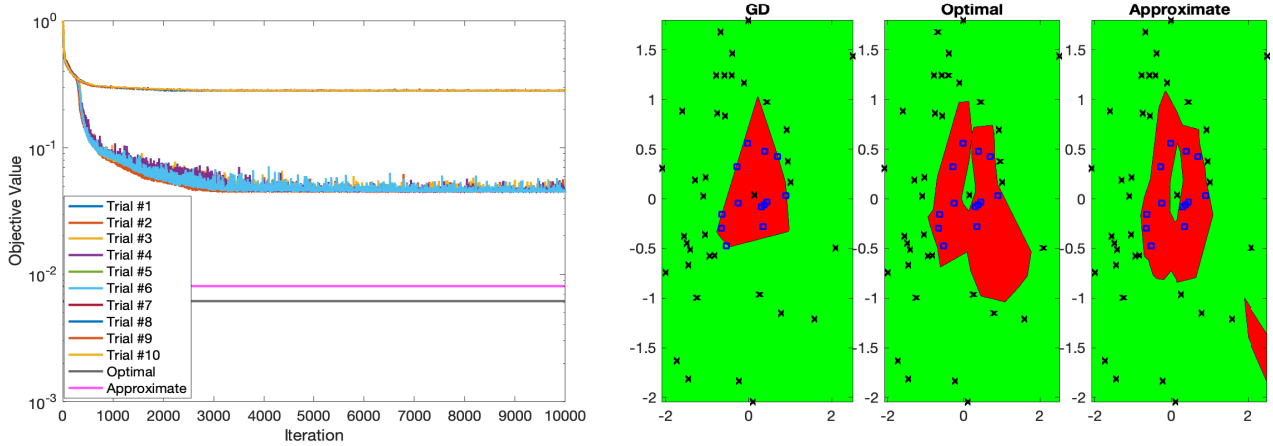
Appendix

Table of Contents

A. Appendix

A.1. Additional numerical results



(a) Independent SGD initialization trials with $m = 50$

(b) Decision boundaries

Figure 6: Training of a two-layer ReLU network with SGD (10 initialization trials) and proposed convex programs on a two-dimensional dataset. Optimal and Approximate denote the objective value obtained by the proposed convex program (8) and its approximation, respectively. Learned decision boundaries are also depicted.

We now present another numerical experiment on a two-dimensional dataset[4], where we place a negative sample ($y = -1$) near the positive samples ($y = +1$) to have a more challenging loss landscape. In Figure 6, we observe that all the SGD

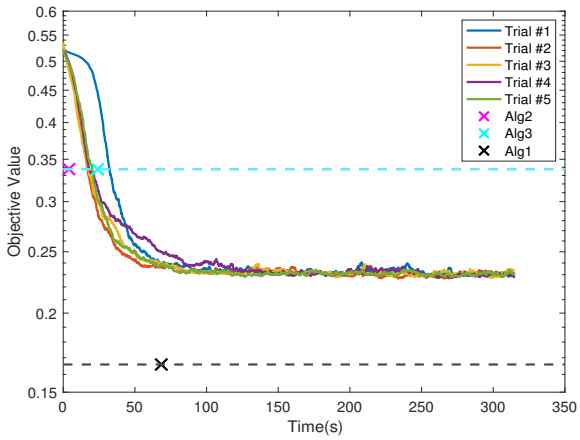
[4]In all the experiments, we use CVX (Grant & Boyd, 2014) and CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018) with the SDPT3 solver (Tütüncü et al., 2001) to solve convex optimization problems.

realizations are stuck at local minima, therefore, achieve a significantly higher objective value compared to both Optimal and Approximate, which are based on convex optimization.

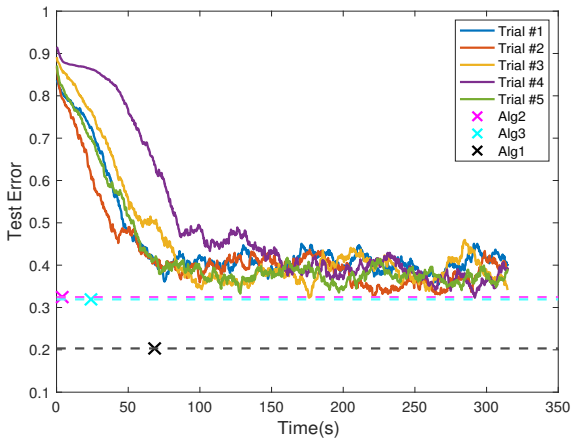
In addition to the classification datasets, we evaluate the performance of the algorithms on three regression datasets, i.e., the Boston Housing, Kinematics, and Bank datasets (Torgo). In Figure 7, we plot the objective value and the corresponding test error of 5 independent initialization trials with respect to time in seconds, where we use squared loss and choose $n = 400$, $d = 13$, $m = 12$, and batch size(bs) 25. Similarly, we plot the objective values and test errors for the Kinematics and Bank datasets in Figure 8 and 9, where $(n, d, m, \text{bs}) = (4000, 8, 12, 25)$ and $(n, d, m, \text{bs}) = (4000, 32, 12, 25)$, respectively. We observe that Alg1 achieves the lowest objective value and test error in both cases.

We also consider the training of a two-layer CNN architecture. In Figure 10, we provide the binary classification performance of the algorithms on a subset of CIFAR-10, where we use hinge loss and choose $(n, d, m, \text{bs}) = (195, 3072, 50, 20)$, filter size $4 \times 4 \times 3$, and stride 4. This experiment also illustrates that Alg1 achieves lower objective value and higher test accuracy compared with the other methods including GD. We also emphasize that in this experiment, we use sign patterns of a clustered subset of patches, specifically 50 clusters, as well as the GD patterns for Alg1. As depicted in Figure 11, the neurons that correspond to the sign patterns of GD matches with the neurons found by GD. Thus, the performance difference stems from the additional sign patterns found by clustering the patches.

In order to evaluate the computational complexity of the introduced approaches, in Table 1, we provide the training time of each algorithm in the main paper. This data clearly shows that the introduced convex programs outperform GD while requiring significantly less training time.



(a) Training objective value



(b) Test error

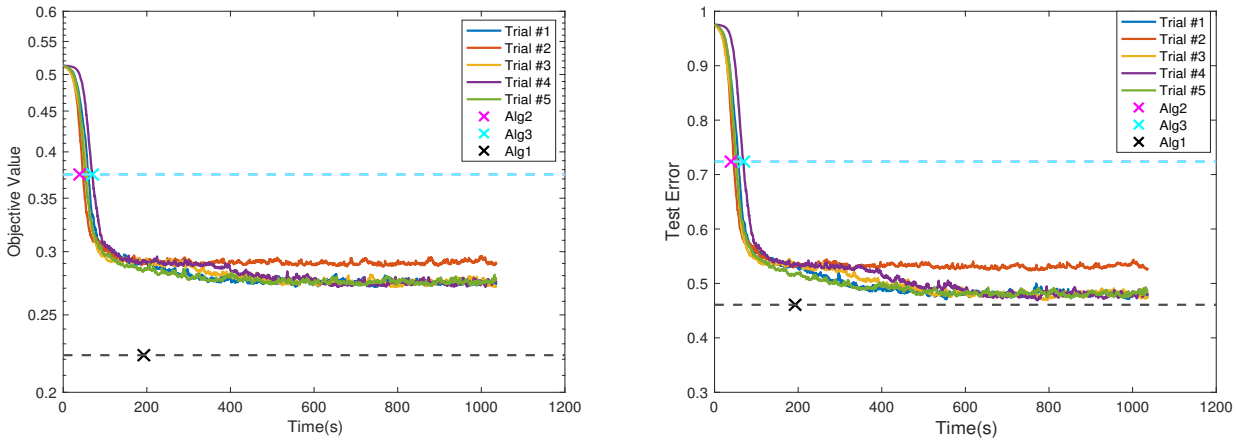
Figure 7: Training and test errors of the algorithms on the Boston Housing dataset ($n = 400$ and $d = 13$) where we run SGD independently in 5 initialization trials. For the convex program (8) approximations (Alg1, Alg2 and Alg3), crossed markers correspond to the total computation time of the convex optimization solver.

Table 1: Training time(in seconds), final objective value and test accuracy(%) of each algorithm in the main paper,where we use the CVX SDPT3 solver to optimize the convex programs.

	Figure 2		Figure 3			Figure 4				Figure 5	
	SGD	Optimal	GD	Approx.	Optimal	GD	Alg1	Alg2	Alg3	GD	L1-Convex
Time(s)	420.663	1.225	890.339	1.498	117.858	624.787	108.065	5.931	12.009	65.365	1.404
Train. Objective	0.001	0.001	0.0032	0.0028	0.0026	0.0042	0.0022	0.0032	0.0032	0.804	0.803
Test Accuracy(%)	-	-	-	-	-	62.75	66.80	60.15	60.20	-	-

A.2. Constructing hyperplane arrangements in polynomial time

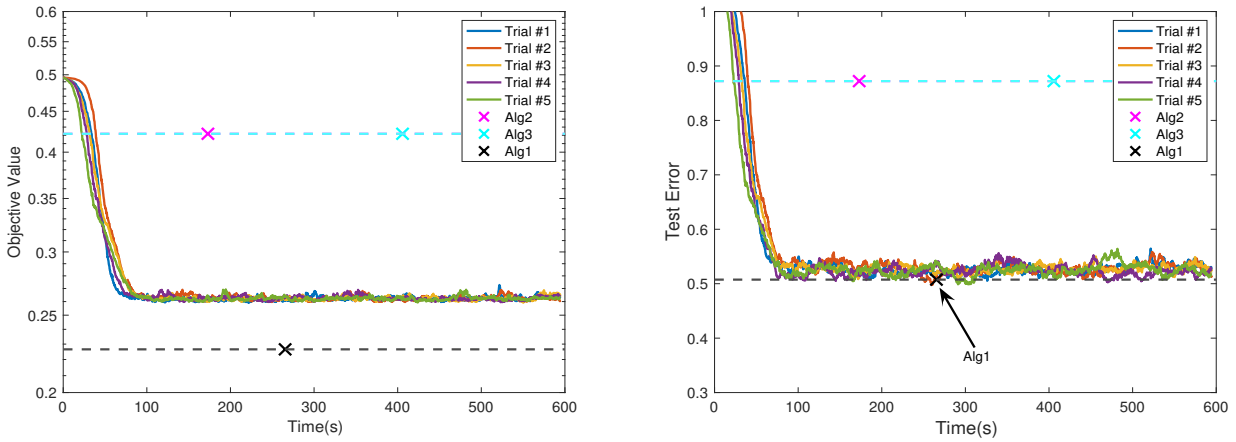
We now consider the number of all distinct sign patterns $\text{sign}(Xz)$ for all possible choices $z \in \mathbb{R}^d$. Note that this number is the number of regions in a partition of $\mathbb{R}^d$ by hyperplanes passing through the origin, and are perpendicular to the rows



(a) Training objective value

(b) Test error

Figure 8: Performance comparison of the algorithms on the Kinematics dataset ($n = 4000$ and $d = 8$) where we run SGD independently in 5 initialization trials. For the convex program (8) approximations (Alg1, Alg2 and Alg3), crossed markers correspond to the total computation time of the convex optimization solver.



(a) Training objective value

(b) Test error

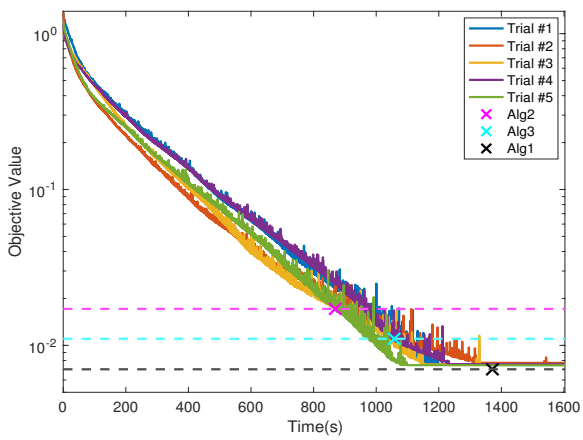
Figure 9: Performance comparison of the algorithms on the Bank dataset ($n = 4000$ and $d = 32$) where we run SGD independently in 5 initialization trials. For the convex program (8) approximations (Alg1, Alg2 and Alg3), crossed markers correspond to the total computation time of the convex optimization solver.

of X. We now show that the dimension d can be replaced with $\text{rank}(X)$ without loss of generality. Suppose that the data matrix X has rank r. We may express $X = U\Sigma V^T$ using its Singular Value Decomposition in compact form, where $U \in \mathbb{R}^{n\times r}, \Sigma \in \mathbb{R}^{r\times r}, V^T \in \mathbb{R}^{r\times d}$. For any vector $z \in \mathbb{R}^d$ we have $Xz = U\Sigma V^T z = Uz'$ for some $z' \in \mathbb{R}^r$. Therefore, the number of distinct sign patterns $\text{sign}(Xz)$ for all possible $z \in \mathbb{R}^d$ is equal to the number of distinct sign patterns $\text{sign}(Uz')$ for all possible $z' \in \mathbb{R}^r$.

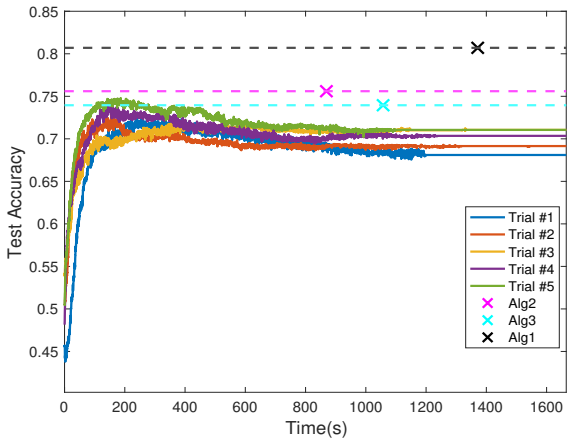
Consider an arrangement of n hyperplanes $\in \mathbb{R}^r$, where $n \geq r$. Let us denote the number of regions in this arrangement by $P_{n,r}$. In (Ojha, 2000; Cover, 1965) it's shown that this number satisfies

$$P_{n,r} \leq 2\sum_{k=0}^{r-1} \binom{n-1}{k}.$$

For hyperplanes in general position, the above inequality is in fact an equality. In (Edelsbrunner et al., 1986), the authors present an algorithm that enumerates all possible hyperplane arrangements $O(n^r)$ time, which can be used to construct the

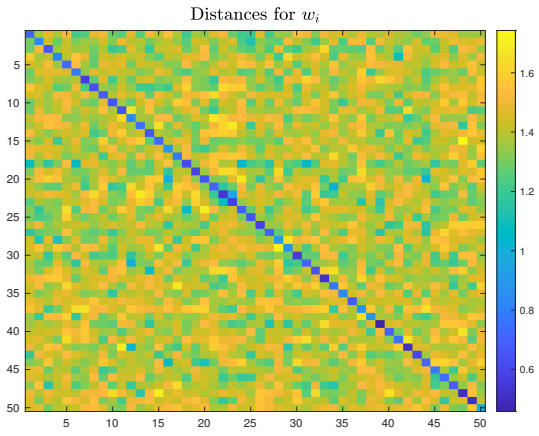


(a) Training objective value

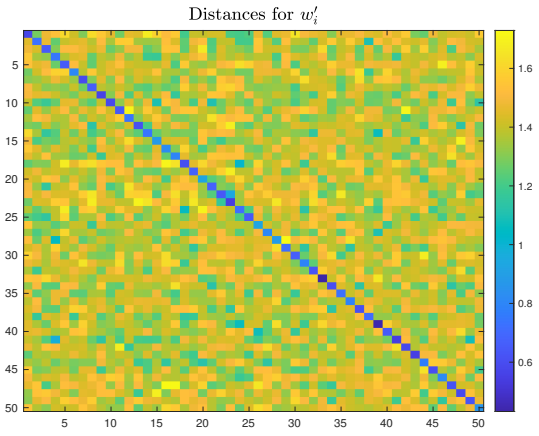


(b) Test accuracy

Figure 10: Performance of the algorithms for two-layer CNN training on a subset of CIFAR-10 ($n = 195$ and filter size $4 \times 4 \times 3$) where we run SGD independently in 5 initialization trials. For the convex program (8) approximations (Alg1, Alg2 and Alg3), crossed markers correspond to the total computation time of the convex optimization solver.



(a)



(b)

Figure 11: Visualization of the distance (using the Euclidean norm of the difference) between the neurons found by GD and our convex program in Figure 10. The ij^{th} entries of the distance plots are $\left\| \frac{w_i}{\|w_i\|_2} - \frac{u_j}{\|u_j\|_2} \right\|_2$ and $\left\| \frac{w_i'}{\|w_i'\|_2} - \frac{u_j}{\|u_j\|_2} \right\|_2$, respectively.

data for the convex program (8).

A.3. Equivalence of the ℓ_1 penalized neural network training cost

In this section, we prove the equivalence between (2) and (3).

Lemma 2 ((Neyshabur et al., 2014; Savarese et al., 2019; Ergen & Pilanci, 2020b;c;d))**.** The following two problems are equivalent:

$$\min_{\{u_j,\alpha_j\}_{j=1}^m} \frac{1}{2} \left\| \sum_{j=1}^m (Xu_j)_+ \alpha_j - y \right\|_2^2 + \frac{\beta}{2} \sum_{j=1}^m (\|u_j\|_2^2 + \alpha_j^2) = \min_{\|u_j\|_2 \leq 1} \min_{\{\alpha_j\}_{j=1}^m} \frac{1}{2} \left\| \sum_{j=1}^m (Xu_j)_+ \alpha_j - y \right\|_2^2 + \beta \sum_{j=1}^m |\alpha_j| \,.$$

[**Proof of Lemma 2**] We can rescale the parameters as $\bar{u}_j = \gamma_j u_j$ and $\bar{\alpha}_j = \alpha_j / \gamma_j$, for any $\gamma_j > 0$. Then, the output

becomes

$$\sum_{j=1}^{m}(X\bar{u}_j)_+\bar{\alpha}_j = \sum_{j=1}^{m}(Xu_j\gamma_j)_+\frac{\alpha_j}{\gamma_j} = \sum_{j=1}^{m}(Xu_j)_+\alpha_j,$$

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

$$\frac{1}{2}\sum_{j=1}^{m}(\alpha_j^2 + \|u_j\|_2^2) \geq \sum_{j=1}^{m}(|\alpha_j|\,\|u_j\|_2),$$

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

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

A.4. Dual problem for (3)

The following lemma proves the dual form of (3).

Lemma 3. The dual form of the following primal problem

$$\min_{\|u_j\|_2\leq 1}\min_{\{\alpha_j\}_{j=1}^m}\frac{1}{2}\Big\|\sum_{j=1}^{m}(Xu_j)_+\alpha_j - y\Big\|_2^2 + \beta\sum_{j=1}^{m}|\alpha_j|\,,$$

is given by the following

$$\min_{\|u_j\|_2\leq 1}\max_{\substack{v\in\mathbb{R}^n \text{ s.t.} \\ |v^T(Xu_j)_+|\leq\beta}} -\frac{1}{2}\|y-v\|_2^2 + \frac{1}{2}\|y\|_2^2\,.$$

[**Proof of Lemma 3**] Let us first reparametrize the primal problem as follows

$$\min_{\|u_j\|_2\leq 1}\min_{r,\{\alpha_j\}_{j=1}^m}\frac{1}{2}\|r\|_2^2 + \beta\sum_{j=1}^{m}|\alpha_j| \text{ s.t. } r = \sum_{j=1}^{m}(Xu_j)_+\alpha_j - y,$$

which has the following Lagrangian

$$L(v,r,u_j,\alpha_j) = \frac{1}{2}\|r\|_2^2 + \beta\sum_{j=1}^{m}|\alpha_j| + v^Tr + v^Ty - v^T\sum_{j=1}^{m}(Xu_j)_+\alpha_j.$$

Then, minimizing over r and α yields the proposed dual form.

A.5. Dual problem for (11)

Let us first reparameterize the primal problem as follows

$$\max_{M,v} -\frac{1}{2}\|v-y\|_2^2 + \frac{1}{2}\|y\|_2^2 \text{ s.t. } \sigma_{\max}(M) \leq \beta,\ M = [X_1^Tv\,...\,X_K^Tv].$$

Then the Lagrangian is as follows

$$L(\lambda,Z,M,v) = -\frac{1}{2}\|v-y\|_2^2 + \frac{1}{2}\|y\|_2^2 + \lambda\left(\beta - \sigma_{\max}(M)\right) + \text{trace}(Z^TM) - \text{trace}(Z^T[X_1^Tv\,...\,X_K^Tv])$$

$$= -\frac{1}{2}\|v-y\|_2^2 + \frac{1}{2}\|y\|_2^2 + \lambda\left(\beta - \sigma_{\max}\left(M\right)\right) + \operatorname{trace}(Z^T M) - v^T \sum_{k=1}^{K} X_k z_k,$$

where $\lambda \geq 0$. Then maximizing over M and v yields the following dual form

$$\min_{z_k \in \mathbb{R}^d, \forall k \in [K]} \frac{1}{2}\Big\| \sum_{k=1}^{K} X_k z_k - y \Big\|_2^2 + \beta \Big\| [z_1, ..., z_K] \Big\|_*,$$

where $\Big\| [z_1, ..., z_K] \Big\|_* = \|Z\|_* = \sum_i \sigma_i(Z)$ is the ℓ_1 norm of singular values, i.e., nuclear norm (Recht et al., 2010).

A.6. Dual problem for (13)

Let us denote the eigenvalue decomposition of U_j as $U_j = F D_j F^H$, where $F \in \mathbb{C}^{d \times d}$ is the Discrete Fourier Transform matrix and $D_j \in \mathbb{C}^{d \times d}$ is a diagonal matrix. Then, applying the scaling in Lemma 2 and then taking the dual as in Lemma 3 yields

$$\max_v -\frac{1}{2}\|v-y\|_2^2 + \frac{1}{2}\|y\|_2^2 \text{ s.t. } \|v^T X F D F^H\|_2 \leq \beta,\ \forall D : \|D\|_F^2 \leq d,$$

which can be equivalently written as

$$\max_v -\frac{1}{2}\|v-y\|_2^2 + \frac{1}{2}\|y\|_2^2 \text{ s.t. } \|v^T \tilde{X} D\|_2 \leq \beta,\ \forall D : \|D\|_F^2 \leq d.$$

Since D is diagonal, $\|D\|_F^2 \leq d$ is equivalent to $\sum_{i=1}^{d} D_{ii}^2 \leq 1$. Therefore, the problem above can be further simplified as

$$\max_v -\frac{1}{2}\|v-y\|_2^2 + \frac{1}{2}\|y\|_2^2 \text{ s.t. } \|v^T \tilde{X}\|_\infty \leq \frac{\beta}{\sqrt{d}}\ .$$

Then, taking the dual of this problem gives the following

$$\min_{z \in \mathbb{C}^d} \frac{1}{2}\Big\| \tilde{X} z - y \Big\|_2^2 + \frac{\beta}{\sqrt{d}}\|z\|_1.$$

A.7. Dual problem for vector output two-layer linear convolutional networks

Vector version of the two-layer linear convolutional network training problem has the following dual

$$\max_V \operatorname{trace} V^T Y \text{ s.t. } \max_{\|u\|_2 \leq 1} \sum_k \|V^T X_k u\|_2^2 \leq 1.$$

Similarly, extreme points are the maximal eigenvectors of $\sum_k X_k^T V V^T X_k$ When $V = Y$, and one-hot encoding is used, these are the right singular vectors of the matrix $[X_{1,c}^T X_{2,c}^T ... X_{K,c}^T]^T$ whose rows contain all the patch vectors for class c.

A.8. Semi-infinite strong duality

Note that the semi-infinite problem (4) is convex. We first show that the optimal value is finite. For $\beta > 0$, it is clear that $v = 0$ is strictly feasible, and achieves 0 objective value. Note that the optimal value p^* satisfies $p^* \leq \|y\|_2^2$ since this value is achieved when all the parameters are zero. Consequently, Theorem 2.2 of (Shapiro, 2009) implies that strong duality holds, i.e., $p^* = d_\infty^*$, if the solution set of the semi-infinite problem in (4) is nonempty and bounded. Next, we note that the solution set of (4) is the Euclidean projection of y onto the polar set $(\mathcal{Q}_X \cup -\mathcal{Q}_X)^\circ$ which is a convex, closed and bounded set since $\big(Xu\big)_+$ can be expressed as the union of finitely many convex closed and bounded sets. □

A.9. Semi-infinite strong gauge duality

Now we prove strong duality for (7). We invoke the semi-infinite optimality conditions for the dual (7), in particular we apply Theorem 7.2 of (Goberna & López-Cerdá, 1998) and use the standard notation therein. We first define the set

$$\mathbf{K} = \mathbf{cone}\left\{ \begin{pmatrix} s\left(Xu\right)_+ \\ 1 \end{pmatrix}, u \in \mathcal{B}_2, s \in \{-1, +1\}; \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\}.$$

Note that $\mathbf{K}$ is the union of finitely many convex closed sets, since $\big(Xu\big)_+$ 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 (15) 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 (15). □

A.10. Neural Gauge function and equivalence to minimum norm networks

Consider the gauge function

$$p^g = \min_{r\geq 0} r \text{ s.t. } ry \in \text{conv}(\mathcal{Q}_X \cup -\mathcal{Q}_X)$$

and its dual representation in terms of the support function of the polar of $\text{conv}(\mathcal{Q}_X \cup -\mathcal{Q}_X)$

$$d^g = \max_{v} v^T y \text{ s.t. } v \in (\mathcal{Q}_X \cup -\mathcal{Q}_X)^\circ.$$

Since the set $\mathcal{Q}_X \cup -\mathcal{Q}_X$ is a closed convex set that contains the origin, we have $p^g = d^g$ (Rockafellar, 1970) and $(\text{conv}(\mathcal{Q}_X \cup -\mathcal{Q}_X))^\circ = (\mathcal{Q}_X \cup -\mathcal{Q}_X)^\circ$. The result in Section A.8 implies that the above value is equal to the semi-infinite dual value, i.e., $p^d = p^g_\infty$, where

$$p^g_\infty := \min_{\mu} \|\mu\|_{TV} \text{ s.t. } \int_{u\in\mathcal{B}_2} (Xu)_+ d\mu(u) = y\,.$$

By Caratheodory's theorem, there exists optimal solutions the above problem consisting of m^* Dirac deltas (Rockafellar, 1970; Rosset et al., 2007), and therefore

$$p^g_\infty = \min_{u_j\in\mathcal{B}_2, j\in[m^*]} \sum_{j=1}^{m^*} |\alpha_j| \text{ s.t. } \sum_{j=1}^{m^*} (Xu_j)_+ d\alpha_j = y\,,$$

where we define m^* as the number of Dirac delta's in the optimal solution to p^g_∞. If the optimizer is non-unique, we define m^* as the minimum cardinality solution among the set of optimal solutions. Now consider the non-convex problem

$$\min_{\{u_j,\alpha_j\}_{j=1}^m} \|\alpha\|_1 \text{ s.t. } \sum_{j=1}^{m} (Xu_j)_+ \alpha_j = y,\ \|u_j\|_2 \leq 1\,.$$

Using the standard parameterization for ℓ_1 norm we get

$$\min_{\{u_j\}_{j=1}^m, s\geq 0, t\geq 0} \sum_{j=1}^{m} (t_j + s_j) \text{ s.t. } \sum_{j=1}^{m} (Xu_j)_+ t_j - (Xu_j)_+ s_j = y,\ \|u_j\|_2 \leq 1\,, \forall j.$$

Introducing a slack variable $r \in \mathbb{R}_+$, an equivalent representation can be written as

$$\min_{\{u_j\}_{j=1}^m, s\geq 0, t\geq 0, r\geq 0} r \text{ s.t. } \sum_{j=1}^{m} (Xu_j)_+ t_j - (Xu_j)_+ s_j = y,\ \sum_{j=1}^{m} (t_j + s_j) = r,\ \|u_j\|_2 \leq 1\,, \forall j.$$

Note that $r > 0$ as long as $y \neq 0$. Rescaling variables by letting $t'_j = t_j/r$, $s'_j = s_j/r$ in the above program, we obtain

$$\min_{\{u_j\}_{j=1}^m, s'\geq 0, t'\geq 0, r\geq 0} r \text{ s.t. } \sum_{j=1}^{m} \big((Xu_j)_+ t'_j - (Xu_j)_+ s'_j\big) = ry,\ \sum_{j=1}^{m} (t'_j + s'_j) = 1,\ \|u_j\|_2 \leq 1\,, \forall j\,.$$

Suppose that $m \geq m^*$. It holds that

$$\exists s', t' \geq 0\,, \{u_j\}_{j=1}^m \text{ s.t. } \sum_{j=1}^{m} (t'_j + s'_j) = 1,\ \|u_j\|_2 \leq 1,\ \forall j,\ \sum_{j=1}^{m} (Xu_j) t'_j - (Xu_j)_+ s'_j = ry \iff ry \in \text{conv}(\mathcal{Q}_X \cup -\mathcal{Q}_X). \tag{17}$$

We conclude that the optimal value of (17) is identical to the gauge function p_g.

A.11. Alternative proof of the semi-infinite strong duality

It holds that $p^* \geq d^*$ by weak duality in (4). Theorem 1 proves that the objective value of (15) is identical to the value of (2) as long as $m \geq m^*$. Therefore we have $p^* = d^*$. □

A.12. Finite dimensional strong duality results for Theorem 1

Lemma 4. Suppose $D(S)$, $D(S^c)$ are fixed diagonal matrices as described earlier, and X is a fixed matrices. The dual of the convex optimization problem

$$\max_{\substack{u \in \mathbb{R}^d \\ \|u\|_2 \leq 1 \\ D(S)Xu \geq 0 \\ D(S^c)Xu \leq 0}} v^T D(S) X u$$

is given by

$$\min_{\substack{\alpha, \beta \in \mathbb{R}^n \\ \alpha, \beta \geq 0}} \|X^T D(S)\big(v + \alpha + \beta\big) - X^T \beta\|_2$$

and strong duality holds.

Note that the linear inequality constraints specify valid hyperplane arrangements. Then there exists strictly feasible points in the constraints of the maximization problem. Standard finite second order cone programming duality implies that strong duality holds (Boyd & Vandenberghe, 2004b) and the dual is as specified. □

A.13. General loss functions

In this section, we extend our derivations to arbitrary convex loss functions.

Consider minimizing the sum of the squared loss objective and squared ℓ_2-norm of all parameters

$$p^* := \min_{\{\alpha_j, u_j\}_{j=1}^m} \ell\left(\sum_{j=1}^m (Xu_j)_+ \alpha_j, y\right) + \frac{\beta}{2} \sum_{j=1}^m (\|u_j\|_2^2 + \alpha_j^2)\,, \tag{18}$$

where $\ell(\cdot, y)$ is a convex loss function. Then, consider the following finite dimensional convex optimization problem

$$\min_{\{v_i, w_i\}_{i=1}^P} \ell\left(\sum_{i=1}^P D_i X (v_i - w_i), y\right) + \beta \sum_{i=1}^P (\|v_i\|_2 + \|w_i\|_2) \text{ s.t. } (2D_i - I) X v_i \geq 0,\ (2D_i - I) X w_i \geq 0. \forall i \in [P], \tag{19}$$

Let us define $m^* = \sum_{i: v_i^* \neq 0}^P 1 + \sum_{i: w_i^* \neq 0}^P 1$, where $\{v_i^*, w_i^*\}_{i=1}^P$ are optimal in (19).

Theorem 5. The convex program (19) and the non-convex problem (18) where $m \geq m^*$ has identical optimal values. Moreover, an optimal solution to (18) can be constructed from an optimal solution to (19) as follows (8) as follows

$$(u_{j_{1i}}^*, \alpha_{j_{1i}}^*) = \left(\frac{v_i^*}{\sqrt{\|v_i^*\|_2}}, \sqrt{\|v_i^*\|_2}\right) \qquad \text{if} \quad v_i^* \neq 0$$
$$(u_{j_{2i}}^*, \alpha_{j_{2i}}^*) = \left(\frac{w_i^*}{\sqrt{\|w_i^*\|_2}}, -\sqrt{\|w_i^*\|_2}\right) \quad \text{if} \quad w_i^* \neq 0\,,$$

where v_i^*, w_i^* are the optimal solutions to (19).

[**Proof of Theorem 5**] The proof parallels the proof of the main result section and Theorem 6. We note that dual constraint set remains the same, and analogous strong duality results apply as we show next.

We also show that our dual characterization holds for arbitrary convex loss functions.

$$\min_{\{u_j,\alpha_j\}_{j=1}^m} \ell\left(\sum_{j=1}^m (Xu_j)_+\alpha_j, y\right) + \beta\|\alpha\|_1 \text{ s.t. } \|u_j\|_2 \leq 1,\ \forall j, \tag{20}$$

where $\ell(\cdot, y)$ is a convex loss function.

Theorem 6. The dual of (20) is given by

$$\max_v -\ell^*(v) \text{ s.t. } |v^T(Xu)_+| \leq \beta,\ \forall u \in \mathcal{B}_2\,,$$

where ℓ^* is the Fenchel conjugate function defined as

$$\ell^*(v) = \max_z z^T v - \ell(z, y)\,.$$

[**Proof of Theorem 6**] The proof follows from classical Fenchel duality (Boyd & Vandenberghe, 2004b). We first describe (20) in an equivalent form as follows

$$\min_{z,\{u_j,\alpha_j\}_{j=1}^m} \ell(z, y) + \beta\|\alpha\|_1 \text{ s.t. } z = \sum_{j=1}^m (Xu_j)_+\alpha_j,\ \|u_j\|_2 \leq 1, \forall j.$$

Then the dual function is

$$g(v) = \min_{z,\{u_j,\alpha_j\}_{j=1}^m} \ell(z, y) - v^T z + v^T \sum_{j=1}^m (Xu_j)_+\alpha_j + \beta\|\alpha\|_1 \text{ s.t. } \|u_j\|_2 \leq 1, \forall j.$$

Therefore, using the classical Fenchel duality (Boyd & Vandenberghe, 2004b) yields the claimed dual form.