}{\|\mathbf{y}_j\|_2}\begin{bmatrix}\|\mathbf{y}_j - \frac{1}{n}\mathbf{1}_{n\times n}\mathbf{y}_j\|_2\\ \frac{1}{\sqrt{n}}\mathbf{1}_n^T\mathbf{y}_j\end{bmatrix}, \quad \forall j\in[K].$$
$$\boldsymbol{\lambda}_j^* = \begin{cases}\beta\frac{\mathbf{y}_j}{\|\mathbf{y}_j\|_2} & \text{if } \beta\leq\|\mathbf{y}_j\|_2\\ \mathbf{y}_j & \text{otherwise}\end{cases}$$

Now let us first denote the set of indices that achieves the extreme point of the dual constraint as $\mathcal{E} = \{j : \beta \leq \|\mathbf{y}_j\|_2, j\in[K]\}$. Then the dual objective in (61) using the optimal dual parameter above

$$\begin{aligned}D_L^* &= -\frac{1}{2}\|\mathbf{\Lambda}^* - \mathbf{Y}\|_F^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2\\ &= -\frac{1}{2}\sum_{j\in\mathcal{E}}(\beta - \|\mathbf{y}_j\|_2)^2 + \frac{1}{2}\sum_{j=1}^{K}\|\mathbf{y}_j\|_2^2\\ &= -\frac{1}{2}\beta^2|\mathcal{E}| + \beta\sum_{j\in\mathcal{E}}\|\mathbf{y}_j\|_2 + \frac{1}{2}\sum_{j\notin\mathcal{E}}\|\mathbf{y}_j\|_2^2.\end{aligned} \tag{62}$$

[10] Notice here we only regularize the last layer's parameters, however, regularizing all the parameters does not change the analysis and conclusion as proven in Appendix A.4.

We next restate the scaled primal problem

$$\begin{aligned}
P_L^* &= \frac{1}{2}\left\|\sum_{j=1}^{K}\left(\frac{(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n})\mathbf{A}_{L-2,j}\mathbf{w}^*_{L-1,j}}{\|(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n})\mathbf{A}_{L-2,j}\mathbf{w}^*_{L-1,j}\|_2}\gamma_j^{(1)^*} + \frac{\mathbf{1}}{\sqrt{n}}\alpha_j^{(1)^*}\right)_+ {\mathbf{w}^*_{L,j}}^T - \mathbf{Y}\right\|_F^2 + \beta\sum_{j=1}^{K}\left\|\mathbf{w}^*_{L,j}\right\|_2 \\
&= \frac{1}{2}\left\|\sum_{j\in\mathcal{E}}\left(\|\mathbf{y}_j\|_2 - \beta\right)\frac{\mathbf{y}_j}{\|\mathbf{y}_j\|_2}\mathbf{e}_j^T - \mathbf{Y}\right\|_F^2 + \beta\sum_{j\in\mathcal{E}}(\|\mathbf{y}_j\|_2 - \beta) \\
&= \frac{1}{2}\sum_{j\in\mathcal{E}}\left\|\beta\frac{\mathbf{y}_j}{\|\mathbf{y}_j\|_2}\mathbf{e}_j^T\right\|_F^2 + \frac{1}{2}\sum_{j\notin\mathcal{E}}\|\mathbf{y}_j\mathbf{e}_j^T\|_F^2 + \beta\sum_{j\in\mathcal{E}}\|\mathbf{y}_j\|_2 - \beta^2|\mathcal{E}| \\
&= -\frac{1}{2}\beta^2|\mathcal{E}| + \frac{1}{2}\sum_{j\notin\mathcal{E}}\|\mathbf{y}_j\|_2^2 + \beta\sum_{j\in\mathcal{E}}\|\mathbf{y}_j\|_2, \qquad (63)
\end{aligned}$$

which is the same with (62). Therefore, strong duality holds, i.e., $P^* = D^*$, and the proposed set of weights is optimal for the primal problem (60).

□

Corollary 4.3. *Computing the last hidden layer activations after BN, i.e., $\mathbf{A}_{L-1} \in \mathbb{R}^{n\times K}$, using the optimal layer weight in Theorem 4.4 and then subtracting their global mean as in (Papyan et al., 2020) yields*

$$\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)\mathbf{A}_{L-1} = \sqrt{\frac{K}{n}}\left(\mathbf{I}_K \otimes \mathbf{1}_{\frac{n}{K}} - \frac{1}{K}\mathbf{1}_{n\times K}\right),$$

where we assume that samples are ordered, i.e., the first n/K samples belong to class 1, next n/K samples belong to class 2 and so on. Therefore, all the activations for a certain class k are the same and their mean is given by $(\sqrt{K/n})(\mathbf{e}_k - \mathbf{1}_K/K)$, which is the k^{th} column of a general simplex ETF with $\alpha = \sqrt{(K-1)/n}$ and $\mathbf{U} = \mathbf{I}_K$.

Proof of Corollary 4.3 . We first restate a crucial assumptions in (Papyan et al., 2020).

Assumption 1. The training dataset has balanced class distribution. Therefore, if we denote the number of data samples as n, then we have $\frac{n}{K}$ samples for each class $j \in [K]$.

Due to Assumption 1 and one-hot encoded labels, we have $\|\mathbf{y}_1\|_2 = \|\mathbf{y}_2\|_2 = \ldots = \|\mathbf{y}_K\|_2 = \sqrt{\frac{n}{K}}$. Now, we assume that $\sqrt{\frac{n}{K}} > \beta$ since otherwise none of the neurons will be optimal as proven in Theorem 4.4. We also remark that $\sqrt{\frac{n}{K}} > 1 \gg \beta$ in practice so that this assumption is trivially satisfied for practical scenarios considered in (Papyan et al., 2020). Therefore, the weights in Theorem 4.4 imply that

$$\begin{aligned}
\mathbf{A}_{(L-1),j} &= \left(\frac{(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n})\mathbf{A}_{L-2,j}\mathbf{w}^*_{(L-1),j}}{\left\|(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n})\mathbf{A}_{L-2,j}\mathbf{w}^*_{(L-1),j}\right\|_2}\gamma^{(L-1)^*} + \frac{\mathbf{1}_n}{\sqrt{n}}\alpha^{(L-1)^*}\right)_+ \\
&= \left(\frac{(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n})\mathbf{y}_j}{\left\|\mathbf{y}_j\right\|_2} + \frac{\mathbf{1}_{n\times n}\mathbf{y}_j}{n\left\|\mathbf{y}_j\right\|_2}\right)_+ \\
&= \frac{\mathbf{y}_j}{\left\|\mathbf{y}_j\right\|_2} \\
&= \frac{\sqrt{K}\mathbf{y}_j}{\sqrt{n}},
\end{aligned}$$

where $\mathbf{A}_{(L-1),j}$ denotes the j^{th} column of the last hidden layer activations after BN and the last equality follows from

Assumptions 1. We then subtract mean from $\mathbf{A}_{L-1}$ as follows

$$\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)\mathbf{A}_{L-1} = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)\frac{\sqrt{K}}{\sqrt{n}}\mathbf{Y} = \frac{\sqrt{K}}{\sqrt{n}}\begin{bmatrix} 1-\frac{1}{K} & -\frac{1}{K} & -\frac{1}{K} & \dots & -\frac{1}{K} \\ 1-\frac{1}{K} & -\frac{1}{K} & -\frac{1}{K} & \dots & -\frac{1}{K} \\ \vdots & & \dots & & \\ -\frac{1}{K} & 1-\frac{1}{K} & -\frac{1}{K} & \dots & -\frac{1}{K} \\ -\frac{1}{K} & 1-\frac{1}{K} & -\frac{1}{K} & \dots & -\frac{1}{K} \\ \vdots & & \dots & & \end{bmatrix}$$

$$= \frac{\sqrt{K}}{\sqrt{n}}\left(\mathbf{I}_K \otimes \mathbf{1}_{\frac{n}{K}} - \frac{1}{K}\mathbf{1}_{n\times K}\right),$$

where we assume that samples are ordered, i.e., the first n/K samples belong to class 1, next n/K samples belong to class 2 and so on. Therefore, all the activations for a certain class k are the same and their mean is given by

$$\frac{\sqrt{K}}{\sqrt{n}}\left[-\frac{1}{K}\ \dots\ \underbrace{1-\frac{1}{K}}_{k^{th}\text{ entry}}\ \dots\ -\frac{1}{K}\right] = \frac{\sqrt{K}}{\sqrt{n}}\left(\mathbf{e}_k^T - \frac{1}{K}\mathbf{1}_K^T\right),$$

which is the k^{th} column of a general simplex ETF with $\alpha = \sqrt{(K-1)/n}$ and $\mathbf{U} = \mathbf{I}_K$ in Definition 2. Hence, our analysis in Theory 4.4 completely explains why the patterns claimed in (Papyan et al., 2020) emerge throughout the training of the state-of-the-art architectures. We also remark that even though we use squared loss for the derivations, this analysis directly applies to the other convex loss functions including cross entropy and hinge loss as proven in Appendix A.1.

□