Supplementary Material

Table of Contents

A. Supplementary Material

A.1. Additional numerical results

In this section, we provide detailed information about our experiments.

We first note that for small scale experiments, i.e., Figure 4 and Table 1, we use CVX (Grant & Boyd, 2014) and CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018) with the SDPT3 solver (Tütüncü et al., 2001) to solve convex optimization problems in (12). Moreover, the training is performed on the CPU of a laptop with i7 processor and 16GB of RAM. For UCI experiments, we use the $80\% - 20\%$ splitting ratio for the training and test sets. Moreover, the learning rate of SGD is tuned via a grid-search on the training split. Specifically, we try different values and choose the best performing learning rate on the validation datasets, which turns out to be 0.3.

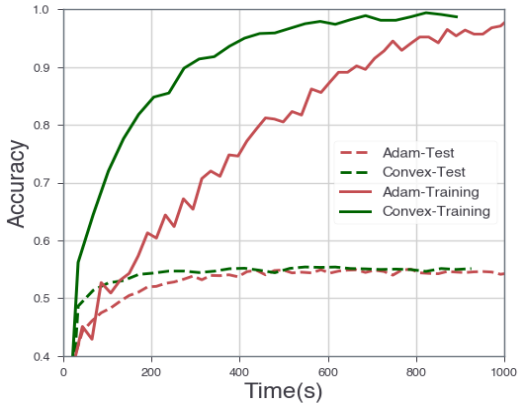
For larger scale experiment in Figure 5, we use a GPU with 50GB of memory. In order to train the constrained convex program in (12), we now introduce an unconstrained version of the convex program as follows

$$\min_{\mathbf{w},\mathbf{w}' \in \mathbb{R}^{2dm_1 P_1 P_2}} \frac{1}{2} \left\| \tilde{\mathbf{X}} \left(\mathbf{w}' - \mathbf{w} \right) - \mathbf{y} \right\|_2^2 + \beta \left(\|\mathbf{w}\|_{2,1} + \|\mathbf{w}'\|_{2,1} \right) + \rho \left(g_{\mathcal{C}}(\mathbf{w}) + g_{\mathcal{C}}(\mathbf{w}') \right) \tag{15}$$

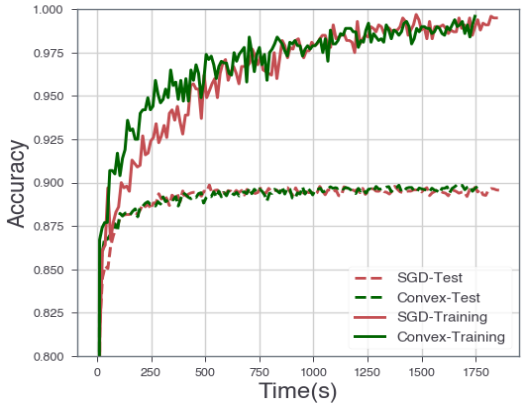
where $\rho > 0$ is a trade-off parameter and

$$g_{\mathcal{C}}(\mathbf{w}) := \mathbf{1}^T \sum_{i,j,l} \left(\left(-(2\mathbf{D}_{1ij} - \mathbf{I}_n) \mathbf{X} \mathbf{w}_{ijl}^{+} \right)_+ + \left(2\mathbf{D}_{1ij} - \mathbf{I}_n) \mathbf{X} \mathbf{w}_{ijl}^{-} \right)_+ \right) + \mathbf{1}^T \sum_{i,l,\pm} \left(-(2\mathbf{D}_{2l} - \mathbf{I}_n) \sum_{j=1}^{m_1} \mathbf{D}_{1ij} \mathbf{X} \mathbf{w}_{ijl}^{\pm} \right)_+ .$$

Since the problem in (15) is in an unconstrained form, we can directly optimize its parameters using conventional local search algorithms such as SGD and Adam. Hence, we are able to use PyTorch to optimize the non-convex objective in (4) and the convex objective in (15) on the conventional benchmark datasets such as CIFAR-10 and Fashion-MNIST datasets with their original training and test splits. For the learning rates of SGD and Adam optimizer (applied to the non-convex formulations), we again follow the same grid-search technique and select 1 and 0.01 as the learning rates, respectively. For SGD, we also use momentum with a parameter of 0.9. Moreover, for the convex programs in both cases, we select the number of hyperplane arrangements for the first and second layer such that $P_1 P_2 = K$ and set the learning rate and the trade-off parameter as 10^{-6} and $\rho = 0.01$, respectively. Then, we run the algorithms on the non-convex and convex formulations for $200, 100$ epochs using SGD in Figure 5a. Similarly, we run the algorithms on the non-convex and convex formulations for $150, 150$ epochs using Adam in Figure 5b. We also note that for all of these experiments, we use an approximated form of the convex program detailed in Remark 3. Therefore, we conjecture that one can even further improve the performance by either sampling more hyperplane arrangements or developing a technique to characterize the set of hyperplane arrangements that generalize well.



(a) CIFAR-10



(b) Fashion-MNIST

Figure 6: Accuracy of a three-layer architecture trained using the non-convex formulation (4) and the proposed convex program (12), where we use (a) CIFAR-10 with $(n, d, m_1, K, \beta, \text{batch size}) = (5\text{x}10^4, 3072, 100, 40, 10^{-3}, 10^3)$ and (b) Fashion-MNIST with $(n, d, m_1, K, \beta, \text{batch size}) = (6\text{x}10^4, 784, 100, 40, 10^{-3}, 10^3)$. We note that the convex model is trained using (a) Adam and (b) SGD.

To complement the experiments in Figure 5, we also conduct a new experiment, where we use Adam and SGD for CIFAR-10 and Fashion-MNIST, respectively. For this case, we use the same setup above except that the learning rates are chosen as $(5\text{x}10^{-7}, 5\text{x}10^{-4})$ and $(10^{-5}, 3)$ for CIFAR-10 and Fashion-MNIST respectively, where the former learning rates belong to the convex problems. We also run the algorithms on the non-convex and convex formulations for $66, 26$ epochs using Adam in Figure 6a and $150, 150$ epochs using SGD in Figure 6b. We plot the accuracy values in Figure 6, where the training on the convex formulation achieves faster convergence and higher (or at least the same) accuracies compared to the training on the original non-convex formulation.

A.2. Proof of Lemma 1

We first note that similar proofs are also presented in (Neyshabur et al., 2014; Savarese et al., 2019; Ergen & Pilanci, 2019; 2020a;b;c;d).

For any $\theta \in \Theta$, we can rescale the parameters as $\bar{\mathbf{w}}_{(L-1)k} = \alpha_k \mathbf{w}_{(L-1)k}$ and $\bar{w}_{Lk} = w_{Lk}/\alpha_k$, for any $\alpha_k > 0$. Then, the network output becomes

$$f_{\bar{\theta},k}(\mathbf{X}) = \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \bar{\mathbf{w}}_{(L-1)k}\right)_+ \bar{w}_{Lk} = \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k}\right)_+ w_{Lk},$$

which proves $f_{\theta,k}(\mathbf{X}) = f_{\bar{\theta},k}(\mathbf{X}), \forall k \in [K]$. In addition to this, we have the following basic inequality

$$\frac{1}{2}\sum_{k=1}^{K}(w_{Lk}^2 + \|\mathbf{w}_{(L-1)k}\|_2^2) \geq \sum_{k=1}^{K}(|w_{Lk}| \, \|\mathbf{w}_{(L-1)k}\|_2),$$

where the equality is achieved with the scaling choice $\alpha_k = \left(\frac{|w_{Lk}|}{\|\mathbf{w}_{(L-1)k}\|_2}\right)^{\frac{1}{2}}$ is used. Since the scaling operation does not change the right-hand side of the inequality, we can set $\|\mathbf{w}_{(L-1)k}\|_2 = 1, \forall k \in [K]$. 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 $\|\mathbf{w}_{(L-1)k}\|_2 \leq 1$. Let us also assume that for a certain index k, we obtain $\|\mathbf{w}_{(L-1)k}\|_2 < 1$ with $w_{Lk} \neq 0$ as an optimal solution. This shows that the unit norm inequality constraint is not active for $\mathbf{w}_{(L-1)k}$, and hence removing the constraint for $\mathbf{w}_{(L-1)k}$ will not change the optimal solution. However, when we remove the constraint, $\|\mathbf{w}_{(L-1)k}\|_2 \to \infty$ reduces the objective value since it yields $w_{Lk} = 0$. Therefore, we have a contradiction, which proves that all the constraints that correspond to a nonzero w_{Lk} must be active for an optimal solution. This also shows that replacing $\|\mathbf{w}_{(L-1)k}\|_2 = 1$ with $\|\mathbf{w}_{(L-1)k}\|_2 \leq 1$ does not change the solution to the problem.

A.3. Constraint on the layer weights (Remark 2)

Here, we prove that changing the unit norm constraint on the first $L-2$ layer weights does not change the structure of the regularization induced by the primal problem (2).

We first note that due to the AM-GM inequality, for each sub-network k, we have

$$\sum_{l=1}^{L} \|\mathbf{W}_{lk}\|_F^2 \geq L \prod_{l=1}^{L} \|\mathbf{W}_{lk}\|_F^{2/L}$$

where the equality is achieved when all the layer weight have the same Frobenius norm, i.e., $\|\mathbf{W}_{1k}\|_F = \ldots = \|\mathbf{W}_{Lk}\|_F$. Therefore, for a given set of arbitrary weight matrices, one can scale them such that their Frobenius norms are equal to each other and reduce the objective function in (16). Based on this observation, for the rest of the derivations, we assume $\|\mathbf{W}_{1k}\|_F = \ldots = \|\mathbf{W}_{(L-2)k}\|_F = s$ without loss of generality.

Now, let us consider the following primal problem instead of (2)

$$P^* = \min_{\theta \in \Theta} \mathcal{L}\left(\sum_{k=1}^{K} f_{\theta,k}(\mathbf{X}), \mathbf{y}\right) + \frac{\beta}{2} \sum_{k=1}^{K} \sum_{l=L-1}^{L} \|\mathbf{W}_{lk}\|_F^2, \tag{16}$$

where $\Theta := \{\theta : \|\mathbf{W}_{lk}\|_F \leq s, \forall l \in [L-2], \forall k \in [K]\}$. For this problem, we can follow all the derivations in the proof of Theorem 1 (see Appendix A.4 below) by changing $\mathbf{1}^T\mathbf{t} = 1$ as $\mathbf{1}^T\mathbf{t} = s^2$ in (17). Then, we will have an additional s factor in the last step of (18). This change will yield β/s instead β in (18). Therefore, if we define a new variable as $\beta' = \beta/s$, following the remaining steps in the proof below will yield the same convex program in (23) with the regularization parameter β'. Hence, the impact of the norm constraint in the primal problem (16) can be reverted by simply setting a new regularization parameter $\beta' = \beta/s$.

A.4. Proof of Theorem 1

We start with rewriting (11) as follows

$$\max_{\mathcal{I}_j \in \{\pm 1\}} \max_{\substack{i \in [P_1] \\ l \in [P_2]}} \max_{\substack{t_j \geq 0 \\ \mathbf{1}^T\mathbf{t} \leq 1}} \max_{\substack{\|\mathbf{w}'_{1j}\|_2^2 / w'_{2j} \leq t_j \\ \mathbf{1}^T\mathbf{w}'_2 \leq 1 \\ \mathbf{w}'_2 \geq 0}} \mathbf{v}^T \mathbf{D}_{2l} \sum_{j=1}^{m_1} \mathcal{I}_j \mathbf{D}_{1ij} \mathbf{X} \mathbf{w}'_{1j} \tag{17}$$

$$\text{s.t. } (2\mathbf{D}_{1ij} - \mathbf{I}_n)\mathbf{X}\mathbf{w}'_{1j} \geq 0,\ \forall j \in [m_1],\ (2\mathbf{D}_{2l} - \mathbf{I}_n) \sum_{j=1}^{m_1} \mathcal{I}_j \mathbf{D}_{1ij} \mathbf{X} \mathbf{w}'_{1j} \geq 0.$$

Then, the Lagrange function of (17) is as follows

$$L(\mathbf{W}'_1, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \mathbf{v}^T \mathbf{D}_{2l} \sum_{j=1}^{m_1} \mathcal{I}_j \mathbf{D}_{1ij} \mathbf{X} \mathbf{w}'_{1j} + \boldsymbol{\alpha}_2^T (2\mathbf{D}_{2l} - \mathbf{I}_n) \sum_{j=1}^{m_1} \mathcal{I}_j \mathbf{D}_{1ij} \mathbf{X} \mathbf{w}'_{1j} + \sum_{j=1}^{m_1} \boldsymbol{\alpha}_{1j}^T (2\mathbf{D}_{1ij} - \mathbf{I}_n) \mathbf{X} \mathbf{w}'_{1j}$$

where $\boldsymbol{\alpha}_2 \geq 0$, $\boldsymbol{\alpha}_{1j} \geq 0$, $\forall j \in [m_1]$. Thus, we have

$$\max_{\substack{i\in[P_1]\\ l\in[P_2]\\ \mathcal{I}_j\in\{\pm1\}}} \max_{\substack{t_j\geq 0\\ \mathbf{1}^T\mathbf{t}\leq 1}} \max_{\substack{\|\mathbf{w}'_{1j}\|_2^2/w'_{2j}\leq t_j\\ \mathbf{1}^T\mathbf{w}'_2\leq 1\\ \mathbf{w}'_2\geq 0}} \min_{\substack{\boldsymbol{\alpha}_2\geq 0\\ \boldsymbol{\alpha}_{1j}\geq 0}} \mathbf{v}^T\mathbf{D}_{2l}\sum_{j=1}^{m_1}\mathcal{I}_j\mathbf{D}_{1ij}\mathbf{X}\mathbf{w}'_{1j} + \boldsymbol{\alpha}_2^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\sum_{j=1}^{m_1}\mathcal{I}_j\mathbf{D}_{1ij}\mathbf{X}\mathbf{w}'_{1j} + \sum_{j=1}^{m_1}\boldsymbol{\alpha}_{1j}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{w}'_{1j}$$

$$= \max_{\substack{i\in[P_1]\\ l\in[P_2]\\ \mathcal{I}_j\in\{\pm1\}}} \min_{\substack{\boldsymbol{\alpha}_2\geq 0\\ \boldsymbol{\alpha}_{1j}\geq 0}} \max_{\substack{t_j\geq 0\\ \mathbf{1}^T\mathbf{t}\leq 1}} \max_{\substack{\|\mathbf{w}'_{1j}\|_2^2/w'_{2j}\leq t_j\\ \mathbf{1}^T\mathbf{w}'_2\leq 1\\ \mathbf{w}'_2\geq 0}} \sum_{j=1}^{m_1}\left(\mathcal{I}_j\mathbf{v}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{X}\mathbf{w}'_{1j} + \mathcal{I}_j\boldsymbol{\alpha}_2^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\mathbf{X}\mathbf{w}'_{1j} + \boldsymbol{\alpha}_{1j}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{w}'_{1j}\right)$$

$$= \max_{\substack{i\in[P_1]\\ l\in[P_2]\\ \mathcal{I}_j\in\{\pm1\}}} \min_{\substack{\boldsymbol{\alpha}_2\geq 0\\ \boldsymbol{\alpha}_{1j}\geq 0}} \max_{\substack{t_j\geq 0\\ \mathbf{1}^T\mathbf{t}\leq 1}} \max_{\substack{\mathbf{1}^T\mathbf{w}'_2\leq 1\\ \mathbf{w}'_2\geq 0}} \sum_{j=1}^{m_1}\left\|\mathcal{I}_j\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v} + \mathcal{I}_j\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_2 + \mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1j}\right\|_2\sqrt{w'_{2j}t_j}$$

$$= \max_{\substack{i\in[P_1]\\ l\in[P_2]\\ \mathcal{I}_j\in\{\pm1\}}} \min_{\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_{1j}\geq 0} \max_{\substack{t_j\geq 0\\ \mathbf{1}^T\mathbf{t}\leq 1}} \left(\sum_{j=1}^{m_1}\left\|\mathcal{I}_j\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v} + \mathcal{I}_j\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_2 + \mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1j}\right\|_2^2 t_j\right)^{\frac{1}{2}}$$

$$= \max_{\substack{i\in[P_1]\\ l\in[P_2]\\ \mathcal{I}_{j_m}\in\{\pm1\}}} \min_{\substack{\boldsymbol{\alpha}_2\geq 0\\ \boldsymbol{\alpha}_{1j_m}\geq 0}} \left\|\mathcal{I}_{j_m}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij_m}\mathbf{v} + \mathcal{I}_{j_m}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij_m}\boldsymbol{\alpha}_2 + \mathbf{X}^T(2\mathbf{D}(S_{1j_m})-\mathbf{I}_n)\boldsymbol{\alpha}_{1j_m}\right\|_2, \tag{18}$$

where j_m denotes the index with the maximum norm. Note that we change the order of max-min for the first equality in (18) since the problem in (17) is convex and there exists a strictly feasible point, therefore strong duality holds, given fixed diagonal matrices $\{\mathbf{D}_{1ij}\}_{j=1}^{m_1}$, $\mathbf{D}_{2l}$ and a fixed set of signs $\{\mathcal{I}_j\}_{j=1}^{m_1}$.

We first enumerate all hyperplane arrangements and signs and index them in an arbitrary order, which are denoted as $\mathbf{D}_{1ij}$ and $\mathbf{D}_{2l}$, where $i \in [P_1]$, $l \in [P_2]$, $P_1 = |\mathcal{S}_1|$, and $P_2 = |\mathcal{S}_2|$. Then we have

$$(9) \iff \max_{\substack{i\in[P_1]\\ l\in[P_2]\\ \mathcal{I}_j\in\{\pm1\}}} \min_{\substack{\boldsymbol{\alpha}_2\geq 0\\ \boldsymbol{\alpha}_{1j}\geq 0}} \max_{j\in[m_1]} \left\|\mathcal{I}_j\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v} + \mathcal{I}_j\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_2 + \mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1j}\right\|_2 \leq \beta,$$

$$\iff \forall i \in [P_1], \forall j \in [m_1], \forall l \in [P_2], \forall \pm\ \exists \boldsymbol{\alpha}_{2il}^{\pm}, \boldsymbol{\alpha}_{1ijl}^{\pm} \geq 0$$

$$\text{s.t. } \left\|\mathcal{I}_{ijl}^{\pm}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v} + \mathcal{I}_{ijl}^{\pm}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_{2il}^{\pm} + \mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1ijl}^{\pm}\right\|_2 \leq \beta,$$

where we introduce the notation $\mathcal{I}_{ijl}^{\pm} = \pm 1$ to enumerate all possible sign patterns. Therefore, the dual problem in (6) can also be written as

$$\max_{\substack{\mathbf{v}\\ \boldsymbol{\alpha}_{2il}^{\pm},\boldsymbol{\alpha}_{1ijl}^{\pm}\geq 0\\ \boldsymbol{\alpha}_{il}^{\pm'},\boldsymbol{\beta}_{ijl}^{\pm'}\geq 0}} -\frac{1}{2}\|\mathbf{v}-\mathbf{y}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2^2 \tag{19}$$

$$\text{s.t. } \left\|\mathcal{I}_{ijl}^{\pm}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v} + \mathcal{I}_{ijl}^{\pm}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_{2il}^{\pm} + \mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1ijl}^{\pm}\right\|_2 \leq \beta,\ \forall i \in [P_1], \forall j \in [m_1], \forall l \in [P_2], \forall \pm$$

$$\left\|-\mathcal{I}_{ijl}^{\pm'}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v} + \mathcal{I}_{ijl}^{\pm'}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_{2il}^{\pm'} + \mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1ijl}^{\pm'}\right\|_2 \leq \beta,\ \forall i \in [P_1], \forall j \in [m_1], \forall l \in [P_2], \forall \pm .$$

We note that the above problem is convex and strictly feasible for $\mathbf{v} = \boldsymbol{\alpha}_{1ijl}^{\pm} = \boldsymbol{\alpha}_{1ijl}^{\pm'} = \boldsymbol{\alpha}_{2il}^{\pm} = \boldsymbol{\alpha}_{2il}^{\pm'} = \mathbf{0}$. Therefore,

Slater's conditions and consequently strong duality holds (Boyd & Vandenberghe, 2004), and (19) can be written as

$$\begin{aligned}
&\min_{\gamma_{ijl}^{\pm},\gamma_{ijl}^{\pm'}\geq 0} \max_{\substack{\mathbf{v} \\ \boldsymbol{\alpha}_{2il}^{\pm},\boldsymbol{\alpha}_{1ijl}^{\pm}\geq 0 \\ \boldsymbol{\alpha}_{2il}^{\pm'},\boldsymbol{\alpha}_{1ijl}^{\pm'}\geq 0}} -\frac{1}{2}\|\mathbf{v}-\mathbf{y}\|_2^2+\frac{1}{2}\|\mathbf{y}\|_2^2 \\
&+\sum_{+,-}\sum_{i=1}^{P_1}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\gamma_{ijl}^{\pm}\left(\beta-\left\|\mathcal{I}_{ijl}^{\pm}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v}+\mathcal{I}_{ijl}^{\pm}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_2 il^{\pm}+\mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1ijl}^{\pm}\right\|_2\right) \\
&+\sum_{+,-}\sum_{i=1}^{P_1}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\gamma_{ijl}^{\pm'}\left(\beta-\left\|-\mathcal{I}_{ijl}^{\pm'}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v}+\mathcal{I}_{ijl}^{\pm'}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_{2il}^{\pm'}+\mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1ijl}^{\pm'}\right\|_2\right),
\end{aligned} \tag{20}$$

where we change the order of max-min since strong duality holds. Next, we introduce variables $\mathbf{r}_{ijl}^{\pm},\mathbf{r}_{ijl}^{\pm'}\in\mathbb{R}^d,\forall i\in[P_1],\forall j\in[m_1],\forall l\in[P_2],\forall\pm$ and represent the dual problem (20) as

$$\begin{aligned}
&\min_{\gamma_{ijl}^{\pm},\gamma_{ijl}^{\pm'}\geq 0} \max_{\substack{\mathbf{v} \\ \boldsymbol{\alpha}_{2il}^{\pm},\boldsymbol{\alpha}_{1ijl}^{\pm}\geq 0 \\ \boldsymbol{\alpha}_{2il}^{\pm'},\boldsymbol{\alpha}_{1ijl}^{\pm'}\geq 0}} \min_{\mathbf{r}_{ijl}^{\pm},\mathbf{r}_{ijl}^{\pm'}\in\mathcal{B}_2} -\frac{1}{2}\|\mathbf{v}-\mathbf{y}\|_2^2+\|\mathbf{y}\|_2^2 \\
&+\sum_{+,-}\sum_{i=1}^{P_1}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\gamma_{ijl}^{\pm}\left(\beta+\mathbf{r}_{ijl}^{\pm T}\left(\mathcal{I}_{ijl}^{\pm}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v}+\mathcal{I}_{ijl}^{\pm}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_{2il}^{\pm}+\mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1ijl}^{\pm}\right)\right) \\
&+\sum_{+,-}\sum_{i=1}^{P_1}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\gamma_{ijl}^{\pm'}\left(\beta+\mathbf{r}_{ijl}^{\pm' T}\left(-\mathcal{I}_{ijl}^{\pm'}\mathbf{X}^T\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{v}+\mathcal{I}_{ijl}^{\pm'}\mathbf{X}^T(2\mathbf{D}_{2l}-\mathbf{I}_n)\mathbf{D}_{1ij}\boldsymbol{\alpha}_{2il}^{\pm'}+\mathbf{X}^T(2\mathbf{D}_{1ij}-\mathbf{I}_n)\boldsymbol{\alpha}_{1ijl}^{\pm'}\right)\right).
\end{aligned} \tag{21}$$

Now, we can change the order of max-min due to Sion's minimax theorem (Sion, 1958) and then compute the maximums with respect to $\mathbf{v},\boldsymbol{\alpha}_{2il}^{\pm},\boldsymbol{\alpha}_{1ijl}^{\pm},\boldsymbol{\alpha}_{2il}^{\pm'},\boldsymbol{\alpha}_{1ijl}^{\pm'}$

$$\min_{\gamma_{ijl}^{\pm},\gamma_{ijl}^{\pm'}\geq 0}\min_{\mathbf{r}_{ijl}^{\pm},\mathbf{r}_{ijl}^{\pm'}\in\mathcal{B}_2}\frac{1}{2}\left\|\sum_{+,-}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\sum_{i=1}^{P_1}\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{X}\left(\mathcal{I}_{ijl}^{\pm'}\gamma_{ijl}^{\pm'}\mathbf{r}_{ijl}^{\pm'}-\mathcal{I}_{ijl}^{\pm}\gamma_{ijl}^{\pm}\mathbf{r}_{ijl}^{\pm}\right)-\mathbf{y}\right\|_2^2+\beta\sum_{+,-}\sum_{i=1}^{P_1}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\left(\gamma_{ijl}^{\pm}+\gamma_{ijl}^{\pm'}\right) \tag{22}$$

$$\text{s.t. } (2\mathbf{D}_{2l}-\mathbf{I}_n)\sum_{j=1}^{m_1}\mathcal{I}_{ijl}^{\pm}\mathbf{D}_{1ij}\mathbf{X}\mathbf{r}_{ijl}^{\pm}\geq 0,\ (2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{r}_{ijl}^{\pm}\geq 0,\ \forall i\in[P_1],\forall j\in[m_1],\forall l\in[P_2],\forall\pm$$

$$(2\mathbf{D}_{2l}-\mathbf{I}_n)\sum_{j=1}^{m_1}\mathcal{I}_{ijl}^{\pm'}\mathbf{D}_{1ij}\mathbf{X}\mathbf{r}_{ijl}^{\pm'}\geq 0,\ (2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{r}_{ijl}^{\pm'}\geq 0,\ \forall i\in[P_1],\forall j\in[m_1],\forall l\in[P_2],\forall\pm.$$

Next, we apply a change of variables as $\mathbf{w}_{ijl}^{\pm}=\mathcal{I}_{ijl}^{\pm}\gamma_{ijl}^{\pm}\mathbf{r}_{ijl}^{\pm}$ and $\mathbf{w}_{ijl}^{\pm'}=\mathcal{I}_{ijl}^{\pm'}\gamma_{ijl}^{\pm'}\mathbf{r}_{ijl}^{\pm'}$, which yields

$$\min_{\mathbf{w}_{ijl}^{\pm},\mathbf{w}_{ijl}^{\pm'}}\frac{1}{2}\left\|\sum_{+,-}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\sum_{i=1}^{P_1}\mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{X}\left(\mathbf{w}_{ijl}^{\pm'}-\mathbf{w}_{ijl}^{\pm}\right)-\mathbf{y}\right\|_2^2+\beta\sum_{+,-}\sum_{i=1}^{P_1}\sum_{j=1}^{m_1}\sum_{l=1}^{P_2}\left(\|\mathbf{w}_{ijl}^{\pm}\|_2+\|\mathbf{w}_{ijl}^{\pm'}\|_2\right) \tag{23}$$

$$\text{s.t. } (2\mathbf{D}_{2l}-\mathbf{I}_n)\sum_{j=1}^{m_1}\mathbf{D}_{1ij}\mathbf{X}\mathbf{w}_{ijl}^{\pm}\geq 0,\ (2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{w}_{ijl}^{+}\geq 0,\ (2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{w}_{ijl}^{-}\leq 0,\ \forall i\in[P_1],\forall j\in[m_1],\forall l\in[P_2],\forall\pm$$

$$(2\mathbf{D}_{2l}-\mathbf{I}_n)\sum_{j=1}^{m_1}\mathbf{D}_{1ij}\mathbf{X}\mathbf{w}_{ijl}^{\pm'}\geq 0,\ (2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{w}_{ijl}^{+'}\geq 0,\ (2\mathbf{D}_{1ij}-\mathbf{I}_n)\mathbf{X}\mathbf{w}_{ijl}^{-'}\leq 0,\ \forall i\in[P_1],\forall j\in[m_1],\forall l\in[P_2],\forall\pm,$$

which is a finite dimensional convex problem with $4dm_1P_1P_2$ variables and $4n(m_1+1)P_1P_2$ constraints.

■

A.5. Proof of Proposition 1

We can construct an optimal solution to the primal problem in (4) from the optimal solution to the convex program in (12), i.e., denoted as $\{\mathbf{w}_{ijl}^{\pm^*}, \mathbf{w}_{ijl}^{\pm'^*}\}_{i,j,l,\pm}$, as follows

$$\mathbf{W}_{1k}^* = \begin{cases} \frac{1}{\sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{+*}\|_2}} \left[\frac{\mathbf{w}_{i1l}^{+*}}{\sqrt{\|\mathbf{w}_{i1l}^{+*}\|_2}} \cdots \frac{\mathbf{w}_{im_1l}^{+*}}{\sqrt{\|\mathbf{w}_{im_1l}^{+*}\|_2}} \right] & \text{if } 1 \le k \le P_1P_2 \\ \frac{1}{\sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{-*}\|_2}} \left[\frac{-\mathbf{w}_{i1l}^{-*}}{\sqrt{\|\mathbf{w}_{i1l}^{-*}\|_2}} \cdots \frac{-\mathbf{w}_{im_1l}^{-*}}{\sqrt{\|\mathbf{w}_{im_1l}^{-*}\|_2}} \right] & \text{if } P_1P_2+1 \le k \le 2P_1P_2 \\ \frac{1}{\sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{+'*}\|_2}} \left[\frac{\mathbf{w}_{i1l}^{+'*}}{\sqrt{\|\mathbf{w}_{i1l}^{+'*}\|_2}} \cdots \frac{\mathbf{w}_{im_1l}^{+'*}}{\sqrt{\|\mathbf{w}_{im_1l}^{+'*}\|_2}} \right] & \text{if } 2P_1P_2+1 \le k \le 3P_1P_2 \\ \frac{1}{\sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{-'*}\|_2}} \left[\frac{-\mathbf{w}_{i1l}^{-'*}}{\sqrt{\|\mathbf{w}_{i1l}^{-'*}\|_2}} \cdots \frac{-\mathbf{w}_{im_1l}^{-'*}}{\sqrt{\|\mathbf{w}_{im_1l}^{-'*}\|_2}} \right] & \text{if } 3P_1P_2+1 \le k \le 4P_1P_2 \end{cases}$$

$$\mathbf{w}_{2k}^* = \begin{cases} \left[\sqrt{\|\mathbf{w}_{i1l}^{+*}\|_2} \cdots \sqrt{\|\mathbf{w}_{im_1l}^{+*}\|_2} \right]^T & \text{if } 1 \le k \le P_1P_2 \\ \left[-\sqrt{\|\mathbf{w}_{i1l}^{-*}\|_2} \cdots -\sqrt{\|\mathbf{w}_{im_1l}^{-*}\|_2} \right]^T & \text{if } P_1P_2+1 \le k \le 2P_1P_2 \\ \left[\sqrt{\|\mathbf{w}_{i1l}^{+'*}\|_2} \cdots \sqrt{\|\mathbf{w}_{im_1l}^{+'*}\|_2} \right]^T & \text{if } 2P_1P_2+1 \le k \le 3P_1P_2 \\ \left[-\sqrt{\|\mathbf{w}_{i1l}^{-'*}\|_2} \cdots -\sqrt{\|\mathbf{w}_{im_1l}^{-'*}\|_2} \right]^T & \text{if } 3P_1P_2+1 \le k \le 4P_1P_2 \end{cases}$$

$$w_{3k}^* = \begin{cases} -\sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{+*}\|_2} & \text{if } 1 \le k \le P_1P_2 \\ -\sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{-*}\|_2} & \text{if } P_1P_2+1 \le k \le 2P_1P_2 \\ \sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{+'*}\|_2} & \text{if } 2P_1P_2+1 \le k \le 3P_1P_2 \\ \sqrt{\sum_{j=1}^{m_1} \|\mathbf{w}_{ijl}^{-'*}\|_2} & \text{if } 3P_1P_2+1 \le k \le 4P_1P_2 \end{cases},$$

where

$$(l,i) = \begin{cases} \left(\left\lfloor \frac{k-1}{P_1} \right\rfloor + 1, k-(l-1)P_1 \right) & \text{if } 1 \le k \le P_1P_2 \\ \left(\left\lfloor \frac{k-1-P_1P_2}{P_1} \right\rfloor + 1, k-1-P_1P_2-(l-1)P_1 \right) & \text{if } P_1P_2+1 \le k \le 2P_1P_2 \\ \left(\left\lfloor \frac{k-1-2P_1P_2}{P_1} \right\rfloor + 1, k-2P_1P_2-(l-1)P_1 \right) & \text{if } 2P_1P_2+1 \le k \le 3P_1P_2 \\ \left(\left\lfloor \frac{k-1-3P_1P_2}{P_1} \right\rfloor + 1, k-3P_1P_2-(l-1)P_1 \right) & \text{if } 3P_1P_2+1 \le k \le 4P_1P_2 \end{cases}.$$

Therefore, we obtain an optimal solution to (4) as $\{\mathbf{W}_{1k}^*, \mathbf{w}_{2k}^*, w_{3k}^*\}_{k=1}^{4P_1P_2}$, where $\mathbf{w}_{1kj}^*$ and w_{2kj}^* are the columns and entries of $\mathbf{W}_{1k}^* \in \mathbb{R}^{d \times m_1}$ and $\mathbf{w}_{2k}^* \in \mathbb{R}^{m_1}$, respectively. The optimality of these parameters can be verified as follows.

We first note that this set of parameters yields the same output with the convex program in (12), i.e.,

$$\sum_{k=1}^{4P_1P_2} \left((\mathbf{X}\mathbf{W}_{1k}^*)_+ \mathbf{w}_{2k}^* \right)_+ w_{3k}^* = \sum_{+,-} \sum_{j=1}^{m_1} \sum_{l=1}^{P_2} \sum_{i=1}^{P_1} \mathbf{D}_{2l}\mathbf{D}_{1ij}\mathbf{X} \left(\mathbf{w}_{ijl}^{\pm'^*} - \mathbf{w}_{ijl}^{\pm^*} \right).$$

We also remark that these parameters are feasible for the original problem (4), i.e., $\|\mathbf{W}_{1k}^*\|_F^2 = 1$, $\forall k \in [4P_1P_2]$, and achieve the same regularization cost with (12)

$$\frac{\beta}{2} \sum_{k=1}^{4P_1P_2} \left(\|\mathbf{w}_{2k}^*\|_2^2 + w_{3k}^{*^2} \right) = \beta \sum_{+,-} \sum_{i=1}^{P_1} \sum_{j=1}^{m_1} \sum_{l=1}^{P_2} \left(\|\mathbf{w}_{ijl}^{\pm^*}\|_2 + \|\mathbf{w}_{ijl}^{\pm'^*}\|_2 \right)$$

Since $\{\mathbf{W}_{1k}^*, \mathbf{w}_{2k}^*, w_{3k}^*\}_{k=1}^{4P_1P_2}$ has the same output, therefore the same prediction error, and regularization cost with the optimal parameters of the convex program in (12), this set of parameters also achieves the optimal objective value P^*, i.e.,

$$P^* = \frac{1}{2} \left\| \sum_{k=1}^{4P_1P_2} \left((\mathbf{X}\mathbf{W}_{1k}^*)_+ \mathbf{w}_{2k}^* \right)_+ w_{3k}^* - \mathbf{y} \right\|_2^2 + \frac{\beta}{2} \sum_{k=1}^{4P_1P_2} \left(\|\mathbf{w}_{2k}^*\|_2^2 + w_{3k}^{*^2} \right).$$

■

A.6. Proof for the dual problem in (3)

The proof follows from classical Fenchel duality (Boyd & Vandenberghe, 2004). We first restate the primal problem after applying the rescaling in Lemma 1

$$P^* = \min_{\hat{\mathbf{y}} \in \mathbb{R}^n, \theta \in \Theta_p} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) + \beta \|\mathbf{w}_L\|_1 \text{ s.t. } \hat{\mathbf{y}} = \sum_{k=1}^{K} \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k} \right)_+ w_{Lk}. \tag{24}$$

Now, we first form the Lagrangian as

$$L(\mathbf{v}, \hat{\mathbf{y}}, \mathbf{w}_L) = \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) - \mathbf{v}^T \hat{\mathbf{y}} + \mathbf{v}^T \sum_{k=1}^{K} \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k} \right)_+ w_{Lk} + \beta \|\mathbf{w}_L\|_1$$

and then formulate the dual function as

$$\begin{aligned} g(\mathbf{v}) &= \min_{\hat{\mathbf{y}}, \mathbf{w}_L} L(\mathbf{v}, \hat{\mathbf{y}}, \mathbf{w}_L) \\ &= \min_{\hat{\mathbf{y}}, \mathbf{w}_L} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) - \mathbf{v}^T \hat{\mathbf{y}} + \mathbf{v}^T \sum_{k=1}^{K} \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k} \right)_+ w_{Lk} + \beta \|\mathbf{w}_L\|_1 \\ &= -\mathcal{L}^*(\mathbf{v}) \text{ s.t. } \left| \mathbf{v}^T \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k} \right)_+ \right| \leq \beta, \, \forall k \in [K], \end{aligned}$$

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

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

Therefore, the dual of (24) with respect to $\mathbf{w}_L$ and $\hat{\mathbf{y}}$ can be written as

$$P^* = \min_{\theta \in \Theta_p} \max_{\mathbf{v}} g(\mathbf{v}) = \min_{\theta \in \Theta_p} \max_{\mathbf{v}} -\mathcal{L}^*(\mathbf{v}) \text{ s.t. } \left| \mathbf{v}^T \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k} \right)_+ \right| \leq \beta, \, \forall k \in [K].$$

We now change the order of min-max to obtain the following lower bound

$$\begin{aligned} P^* \geq D^* &= \max_{\mathbf{v}} \min_{\theta \in \Theta_p} -\mathcal{L}^*(\mathbf{v}) \text{ s.t. } \left| \mathbf{v}^T \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k} \right)_+ \right| \leq \beta, \, \forall k \in [K] \\ &= \max_{\mathbf{v}} -\mathcal{L}^*(\mathbf{v}) \text{ s.t. } \max_{\theta \in \Theta_p} \left| \mathbf{v}^T \left((\mathbf{X}\mathbf{W}_1)_+ \dots \mathbf{w}_{(L-1)} \right)_+ \right| \leq \beta. \end{aligned}$$

■

A.7. Extension to vector outputs

Here, we present the extensions of our approach to vector outputs, i.e., $\mathbf{Y} \in \mathbb{R}^{n \times C}$. The original training problem in this case is as follows

$$P_v^* := \min_{\theta \in \Theta} \mathcal{L} \left(\sum_{k=1}^{K} f_{\theta,k}(\mathbf{X}), \mathbf{Y} \right) + \frac{\beta}{2} \sum_{k=1}^{K} \sum_{l=L-1}^{L} \|\mathbf{W}_{lk}\|_F^2.$$

Using the same scaling in Lemma 1 and following the steps in the scalar output case yields the following dual problem

$$D_v^* := \max_{\mathbf{V}} \min_{\theta \in \Theta_p} -\mathcal{L}^*(\mathbf{V}) \text{s.t. } \left\| \mathbf{V}^T \left((\mathbf{X}\mathbf{W}_{1k})_+ \dots \mathbf{w}_{(L-1)k} \right)_+ \right\|_2 \leq \beta, \, \forall k \in [K],$$

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

$$\mathcal{L}^*(\mathbf{V}) := \max_{\mathbf{Z}} \operatorname{trace} \left(\mathbf{Z}^T \mathbf{V} \right) - \mathcal{L}(\mathbf{Z}, \mathbf{Y}) \,.$$

The rest of the derivations directly follows the steps in Section A.4 and (Sahiner et al., 2021).