Optimal Sets and Solution Paths of ReLU Networks

Aaron Mishkin [1] **Mert Pilanci** [2]

Abstract

We develop an analytical framework to characterize the set of optimal ReLU neural networks by reformulating the non-convex training problem as a convex program. We show that the global optima of the convex parameterization are given by a polyhedral set and then extend this characterization to the optimal set of the non-convex training objective. Since all stationary points of the ReLU training problem can be represented as optima of sub-sampled convex programs, our work provides a general expression for all critical points of the non-convex objective. We then leverage our results to provide an optimal pruning algorithm for computing minimal networks, establish conditions for the regularization path of ReLU networks to be continuous, and develop sensitivity results for minimal ReLU networks.

1. Introduction

Neural networks have transformed machine learning. Despite their success, little is known about the global optima for typical non-convex training problems, the solution path of regularized networks, or how to prune networks without degrading the model fit. This is in stark contrast to generalized linear models with ℓ_2 or ℓ_1 penalties; for example, it is well-known that the lasso (Tibshirani, 1996) has a piecewise linear path (Osborne et al., 2000; Efron et al., 2004), a polyhedral solution set (Tibshirani, 2013), and admits efficient algorithms for computing minimal solutions (Tibshirani, 2013). In this paper, we close the gap by studying neural networks through the lens of convex reformulations.

One of the main challenges of neural networks is non-convexity. For non-convex problems, stationarity of the training objective does not imply optimality of the network weights and so, to the best of our knowledge, no work has been able to derive an analytical expression for the optimal set. Convex reformulations provide a solution by rewriting the non-convex optimization problem as a convex program in a lifted parameter space (Pilanci & Ergen, 2020). We focus on the convex reformulation for two-layer networks with ReLU activation functions and weight decay regularization. The resulting problem is related to the group lasso (Yuan & Lin, 2006) and induces *neuron sparsity* in the network.

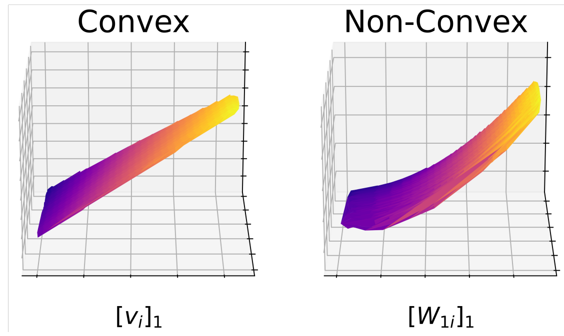

Figure 1. Convex vs non-convex solution spaces for two-layer ReLU networks. We plot the first feature of three different neurons; the non-convex parameterization maps the compact polytope of solutions for the convex parameterization into a curved manifold.

Let $Z \in \mathbb{R}^{n \times d}$ be a data matrix and $y \in \mathbb{R}^n$ associated targets. The prediction function for two-layer ReLU networks is

$$f_{W_1,w_2}(Z) = \sum_{i=1}^{m} (ZW_{1i})_+ \, w_{2i},$$

where $W_1 \in \mathbb{R}^{m \times d}$, $w_2 \in \mathbb{R}^m$ are the weights of the first and second layers, m is the number of hidden units, and $(\cdot)_+ = \max\{\cdot, 0\}$ is the ReLU activation. Fitting f_{W_1,w_2} with convex loss L with weight decay (ℓ_2) regularization leads to the standard non-convex optimization problem:

$$\min_{W_1,w_2} L\Big(f_{W_1,w_2}(Z), y\Big) + \frac{\lambda}{2} \underbrace{\left(\|W_1\|_F^2 + \|w_2\|_2^2\right)}_{R(W_1,w_2)}. \quad (1)$$

The regularization path or *solution function* of this training problem is the mapping between the regularization parameter λ and the set of optimal model weights,

$$\mathcal{O}^*(\lambda) = \arg\min_{W_1,w_2} L\Big(f_{W_1,w_2}(Z), y\Big) + \frac{\lambda}{2} R(W_1, w_2). \quad (2)$$

In general, the optimal neural network is not unique and $\mathcal{O}^*(\lambda)$ will be set valued. Indeed, there are always at least

[1]Department of Computer Science, Stanford University [2]Department of Electrical Engineering, Stanford University. Correspondence to: Aaron Mishkin <amishkin@cs.stanford.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

$m!$ solutions since any permutation of the hidden units yields an identical model. We call the solution to a ReLU training *p-unique* when it is unique up to permutations.

We study $\mathcal{O}^*$ by re-writing Equation (1) as an instance of the *constrained group lasso* (CGL). We make the following contributions by analyzing CGL:

- We derive an analytical expression for the solution set of two-layer ReLU networks and simple criteria for the solution to be p-unique, i.e. unique up to permutation.
- We extend this characterization to show that the set of stationary points of a two-layer ReLU model is exactly
$$\begin{aligned} \mathcal{C}_\lambda = \big\{(W_1, w_2) : \tilde{\mathcal{D}} \subseteq \mathcal{D}_Z,\ & f_{W_1,w_2}(Z) = \hat{y}_{\tilde{\mathcal{D}}}, \\ & W_{1i} = (\alpha_i/\lambda)^{1/2} v_i(\tilde{\mathcal{D}}),\ w_{2i} = (\alpha_i\lambda)^{1/2}, \\ & \alpha_i \geq 0,\ i \in [m] \setminus \mathcal{S}_\lambda \implies \alpha_i = 0\big\}, \end{aligned} \tag{3}$$
where $\tilde{\mathcal{D}}$ is a set of sub-sampled activation patterns, $\hat{y}_{\tilde{\mathcal{D}}}$ is the unique optimal model fit using those patterns, and $v_i(\tilde{\mathcal{D}})$ are uniquely given by optimal parameters for the dual of the convex reformulation. See Figure 1.
- We provide an optimal pruning algorithm that can be used to compute minimal models — the smallest-width neural networks which are optimal for a given dataset and regularization parameter — and an intuitive extension for pruning beyond minimal models.
- We prove that the regularization path of ReLU networks is discontinuous in general and establish sufficient conditions for path to be closed/continuous.
- We give a simple algorithm for computing the unique ReLU network corresponding to the min-norm model in the convex lifting and, under additional constraint qualifications, develop differential sensitivity results for minimal ReLU networks.

In many cases, we obtain strictly stronger results for gated ReLU networks (Fiat et al., 2019), which correspond directly to an unconstrained group lasso problem (Mishkin et al., 2022). In particular, we give new sufficient conditions for (i) the group lasso to be unique, (ii) global continuity of the group lasso model fit, and (iii) weak differentiability of the solution function for gated ReLU networks.

The paper is structured as follows: we cover related work in Section 1.1 and introduce notation in Section 1.2. Then we provide details for convex reformulations of neural network in Section 2. Section 3 analyzes CGL and Section 4 interprets these results in the specific context of two-layer ReLU networks. Section 5 concludes with experiments.

1.1. Related Work

The Lasso and Group Lasso: Our work is most similar to Hastie et al. (2007), who consider homotopy methods, and Tibshirani (2013), who characterize the lasso solution set. Limited results exist beyond the lasso. Tibshirani & Taylor (2011) analyze the generalized lasso, while Yuan & Lin (2006) show the group lasso is piece-wise linear when X is orthogonal. Roth & Fischer (2008) partially characterize the group lasso solution set, while Vaiter et al. (2012) derive stability results and the degrees-of-freedom.

Convex Reformulations: Convex reformulations for neural networks have rapidly advanced since Pilanci & Ergen (2020); convolutions (Ergen & Pilanci, 2021b; Gupta et al., 2021), vector-outputs (Sahiner et al., 2021), batch-norm (Ergen et al., 2021), and deeper networks (Ergen & Pilanci, 2021a) have all been explored.

Neural Network Solution Sets: Characterizations of solution sets are largely empirical. Mode connectivity has been studied extensively, Garipov et al. (2018); Draxler et al. (2018). Nguyen (2019); Kuditipudi et al. (2019) attempt to theoretically explain mode connectivity. Sensitivity is connected to differentiable optimization layers (Agrawal et al., 2019) and hypergradient descent (Baydin et al., 2017) We refer to Blalock et al. (2020) for an overview on pruning.

1.2. Notation

We use lower-case a to denote vectors and upper-case A for matrices. For $d \in \mathbb{N}$, $[d] = \{1, \ldots, d\}$. Calligraphic letters $\mathcal{C}$ denote sets. For a block of indices $b_i \subseteq [d]$, we write A_{b_i} for the sub-matrix of columns indexed by b_i. Similarly, a_{b_i} is the sub-vector indexed by b_i. If $\mathcal{M}$ is a collection of blocks, then $A_{\mathcal{M}}$ is the submatrix and $a_{\mathcal{M}}$ the sub-vector with columns/elements indexed by blocks in the collection. Finally, $|\mathcal{M}|$ is cardinality of the union of blocks in $\mathcal{M}$.

2. Convex Reformulations

Now we introduce background on convex reformulations. Convex reformulations re-write Equation (1) as a convex program by enumerating the activations a single neuron in the hidden layer can take on for fixed Z as follows:

$$\mathcal{D}_Z = \left\{D = \mathrm{diag}(\mathbb{1}(Zu \geq 0)) : u \in \mathbb{R}^d\right\}.$$

This set grows as $|\mathcal{D}_Z| \leq O(r(n/r)^r)$, where $r := \mathrm{rank}(Z)$ (Pilanci & Ergen, 2020). Each "activation pattern" $D_i \in \mathcal{D}_Z$ is associated with a convex cone,

$$\mathcal{K}_i = \left\{u \in \mathbb{R}^d : (2D_i - I)Zu \succeq 0\right\}.$$

If $u \in \mathcal{K}_i$, then u matches D_i, meaning $D_i Z u = (Zu)_+$. For any subset $\tilde{\mathcal{D}} \subseteq \mathcal{D}_Z$, the convex reformulation is,

$$\begin{aligned} \min_{v,w} L\Big(\sum_{D_i \in \tilde{\mathcal{D}}} D_i Z(v_i - u_i), y\Big) + \lambda \sum_{D_i \in \tilde{\mathcal{D}}} \|v_i\|_2 + \|u_i\|_2 \\ \text{s.t.} \quad v_i, u_i \in \mathcal{K}_i. \end{aligned} \tag{4}$$

Pilanci & Ergen (2020) prove that this program and Equation (1) are equivalent in the following sense: if $\tilde{\mathcal{D}} = \mathcal{D}_Z$

and $m \geq m^*$ for some $m^* \leq n+1$, then the two programs have the same optimal value and every solution to the convex program can be mapped to a solution of the non-convex training problem and vice-versa. Given a solution (v^*, u^*), optimal weights for the ReLU problem are given by

$$\begin{aligned} W_{1i} &= v_i^* / \sqrt{\|v_i^*\|}, \quad w_{2i} = \sqrt{\|v_i^*\|} \\ W_{1j} &= u_i^* / \sqrt{\|u_i^*\|}, \quad w_{2j} = -\sqrt{\|u_i^*\|}, \end{aligned} \tag{5}$$

where we use the convention that $0/0 = 0$.

In practice, learning with $\mathcal{D}_Z$ is intractable except when the data are low rank. Mishkin et al. (2022) provide refined conditions on $\tilde{\mathcal{D}}$ which are sufficient for Equation (4) to be equivalent to the non-convex problem, while Wang et al. (2021) show that the minimum of every sub-sampled convex program is a stationary point of the ReLU training problem.

2.1. Gated ReLU Networks

An alternative is the gated ReLU activation function,

$$\phi_g(Z, u) = \text{diag}(\mathbb{1}(Zg \geq 0))Zu,$$

where $g \in \mathbb{R}^d$ is a "gate" vector, which is also optimized. The gated ReLU activation modifies the ReLU activation to decouple the thresholding operator from the neuron weights. Two-layer gated ReLU networks predict as follows:

$$h_{W_1, w_2}(Z) = \sum_{i=1}^{m} \phi_{g_i}(Z, W_{1i}) w_{2i}. \tag{6}$$

Mishkin et al. (2022) show that this gated ReLU neural network has the convex reformulation,

$$\min_u \; L\Big(\sum_{D_i \in \tilde{\mathcal{D}}} D_i Z w_i, y\Big) + \lambda \sum_{D_i \in \tilde{\mathcal{D}}} \|w_i\|_2, \tag{7}$$

where decoupling the activations from the neuron weights allows $u_i, v_i \in \mathcal{K}_i$ to be merged. The solution mapping for w^* and conditions for for the convex program to be equivalent to Equation (6) are similar to the ReLU case.

3. The Constrained Group Lasso

In this section, we develop properties of CGL, a generalized linear model which captures both the convex ReLU and convex gated ReLU programs. Let $\mathcal{B} = \{b_1, \ldots, b_m\}$ be a disjoint partition of the feature indices $[d]$. Given regularization parameter $\lambda \geq 0$, CGL solves the program:

$$\begin{aligned} p^*(\lambda) = \min_w F_\lambda(w) := \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \\ \text{s.t.} \quad K_{b_i}^\top w_{b_i} \leq 0 \text{ for all } b_i \in \mathcal{B}, \end{aligned} \tag{8}$$

where $K_{b_i} \in \mathbb{R}^{|b_i| \times a_{b_i}}$. Solutions to Equation (8) are block sparse when λ is sufficiently large, meaning $w_{b_i} = 0$ subset of b_i. This is similar to the feature sparsity given by the lasso, to which CGL naturally reduces when $b_i = \{i\}$ and $K_{b_i} = 0$ for each $b_i \in \mathcal{B}$. Although we consider squared-error, our results generalize to strictly convex losses — see Appendix C for comments.

The convex reformulations introduced in the previous section are instances of CGL using the basis function $X = [D_1 Z \ldots D_p Z]$, where $p = |\tilde{\mathcal{D}}_Z|$. For gated ReLU models, $K_{b_i} = 0$ while ReLU models set $K_{b_i} = -Z^\top(2D_i - I)$. For both problems, block sparsity from the group ℓ_1 penalty induces neuron sparsity in the final solution.

Our goal is to characterize the solution function of CGL,

$$\mathcal{W}^*(\lambda) := \mathop{\arg\min}_{w \,:\, K_{b_i}^\top w_{b_i}} F_\lambda(w)$$

For a general data matrix, F_λ is not strictly convex and CGL may admit multiple solutions — these correspond to networks which are not related by permutation. As such, $\mathcal{W}^*$ is a point-to-set map and we must use a criterion to define a function; for instance, the min-norm solution mapping

$$w^*(\lambda) = \arg\min \{\|w\|_2 : w \in \mathcal{W}^*(\lambda)\},$$

defines a function for all $\lambda \geq 0$.

Now we introduce notation that will be used throughout this section. Let $\hat{y}(\lambda) = Xw$ for $w \in \mathcal{W}^*(\lambda)$ denote the optimal model fit, which is the same for any choice of optimal w (Lemma A.1). Similarly, define the optimal residual $r(\lambda) := y - \hat{y}(\lambda)$ and $c_{b_i}(\lambda) := X_{b_i}^\top r(\lambda)$ as the correlation vector for block b_i. We write $c \in \mathbb{R}^d$ for the concatenation of these block-vectors. Finally, let ρ_{b_i} be the dual parameters for the constraint $K_{b_i}^\top w_{b_i} \leq 0$, ρ their concatenation, and K the block-diagonal matrix with blocks given by K_{b_i}.

The Lagrangian associated with Equation (8) is

$$\mathcal{L}(w, \rho) = \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 + \langle K\rho, w \rangle. \tag{9}$$

The constraints are linear, strong duality attains if feasibility holds, and the necessary and sufficient conditions for primal-dual pair (w, ρ) to be optimal are the KKT conditions:

$$\begin{aligned} X_{b_i}^\top(Xw - y) + K_{b_i}\rho_{b_i} + s_{b_i} &= 0 \\ K_{b_i}^\top w_{b_i} &\leq 0 \\ [\rho_{b_i}]_j \cdot [K_{b_i}]_j^\top w_{b_i} &= 0 \quad \forall j \in [a_{b_i}] \\ \rho_{b_i} &\geq 0, \end{aligned} \tag{10}$$

where $s_{b_i} \in \partial \lambda \|w_{b_i}\|_2$. Since the KKT conditions hold for every combination of optimal primal-dual pair (Boyd & Vandenberghe, 2014), we always use the min-norm dual optimal parameter ρ^* with no loss of generality. To simplify our notations, we define $v_{b_i} := c_{b_i} - K_{b_i}\rho_{b_i}^*$. In what follows, all proofs are deferred to Appendix A.

3.1. Describing the Optimal Set

Stationary of the Lagrangian implies the equicorrelation set

$$\mathcal{E}_\lambda = \left\{ b_i \in \mathcal{B} : \|v_{b_i}\|_2 = \lambda \right\}.$$

contains all blocks which may be active for fixed λ. That is, the active set $\mathcal{A}_\lambda(w) = \{b_i : w_{b_i} \neq 0\}$ satisfies $\mathcal{A}_\lambda(w) \subseteq \mathcal{E}_\lambda$ for every $w \in \mathcal{W}^*(\lambda)$. However, not all blocks in $\mathcal{E}_\lambda$ may be non-zero for some solution. Thus, we define

$$\mathcal{S}_\lambda = \left\{ b_i \in \mathcal{B} : \exists w \in \mathcal{W}^*(\lambda), w_{b_i} \neq 0 \right\},$$

which is the set of blocks supported by some solution.

Our first result combines KKT conditions with uniqueness of $\hat{y}(\lambda)$ to characterize the solution set for fixed $\lambda > 0$.

Proposition 3.1. *Fix $\lambda > 0$. The optimal set for the CGL problem is given by*

$$\begin{aligned} \mathcal{W}^*(\lambda) = \big\{ w \in \mathbb{R}^d : \forall\, b_i \in \mathcal{S}_\lambda, w_{b_i} = \alpha_{b_i} v_{b_i}, \alpha_{b_i} \geq 0, \\ \forall\, b_j \in \mathcal{B} \setminus \mathcal{S}_\lambda, w_{b_j} = 0,\, Xw = \hat{y} \big\} \end{aligned} \tag{11}$$

Since this characterization is implicit due to the dependence on $\mathcal{S}_\lambda$, we also give an alternative and explicit construction in Proposition A.2, which shows that when $K = 0$ we may replace $\mathcal{S}_\lambda$ with $\mathcal{E}_\lambda$; We prefer Proposition 3.1 to Proposition A.2 since it better mirrors this simpler setting. However, Proposition A.2 can be substituted wherever desired.

Now that we know "shape" of the solution set, it is possible to obtain simple conditions for existence of a unique solution. As an immediate consequence of Proposition 3.1, the solution map is a subset of directions in $\text{Null}(X_{\mathcal{E}_\lambda})$.

Corollary 3.2. *If $w, w' \in \mathcal{W}^*(\lambda)$ and $z' = w - w'$, then*

$$z'_{\mathcal{E}_\lambda} \in \mathcal{N}_\lambda := \textit{Null}(X_{\mathcal{E}_\lambda}) \cap \left\{ z_{\mathcal{E}_\lambda} : \forall b_i \in \mathcal{E}_\lambda,\, z_{b_i} = \alpha_{b_i} v_{b_i} \right\}.$$

As a result, the group lasso solution is unique if $\mathcal{N}_\lambda = \{0\}$.

Corollary 3.2 extends a similar result for the lasso to CGL (Tibshirani, 2013, Eq. 9) and implies the solution is unique for all $\lambda \geq 0$ if the columns of X are linearly independent. The corollary also provides a simple check for primal uniqueness given a primal-dual solution pair.

Lemma 3.3. *Fix $\lambda > 0$. The solution to CGL problem is unique if and only if $\{X_{b_i} v_{b_i}\}_{\mathcal{S}_\lambda}$ are linearly independent.*

Note that a dual solution ρ is necessary to compute v in general; By uniformizing over v_{b_i}, we obtain a stronger condition that can be checked whenever $\mathcal{E}_\lambda$ is known, yet is still weaker than linear independence of the columns of X.

Corollary 3.4. *If the columns of $X_{\mathcal{E}_\lambda}$ are linearly independent, then CGL problem has a unique solution.*

Finally, we consider the special case when there are no constraints and $K = 0$. In this setting, $v_{b_i} = c_{b_i}$ — the dual parameters are trivially zero — and we can provide a global condition which is much stronger than linear independence.

Proposition 3.5. *[Group General Position] Suppose for every $\mathcal{E} \subseteq \mathcal{B}$, $|\mathcal{E}| \leq n + 1$, there do not exist unit vectors $z_{b_i} \in \mathbb{R}^{|b_i|}$ such that for any $j \in \mathcal{E}$,*

$$X_{b_j} z_{b_j} \in \textit{affine}(\{X_{b_i} z_{b_i} : b_i \in \mathcal{E} \setminus b_j\}).$$

Then the group lasso solution is unique for all $\lambda > 0$.

We call this uniqueness condition *group general position* (GGP) because it naturally extends general position to groups of vectors. General position itself is an extension of affine independence and is sufficient for the lasso solution to be unique (Tibshirani, 2013). GGP is strictly weaker than linear independence of the columns of X, but neither implies nor is implied by general position (Proposition A.3).

3.2. Computing Dual Optimal Parameters

The main difficulty of Lemma 3.3 is that knowledge of a dual optimal parameter is required to check if a unique solution exists. A dual optimal parameter is also required to fully leverage our characterization of the optimal set. As such, now we turn to computing optimal dual parameters.

We give one Lagrange dual problem for CGL in Lemma A.4. A nice feature of this dual problem is v_{b_i} attains an alternative interpretation as dual variable. However, evaluating the dual requires computing $(X^\top X)^+$, which may be difficult even if X is structured, as in the convex ReLU program. Instead, we focus on computing ρ given a primal solution.

Let $w \in \mathcal{W}^*(\lambda)$. If $w_{b_i} \neq 0$, then KKT conditions imply

$$K_{b_i} \rho_{b_i} = c_{b_i} - \lambda \frac{w_{b_i}}{\|w_{b_i}\|_2}, \tag{12}$$

so that the "dual fit" $\hat{d}_{b_i} = K_{b_i} \rho_{b_i}$ is easily computed. Recovering the dual parameter is a linear feasibility problem:

$$\rho_{b_i} \in \left\{ \rho_{b_i} \geq 0 : K_{b_i} \rho_{b_i} = \hat{d}_{b_i} \right\}. \tag{13}$$

If $w_{b_i} = 0$, then complementary slackness is trivially satisfied and we compute the min-norm dual parameter by solving the following program:

$$\min \left\{ \|\rho_{b_i}\|_2 : \|c_{b_i} - K_{b_i} \rho_{b_i}\|_2 \leq \lambda, \rho_{b_i} \geq 0 \right\}. \tag{14}$$

In general, however, we only need *some* dual optimal parameter for our results to hold; thus, is is typically easier to find ρ by solving the following non-negative regression:

$$\rho_{b_i} = \arg\min \left\{ \|K_{b_i} \rho_{b_i} - c_{b_i}\|_2^2 : \rho_{b_i} \geq 0 \right\}. \tag{15}$$

See Proposition A.5 for details.

3.3. Minimal Solutions and Optimal Pruning

Often we want the most parsimonious solution, i.e. the one using the fewest feature groups. We say a primal solution w is *minimal* if there does not exist $w' \in \mathcal{W}^*(\lambda)$ such that $\mathcal{A}_\lambda(w') \subsetneq \mathcal{A}_\lambda(w)$. Building on the previous section, we start with a sufficient condition for w to be minimal.

Algorithm 1 Optimal Solution Pruning

Input: data matrix X, solution w.
$k \leftarrow 0$.
$w^k \leftarrow w$.
while $\exists \beta \neq 0$ s.t. $\sum_{b_i \in \mathcal{A}_\lambda(w^k)} \beta_{b_i} X_{b_i} w_{b_i}^k = 0$ **do**
  ${b_i}^k \leftarrow \arg\max_{b_i} \left\{ |\beta_{b_i}| : b_i \in \mathcal{A}_\lambda(w^k) \right\}$
  $t^k \leftarrow 1/|\beta_{{b_i}^k}|$
  $w^{k+1} \leftarrow w^k(1 - t^k \beta_{b_i})$
  $k \leftarrow k + 1$
end while
Output: final weights w^k

Proposition 3.6. *For $\lambda > 0$, $w \in \mathcal{W}^*(\lambda)$ is minimal if and only if the vectors $\{X_{b_i} w_{b_i}\}_{\mathcal{A}(w)}$ are linearly independent.*

Linear independence of $\{X_{b_i} w_{b_i}\}_{\mathcal{A}(w)}$ also identifies w as a vertex of $\mathcal{W}^*(\lambda)$ (Bertsekas, 2009), meaning minimal models are exactly the extreme points of the optimal set. Combining this characterization with our condition for uniqueness of a solution (Lemma 3.3) shows that minimal solutions are the only solution on their support.

Corollary 3.7. *Suppose w is a minimal solution. Then w is the unique solution with support $\mathcal{A}_\lambda(w)$.*

Furthermore, all minimal solutions are equivalent in the sense that they have the same number of active blocks.

Proposition 3.8. *Let $\mathcal{V} = Span(\{X_{b_i} \bar{w}_{b_i}\})$ for $\bar{w} \in \mathcal{W}^*(\lambda)$. Every minimal solution has $c = dim(\mathcal{V})$ active blocks.*

Algorithm 1 gives a procedure which, starting from any optimal solution w, computes a optimal model with the smallest possible number of active blocks in $O((n^3 l + nd)$ time, where l is the number active blocks in w. (see Proposition A.6). Our algorithm can also be used to verify a minimal solution, since if w minimal then it is unique on its support and Algorithm 1 must return w immediately. This procedure also implies the existence of at least one minimal solution.

Corollary 3.9. *There exists $w \in \mathcal{W}^*(\lambda)$ for which the vectors $\{X_{b_i} w(\lambda) : b_i \in \mathcal{A}(w)\}$ are linearly independent.*

Corollary 3.9 will be useful tool later when we study sensitivity of the model fit to perturbations in y and λ.

A disadvantage of Algorithm 1 is that it cannot continue beyond a minimal solution. However, minimal models may still be quite large. We can perform approximate pruning in such cases using the least squares fit to approximate β,

$$\tilde{\beta} = \arg\min \|A\beta - X_{b_j} w_{b_j}\|_2^2,$$

where $A = [X_{b_i} w_{b_i}]_{\mathcal{A}_\lambda \setminus b_j}$ and $b_j \in \mathcal{A}_\lambda$ is chosen randomly. Using $\tilde{\beta}$ in Algorithm 1 is optimal when $\{X_{b_i} w_{b_i}\}_{\mathcal{A}_\lambda}$ are dependent and chooses the update parameters to minimize degradation of the model fit otherwise.

3.4. Continuity of the Solution Path

A major concern when learning with regularizers is how to tune the parameter λ. Typical strategies like grid-search on the (cross) validation loss are effective only if the solution function satisfies basic continuity properties. For example, if $\mathcal{W}^*$ is single-valued but discontinuous in λ, then the sample complexity of grid-search can be made arbitrarily poor by "hiding" the optimal λ in a discontinuity (Nesterov et al., 2018, Sec. 1.1). In this section, we justify grid-search for CGL by proving several continuity properties of the solution function, particularly when the solution is unique. We start with basic definitions of continuity for point-to-set maps.

Definition 3.10 (Closed)**.** $T : \mathcal{X} \to 2^{\mathcal{Z}}$ is closed if $\{x_k\} \subset \mathcal{X}$, $x_k \to \bar{x}$ and $z_k \in T(x_k)$, $z_k \to \bar{z}$ implies $\bar{z} \in T(\bar{x})$.

Definition 3.11 (Open)**.** $T : \mathcal{X} \to 2^{\mathcal{Z}}$ is open if $\{x_k\} \subset \mathcal{X}$, $x_k \to \bar{x}$ and $\bar{z} \in T(\bar{x})$, implies there exists $k' \in \mathbb{N}$, $z_k \in T(x_k)$ for $k \geq k'$, such that $z_k \to \bar{z}$.

We say that T is *continuous* if it is both closed and open. If $T(x)$ is a singleton for all $x \in \mathcal{X}$, then openness/closedness are equivalent and imply continuity. We start with (functional) continuity of the optimal objective.

Proposition 3.12. *$\lambda \mapsto p^*(\lambda)$ is continuous for all $\lambda \geq 0$.*

While standard sensitivity results imply that $\mathcal{W}^*$ is closed, unfortunately openness is not possible in the general setting.

Proposition 3.13. *While $\mathcal{W}^*$ is closed on $\mathbb{R}_+$, it is open 0 if only if X is full column rank. However, if the solution is unique on $\Lambda \subset \mathbb{R}^+$, then $\mathcal{W}^*$ is open at every $\lambda \in \Lambda$.*

As a corollary of Proposition 3.13, $\mathcal{W}^*$ is open on $\mathbb{R}^+$ if and only if X is full column rank. Continuity of $\mathcal{W}^*$ is impossible in general because, as Hogan (1973) shows, openness is a local stability property; since $\mathcal{W}^*(0)$ is unbounded, many "unstable" solutions exist at $\lambda = 0$ which are not limit points of other solutions. Continuity of the unique solution path is an immediate corollary of Proposition 3.13.

Corollary 3.14. *If the CGL solution is unique on an interval $\Lambda \subset \mathbb{R}_+$, then it is also continuous on Λ.*

In particular, if $K = 0$ and GGP holds, then the group lasso solution is continuous for all $\lambda > 0$. We can strengthen our continuity results when $K_{b_i} = 0$ in another way: by analyzing the dual of the group lasso problem, we extend continuity from p^* to the optimal model fit.

Proposition 3.15. *If $K = 0$, then $\hat{y}(\lambda)$ is continuous on $\mathbb{R}_+$ and the penalty $\sum_{b_i \in \mathcal{B}} \|w_{b_i}(\lambda)\|_2$ is continuous for $\lambda > 0$.*

3.5. The Min-Norm Path

Now we turn our attention to the min-norm solution path. Min-norm solutions are typically used in under-determined problems and the norm of the solution is connected to generalization performance (Neyshabur et al., 2015; Gunasekar et al., 2017) Furthermore, the min-norm solution is a function of λ, unlike $\mathcal{W}^*$. Throughout this section, $\mathcal{A}^*_\lambda = \mathcal{A}_\lambda(w^*)$ denotes the active set of the min-norm solution.

Unfortunately, studying the min-norm path immediately encounters a surprising difficulty: as opposed to least-squares problems or the lasso (see Tibshirani (2013)), the min-norm solution may not lie in the row space of the active set.

Proposition 3.16. *Suppose $K_{b_i} = 0$. There exists (X, y) and $\lambda > 0$ such that $w^*_{\mathcal{A}^*_\lambda}(\lambda) \notin \text{Row}(X_{\mathcal{A}^*_\lambda})$.*

Since the min-norm solution is not given by projecting onto $\text{Row}(X_{\mathcal{A}^*_\lambda})$, how can we compute and study it? Again, our characterization for the solution set provides a way forward.

Proposition 3.17. *Let $\lambda > 0$ and consider the program:*

$$\alpha^* = \mathop{\arg\min}_{\alpha \geq 0} \|\alpha\|_2^2 \;\; \text{s.t.} \sum_{b_i \in \mathcal{S}_\lambda} \alpha_{b_i} X_{b_i} v_{b_i} = \hat{y}. \tag{16}$$

Then the min-norm solution is given by $w^*_{b_i} = \alpha^*_{b_i} v_{b_i}$.

Equation (16) is a quadratic program (QP) that can be solved with off-the-shelf software like CVXPY (Diamond & Boyd, 2016). If $|\mathcal{B}|$ and d are large, this QP may be too expensive to handle directly. In such situations, we propose to the solve the following elastic-net-type problem

$$\begin{aligned} \min_w \; & \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 + \frac{\delta}{2}\|w\|_2^2 \\ \text{s.t.} \;\; & K_{b_i}^\top w_{b_i} \leq 0 \text{ for all } b_i \in \mathcal{B}. \end{aligned} \tag{17}$$

This ℓ_2-penalized CGL problem is equivalent to CGL with modified dataset $(\tilde{X}, \tilde{y})$ (Lemma A.14). Since the optimization problem is strongly convex ($\tilde{X}$ is full column rank), invoking Proposition 3.13 implies the solution $w^\delta(\lambda)$ is continuous for all $\lambda \geq 0$. Moreover, as $\delta \to 0$, the penalized solution converges to the min-norm solution to CGL.

Proposition 3.18. *The solution to the ℓ_2-penalized problem converges to the min-norm solution as $\delta \to 0$. That is,*

$$\lim_{\delta \to 0} w^\delta(\lambda) = w^*(\lambda).$$

Uniqueness and continuity of the solution path for the penalized CGL problem mean we may prefer to solve Problem (17) with small $\delta > 0$ when tuning λ. Proposition 3.18 guarantees that the bias induced by δ will be small and a polishing step with $\delta = 0$ can always be used. Finally, nonzero δ ensures the objective is strongly convex, meaning we can use linearly convergent methods to solve the problem.

3.6. Sensitivity

Now we move onto the problem of sensitivity of a solution $w \in \mathcal{W}^*(\lambda, y)$ to perturbations, either in λ or the targets y. The main tool for measuring such perturbations are the gradients, for example $\nabla_\lambda w(\lambda, y)$. However, since the solution path of the group lasso is non-smooth, we must cope with the fact that gradients are not available everywhere.

We show that the gradients of minimal solutions exist almost everywhere under additional constraint qualifications (CQs). We do so by considering a *reduced* problem and showing that the solution to this reduced problem is exactly $w_{\mathcal{A}_\lambda}$. Define the reduced problem as follows:

$$\begin{aligned} \min_{w_{\mathcal{A}_\lambda}} \; & \frac{1}{2}\|X_{\mathcal{A}_\lambda} w_{\mathcal{A}_\lambda} - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{A}_\lambda} \|w_{b_i}\|_2 \\ \text{s.t.} \;\; & K_{\mathcal{A}_\lambda}^\top w_{\mathcal{A}_\lambda} \leq 0 \end{aligned} \tag{18}$$

If $\mathcal{A}_\lambda(w)$ is the support of a minimal solution, then w is the only solution with support $\mathcal{A}_\lambda$ and Equation (18) can be used to compute the unique active weights.

Proposition 3.19. *Let $w \in \mathcal{W}^*(\lambda, y)$ be minimal. The active blocks $w_{\mathcal{A}_\lambda}$ are the unique solution to Problem (18).*

We use this fact to obtain a local solution function for CGL using the implicit function theorem. Given a solution w, let

$$\mathcal{B}(w) = \bigcup_{b_i \in \mathcal{A}_\lambda} \left\{ j \in [a_{b_i}] : [K_{b_i}]_j^\top w_{b_i} = 0 \right\},$$

be the active constraints. We now need two classical CQs.

Definition 3.20 (LICQ)**.** $w \in \mathcal{W}^*(\lambda, y)$ satisfies linear independence CQ if $\{[K]_j : j \in \mathcal{B}(w)\}$ are linearly independent.

Definition 3.21 (SCS)**.** Primal solution $w \in \mathcal{W}^*(\lambda)$ satisfies strict complementary slackness if there exists a dual optimal parameter ρ such that $[\rho]_j > 0$ for every $j \in \mathcal{B}$.

Now we can state our main differential sensitivity result.

Proposition 3.22. *Let $w \in \mathcal{W}^*(\bar{\lambda}, \bar{y})$ be minimal and suppose w satisfies LICQ on the active set $\mathcal{A}_\lambda$ and SCS on the equicorrelation set $\mathcal{E}_\lambda$. Then w has a locally continuous solution function $(\lambda, y) \mapsto w(\lambda, y)$. Moreover, if*

$$D = \begin{bmatrix} X_{\mathcal{A}_\lambda}^\top X_{\mathcal{A}_\lambda} + M(\bar{w}) & K_{\mathcal{A}_\lambda} \\ \rho_{\bar{\mathcal{A}}_\lambda} \odot K_{\mathcal{A}_\lambda} & \text{diag}(K_{\mathcal{A}_\lambda}^\top w_{\bar{\mathcal{A}}_\lambda}) \end{bmatrix},$$

where $\odot$ is the element-wise product, $u_{b_i} = \frac{w_{b_i}}{\|w_{b_i}\|_2}$, u is the concatenation of these vectors, and M is block-diagonal projection matrix in Equation (26), then the Jacobians of $w(\bar{\lambda}, \bar{y})$ with respect to λ and y are given as follows:

$$\nabla_\lambda w(\bar{\lambda}, \bar{y}) = -[D^{-1}]_{\mathcal{A}_\lambda} u_{\mathcal{A}_\lambda} \;\; \nabla_y w(\bar{\lambda}, \bar{y}) = [D^{-1}]_{\mathcal{A}_\lambda} X_{\mathcal{A}_\lambda}^\top,$$

where $[D^{-1}_{\mathcal{A}_\lambda}]_{\mathcal{A}_\lambda}$ is the $|\mathcal{A}_\lambda| \times |\mathcal{A}_\lambda|$ dimensional leading principle submatrix of D.

Algorithm 2 Approximate ReLU Pruning

Input: data matrix Z, weights W_1, w_2, score function s.
$m \leftarrow |\mathcal{A}_\lambda(W_1)|$
$(W_1^0, w_2^0) \leftarrow (W_1, w_2)$.
$q_i^0 \leftarrow (XW_{1i}^0)_+ w_2^0$
for $k = 0$ to $m-1$ **do**
  $j^k = \arg\min_{i \in \mathcal{A}_\lambda(W_{1i}^k)} s(W_{1i}^k)$
  $\beta^k = \arg\min_\beta \| \sum_{i \neq j^k} \beta_i q_i^k - q_{j^k}^k \|_2^2$
  $i^k \leftarrow \arg\max_i \left\{ |\beta_i| : i \in \mathcal{A}_\lambda(W_{1i}^k) \right\}$
  $t^k \leftarrow 1/|\beta_{i^k}|$
  $(W_{1i}^{k+1}, w_{2i}^{k+1}) \leftarrow (W_{1i}^k, w_{2i}^k) \cdot (1 - t^k \beta_i)^{1/2}$
  $q_i^{k+1} \leftarrow q_i^k \cdot (1 - t^k \beta_i)$
end for
Output: final weights W_1^k, w_2^k

4. Specialization to Neural Networks

Now we specialize our results for CGL to two-layer neural networks with ReLU or gated ReLU activations. We state and prove our results for ReLU networks, but they are easily adapted to gated ReLUs. We start by interpreting conditions for uniqueness in the context of non-convex ReLU models and then move on to discussing optimal pruning for ReLU networks and continuity properties of the solution function. Proofs are deferred to Appendix B.

Optimal Sets and Uniqueness: Combining the mapping between solutions for the convex reformulation and the original non-convex training problem (Equation (5)) and Proposition 3.1 immediately allows us to characterize the solution set for the full ReLU problem:

Corollary 4.1. *Suppose $m \geq m^*$ and $\tilde{D} = \mathcal{D}_Z$ (no sub-sampling), with $p = |\tilde{\mathcal{D}}_Z|$. Then the optimal set for the ReLU problem is*

$$\begin{aligned}\mathcal{O}_\lambda = \big\{(W_1, w_2)\colon & f_{W_1,w_2}(Z) = \hat{y}, W_{1i} = (\alpha_i/\lambda)^{1/2} v_i, \\ & w_{2i} = (\alpha_i \lambda)^{1/2}, \alpha_i \geq 0,\ i \in [2p] \setminus \mathcal{S}_\lambda \Rightarrow \alpha_i = 0\big\},\end{aligned} \tag{19}$$

Eq. (19) abuses notation slightly by using $\mathcal{S}_\lambda$ as a subset set of the neuron indices $\{1, \ldots 2p\}$, where indices $\{1, \ldots p\}$ index the positive neurons (neurons corresponding to blocks $D_i Z$ in the convex program) and $\{p+1, \ldots, 2p\}$ index the negative neurons ($-D_i Z$). Figure 1 plots the first feature of three neurons as they vary over this solution set. The mapping from convex to non-convex parameterization transforms the flat polytope of solutions into a curved manifold.

Choosing a sub-sampled set of patterns $\tilde{\mathcal{D}} \subset \mathcal{D}_Z$ corresponds to finding a stationary point of the non-convex training problem (Wang et al., 2021). Using this fact with Corollary 4.1 finally justifies description of all stationary points of the ReLU problem given in the introduction.

Proposition 4.2. *The set of stationary points of two-layer ReLU networks is given by*

$$\begin{aligned}\mathcal{C}_\lambda = \big\{(W_1, w_2) : & \tilde{\mathcal{D}} \subseteq \mathcal{D}_Z,\ f_{W_1,w_2}(Z) = \hat{y}_{\tilde{\mathcal{D}}}, \\ & W_{1i} = (\alpha_i/\lambda)^{1/2} v_i(\tilde{\mathcal{D}}),\ w_{2i} = (\alpha_i \lambda)^{1/2}, \\ & \alpha_i \geq 0,\ i \in [2|\tilde{\mathcal{D}}|] \setminus \mathcal{S}_\lambda \implies \alpha_i = 0\big\},\end{aligned}$$

where $\tilde{\mathcal{D}}$ are sub-sampled activation patterns, $\hat{y}_{\tilde{\mathcal{D}}}$ is the optimal model fit using those patterns, and $v_i(\tilde{\mathcal{D}}) = c_{b_i}(\tilde{\mathcal{D}}) - K_{b_i} \rho_{b_i}(\tilde{\mathcal{D}})$ is determined by the fit and the dual parameters.

We note that since deeper networks are also related to CGL through convex reformulations (Ergen & Pilanci, 2021a), Proposition 4.2 may also be applied beyond two layers.

Recall that the solution set for a two-layer ReLU network is typically not unique due to permutation symmetries. However, if the convex solution is unique, then the non-convex ReLU training problem is p-unique. Combining this with our results for CGL gives the following sufficient conditions.

Proposition 4.3. *Let $\lambda > 0$ and suppose that the convex ReLU problem has a unique solution. Then the ReLU model solution is p-unique. In particular, if $\{D_i Z v_{b_i}\}_{\mathcal{E}_\lambda}$ are linearly independent, then the non-convex solution is p-unique.*

For gated ReLU networks, it is also sufficient to check the blocks $[D_i Z]_{D_i \in \tilde{\mathcal{D}}}$ to see if they satisfy GGP. By looking at the structure of $D_i X$, we give simple sufficient conditions for sub-sampled convex ReLU programs to be unique.

Proposition 4.4. *Let $\lambda > 0$ and $p = |\tilde{\mathcal{D}}|$. Suppose Z follows a continuous probability distribution and nnz$(D_i) \geq p \cdot d$ for every $D_i \in \tilde{\mathcal{D}}$. If $\mathcal{E}_\lambda$ does not contain two blocks with the same activation pattern, then the sub-sampled convex ReLU program has a unique solution almost surely.*

Proposition 4.4 requires n to be much greater than d to be useful due to the trivial bound nnz$(D_i) \leq n$. In practice, the condition on activation patterns can be enforced by constraining $v_i = 0$ or $w_i = 0$ for each activation pattern before solving the convex reformulation.

Pruning: If (v, u) is a minimal solution to the convex reformulation, then the corresponding ReLU network is the p-unique model using only those activation patterns (Propositions 3.6 and 4.3). Thus, Algorithm 1 can be used to prune any solution to obtain the "narrowest" neural network achieving the optimal training objective. Algorithm 2 specializes our pruning algorithm to the ReLU problem and extends it to support approximate pruning. Note the resulting procedure is completely independent of the convex reformulation. The complexity of this method is as follows.

Proposition 4.5. *Suppose $r = rank(X)$. Then an optimal and minimal ReLU network with at most $m^* \leq n$ non-zero neurons can be computed in $O\left(d^3 r^3 (n/r)^{3r}\right)$ time.*

As a consequence, the complexity of computing an optimal and minimal ReLU network is fully polynomial when r is bounded. We also have a more sensitive statement for

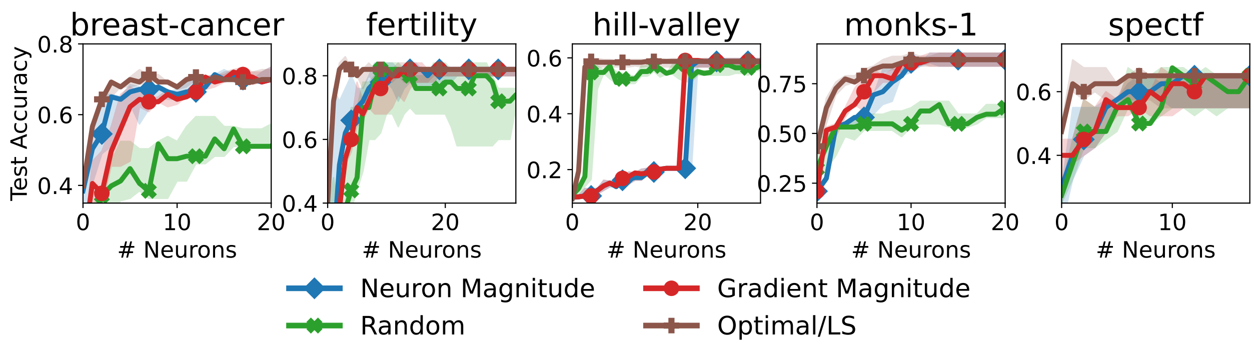

Figure 2. Pruning neurons from two-layer ReLU networks on binary classification tasks from the UCI repository. We compare our theory-inspired approach (**Optimal/LS**), against removing the neuron with smallest ℓ_2 norm (**Neuron Magnitude**), removing the neuron with the smallest weighted gradient norm (**Gradient Magnitude**), and random pruning (**Random**). For Optimal/LS, we use Algorithm 2, which begins with optimal pruning and then switches to a least-squares heuristic. We plot test accuracy against number of active neurons. Optimal/LS dominates the baseline methods on every dataset and even improves test accuracy on `breast-cancer` and `fertility`.

Table 1. Tuning neural networks by searching over the optimal set. We fit two-layer ReLU networks on the training set and compute the minimum ℓ_2 norm solution (Min L$_2$). Then we tune by finding an extreme point approximating the maximum ℓ_2-norm solution (EP), minimizing validation MSE over the optimal set (V-MSE), and minimizing test MSE over the optimal set (T-MSE). Results show median test accuracy; Max Diff. reports the difference between the best and worse models found. Exploring the optimal set reveals a huge disparity in the performance of optimal networks, with the generalization gap exceeding 20 points on four datasets.

Dataset	**Min L$_2$**	**EP**	**V-MSE**	**T-MSE**	**Max Diff.**
fertility	0.66	0.69	0.65	0.64	0.05
heart-hung.	0.75	0.75	0.71	0.85	0.14
mammogr.	0.77	0.77	0.57	0.78	0.21
monks-1	0.67	0.66	0.49	0.57	0.17
planning	0.53	0.52	0.53	0.7	0.17
spectf	0.64	0.64	0.56	0.58	0.08
horse-colic	0.75	0.59	0.74	0.85	0.26
ilpd-indian	0.59	0.59	0.53	0.72	0.19
parkinsons	0.74	0.74	0.65	0.88	0.23
pima	0.68	0.68	0.68	0.87	0.2

the minimal width: if (W_1^*, w_2^*) are optimal weights for the ReLU model, then m^* is exactly the dimensional of the span of the optimal activations $\{(XW_{1i}^*)_+\}_i$ (Proposition 3.8). We experiment with pruning ReLU networks using this approach in Section 5 and that show it is more effective than naive pruning strategies.

Continuity: First we give a negative result for singular networks, that is, models where $m < m^*$ and no convex reformulation exists. In this setting, the solution map can be made to behave arbitrarily poorly.

Proposition 4.6. *There exists (Z, y) for which $\mathcal{O}^*$ is not open nor is the model fit $f_{W_1,w_2}(Z)$ continuous in λ.*

Combined with the next result, Proposition 4.6 indicates that the threshold m^* may be crucial for continuity to extend to the non-convex parameterization.

Corollary 4.7. *Suppose $m \geq m^*$. Then the optimal model fit for two-layer gated ReLU networks is continuous at all $\lambda > 0$. Similarly, if the (gated) ReLU solution is p-unique on an open interval Λ, then the regularization path is also continuous on Λ up to permutations of the weights.*

Together, Corollary 4.7 and Proposition 4.4 are concrete conditions for the model fit and regularization path of a sub-sampled problem to be continuous.

Min-Norm Solutions: In Section 3.5, we examined the minimum ℓ_2-norm solution to CGL. However, all optimal ReLU networks have the same ℓ_2-norm when $\lambda > 0$. Minimizing the Euclidean norm of solutions to the convex reformulation instead selects for the network which minimizes the sum of neuron norms to the fourth power.

Lemma 4.8. *The minimum ℓ_2-norm solution to the convex reformulation of a (gated) ReLU model corresponds to the p-unique optimal neural network which minimizes*

$$r(W_1, w_2) = \sum_{i=1}^{m} \|W_{1i}\|_2^4 + \|w_{2i}\|_2^4.$$

As a result, we can compute the r-minimal optimal ReLU network by solving Problem (16). If $\mathcal{S}_\lambda$ is unknown, then using $\mathcal{A}_\lambda(w)$ for some solution w as an approximation gives the r-minimal network using a subset of those activations.

Sensitivity: Proposition 3.22 extends similar results for the group lasso by Vaiter et al. (2012) to CGL using standard CQs. Since K is block-diagonal, LICQ will be satisfied whenever the rows of Z are linearly independent. SCS is more challenging; while the classical theorem of Goldman & Tucker (2016) establishes that SCS is satisfied for linear programs, it is known that SCS can fail for general cone programs (Tunçel & Wolkowicz, 2012). As such, SCS must be checked on a per-problem basis in general.

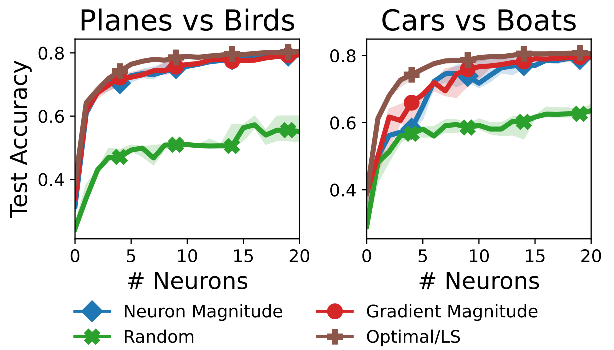

Figure 3. Pruning neurons from two-layer ReLU networks on two binary classification tasks drawn from the CIFAR-10 dataset. We compare our method (**Optimal/LS**) against baselines; see Figure 2 for details. Our approach, which makes use of a weight correction after pruning, outperforms every baseline.

In the context of gated ReLU problems, $K = 0$ and there is no requirement for CSC/LICQ. Minimal models $w(\lambda, y)$ are weakly differentiable, which Vaiter et al. (2012) uses to compute the degrees of freedom of w via Stein's Lemma (Stein, 1981). It is straightforward to extend this calculation to the gated ReLU weights using the chain rule, which can then be used to calculate Stein's unbiased risk estimator.

5. Experiments

Through convex reformulations, we have characterized the optimal sets of ReLU networks, minimal networks, and sensitivity results . Our goal in this section is to illustrate the power of our framework for analyzing ReLU networks and developing new algorithms.

Tuning: We first consider a tuning task on 10 binary classification datasets from the UCI repository (Dua & Graff, 2017). For each dataset, we do a train/validation/test split, fit a two-layer ReLU model on the training set, and then compute the minimum ℓ_2-norm model. We use this to explore the optimal set in three ways: (i) we compute an extreme point that (approximately) maximizes the model's ℓ_2-norm; (ii) we minimize the validation MSE over $\mathcal{W}^*(\lambda)$; (iii) we minimize test MSE over $\mathcal{W}^*(\lambda)$. These procedures select for different optimal models, have no effect on the training objective, and are only possible because we know $\mathcal{W}^*$.

The results are summarized in Table 1. We see that optimal models can perform very differently at test time despite having exactly the same training error and model norm. Indeed, 9/10 datasets show at least a 10 percent gap between the best and worst models and 4/10 have a gap exceeding 20 percent accuracy. We conclude that the training objective is badly under-determined even for shallow neural networks, implying that implicit regularization is critical in practice. See Appendix D for results on additional datasets.

Pruning: We also consider several neuron pruning task. We use two-layer ReLU networks and start pruning from the model given by optimizing the convex reformulation. We compare four strategies: (i) pruning neurons optimally using Algorithm 2 until $\{(XW_{1i})_+\}_{\mathcal{A}_\lambda}$ are linearly independent and then approximately using least-squares fits; and (ii) by removing the neuron with the smallest magnitude, $\|W_{1i} \cdot w_{2i}\|$; (iii) by remove the neuron with the smallest weighted gradient; and (iv) by random pruning.

Figure 2 shows test performance of the two methods for five UCI datasets. Our theory-based pruning method has better test performance than the baselines on every dataset considered; on `hill-valley`, the gap between our approach and magnitude-based pruning is approximately 40%. Figure 3 presents similar results for two binary tasks taken from the CIFAR-10 dataset (Krizhevsky et al., 2009). We provide experiments on additional datasets, including MNIST (LeCun et al., 1998), and experimental details in Appendix D.

6. Conclusion

We study the structure and properties of solution sets for shallow neural networks with (gated) ReLU activations. Unlike previous work, we avoid non-convexity of neural networks by studying the constrained group lasso, a generalized linear model which unifies the convex reformulations of both ReLU and gated ReLU networks. We derive analytical expressions for the optima and all stationary points of the training objective for two-layer ReLU networks. Building on this characterization, we develop conditions for the optimal neural network to be permutation unique, an algorithm for optimal pruning of neural networks, and sensitivity results. We demonstrate the utility of our framework in experiments on MNIST, CIFAR-10, and UCI datasets.

There is still much work to do in this area. For example, we conjecture that the min-norm CGL solution, which corresponds to the network minimizing a fourth-power penalty, always has a continuous regularization path. More generally, it remains to extend our characterization of the solution set to deeper networks and vector-output models.

Acknowledgements

This work was supported by the National Science Foundation (NSF), Grants ECCS-2037304, DMS-2134248; by the NSF CAREER Award, Grant CCF-2236829; by the U.S. Army Research Office Early Career Award, Grant W911NF-21-1-0242; by the Stanford Precourt Institute; by the Stanford Research Computing Center; and by the ACCESS—AI Chip Center for Emerging Smart Systems through InnoHK, Hong Kong, SAR. Aaron Mishkin was supported by NSF Grant DGE-1656518 and by NSERC Grant PGSD3-547242-2020. We thank Frederik Kunstner and Victor Sanches Portella for many insightful discussions.

References

Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., and Kolter, J. Z. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

ApS, M. Mosek optimizer API for python. *Version*, 9(17): 6–4, 2022.

Baydin, A. G., Cornish, R., Rubio, D. M., Schmidt, M., and Wood, F. Online learning rate adaptation with hypergradient descent. *arXiv preprint arXiv:1703.04782*, 2017.

Berge, C. *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Courier Corporation, 1997.

Bertsekas, D. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

Blalock, D. W., Ortiz, J. J. G., Frankle, J., and Guttag, J. V. What is the state of neural network pruning? In Dhillon, I. S., Papailiopoulos, D. S., and Sze, V. (eds.), *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org, 2020.

Boyd, S. P. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2014.

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990.

Delgado, M. F., Cernadas, E., Barro, S., and Amorim, D. G. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.*, 15(1): 3133–3181, 2014.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. Essentially no barriers in neural network energy landscape. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1308–1317. PMLR, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Ergen, T. and Pilanci, M. Global optimality beyond two layers: Training deep ReLU networks via convex programs. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2993–3003. PMLR, 2021a.

Ergen, T. and Pilanci, M. Implicit convex regularizers of CNN architectures: Convex optimization of two- and three-layer networks in polynomial time. In *International Conference on Learning Representations: ICLR 2021*, 2021b.

Ergen, T., Sahiner, A., Ozturkler, B., Pauly, J. M., Mardani, M., and Pilanci, M. Demystifying batch normalization in ReLU networks: Equivalent convex optimization models and implicit regularization. In *International Conference on Learning Representations*, 2021.

Fiacco, A. V. and Ishizuka, Y. Sensitivity and stability analysis for nonlinear programming. *Annals of Operations Research*, 27(1):215–235, 1990.

Fiat, J., Malach, E., and Shalev-Shwartz, S. Decoupling gating from linearity. *arXiv preprint arXiv:1906.05032*, 2019.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8803–8812, 2018.

Goldman, A. J. and Tucker, A. W. 4. theory of linear programming. In *Linear Inequalities and Related Systems.(AM-38), Volume 38*, pp. 53–98. Princeton University Press, 2016.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6151–6159, 2017.

Gupta, V., Bartan, B., Ergen, T., and Pilanci, M. Exact and relaxed convex formulations for shallow neural autoregressive models. In *International Conference on Acoustics, Speech, and Signal Processing*, 2021.

Hastie, T., Taylor, J., Tibshirani, R., and Walther, G. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

Hogan, W. W. Point-to-set maps in mathematical programming. *SIAM review*, 15(3):591–603, 1973.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuditipudi, R., Wang, X., Lee, H., Zhang, Y., Li, Z., Hu, W., Ge, R., and Arora, S. Explaining landscape connectivity of low-cost solutions for multilayer nets. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14574–14583, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Mishkin, A., Sahiner, A., and Pilanci, M. Fast convex optimization for two-layer ReLU networks: Equivalent model classes and cone decompositions. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15770–15816. PMLR, 2022.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

Nguyen, Q. On connected sublevel sets in deep learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4790–4799. PMLR, 2019.

Osborne, M. R., Presnell, B., and Turlach, B. A. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

Pilanci, M. and Ergen, T. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7695–7705. PMLR, 2020.

Robinson, S. M. and Day, R. H. A sufficient condition for continuity of optimal sets in mathematical programming. *Journal of Mathematical Analysis and Applications*, 45 (2):506–511, 1974.

Roth, V. and Fischer, B. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In Cohen, W. W., McCallum, A., and Roweis, S. T. (eds.), *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pp. 848–855. ACM, 2008.

Sahiner, A., Ergen, T., Pauly, J. M., and Pilanci, M. Vector-output ReLU neural network problems are copositive programs: Convex analysis of two layer networks and polynomial-time algorithms. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tibshirani, R. J. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371, 2011.

Tunçel, L. and Wolkowicz, H. Strong duality and minimal representations for cone optimization. *Computational optimization and applications*, 53(2):619–648, 2012.

Vaiter, S., Deledalle, C., Peyré, G., Fadili, J., and Dossal, C. The degrees of freedom of the group lasso for a general design. *CoRR*, abs/1212.6478, 2012.

Wang, Y., Lacotte, J., and Pilanci, M. The hidden convex optimization landscape of regularized two-layer ReLU networks: an exact characterization of optimal solutions. In *International Conference on Learning Representations*, 2021.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67, 2006.

A. Constrained Group Lasso: Proofs

A.1. Describing the Optimal Set

Lemma A.1. *The model fit is the same for all optimal solutions to the CGL problem. That is,*

$$Xw = Xw' \quad \textit{for all } w, w' \in \mathcal{W}^*(\lambda).$$

As a consequence, the sum of group norms is also constant when $\lambda > 0$.

Proof. This follows in a similar fashion to the classic result for the group lasso. Let $w, w' \in \mathcal{W}^*(\lambda)$, $\bar{w} = \frac{1}{2}w + \frac{1}{2}w'$, and suppose p^* is the optimal value of the constrained program. By convexity, we have

$$\frac{1}{2}\|X\bar{w} - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|\bar{w}_{b_i}\|_2 \leq \frac{1}{2}p^* + \frac{1}{2}p^* = p^*,$$

where the inequality is strict if $Xw \neq Xw'$ by strong convexity of $f(u) = \|u - y\|_2^2$. Since w, w' are both feasible, $\bar{w}$ is also feasible and clearly $\bar{w}$ cannot obtain an objective value less than p^*. Thus, $Xw = Xw'$ must hold.

To see the second part of the result, observe that

$$\lambda \sum_{b_i \in \mathcal{B}} \|\bar{w}_{b_i}\|_2 = p^* - \frac{1}{2}\|\hat{y}(\lambda) - y\|_2^2,$$

is also constant over $\mathcal{W}^*(\lambda)$. □

Proposition 3.1. *Fix $\lambda > 0$. The optimal set for the CGL problem is given by*

$$\begin{aligned} \mathcal{W}^*(\lambda) = \big\{ & w \in \mathbb{R}^d : \forall\, b_i \in \mathcal{S}_\lambda, w_{b_i} = \alpha_{b_i} v_{b_i}, \alpha_{b_i} \geq 0, \\ & \forall\, b_j \in \mathcal{B} \setminus \mathcal{S}_\lambda, w_{b_j} = 0,\ Xw = \hat{y} \big\} \end{aligned} \tag{11}$$

Proof. Fix $\lambda > 0$ and let $w \in \mathcal{W}^*(\lambda)$. If $w_{b_i} \neq 0$, then the KKT conditions require

$$\lambda \frac{w_{b_i}}{\|w_{b_i}\|_2} = v_{b_i} \implies w_{b_i} = \alpha_{b_i} v_{b_i},$$

for $\alpha > 0$. If $w_{b_i} = 0$, then $w_{b_i} = \alpha_{b_i} v_{b_i}$ holds trivially for $\alpha_{b_i} = 0$. Since w is optimal, it must satisfy

$$Xw = \hat{y},$$

by Lemma A.1. Finally, $\mathcal{A}_\lambda(w) \subseteq \mathcal{S}_\lambda$ so that w satisfies the characterization.

For the reverse direction, we start by defining

$$\mathcal{X} = \Bigg\{ w \in \mathbb{R}^d : \forall\, b_i \in \mathcal{S}_\lambda,\ w_{b_i} = \alpha_{b_i} v_{b_i}, \alpha_{b_i} \geq 0, \\ \forall\, b_j \in \mathcal{B} \setminus \mathcal{S}_\lambda, w_{b_j} = 0,\ Xw = \hat{y} \Bigg\}$$

Take any $w' \in \mathcal{X}$. If $w'_{b_i} \neq 0$, then

$$\begin{aligned} w'_{b_i} = \alpha v_{b_i} \implies & \frac{w'_{b_i}}{\|w'_{b_i}\|_2} = \frac{v_{b_i}}{\lambda} \\ \implies & \lambda \frac{w'_{b_i}}{\|w'_{b_i}\|_2} = v_{b_i}, \end{aligned}$$

where we have used the fact that $\|v_{b_i}\|_2 = \lambda$ for all $b_i \in \mathcal{E}_\lambda$. That is,

$$X_{b_i}^\top (Xw - y) + \lambda \frac{w'_{b_i}}{\|w'_{b_i}\|_2} + K_{b_i} \rho^*_{b_i} = 0,$$

which is exactly stationarity of the Lagrangian.

If $w'_{b_i} = 0$, then

$$Xw' = \hat{y} \implies \|X_{b_i}^\top (y - Xw') + K\rho^*_{b_i}\|_2 \leq \lambda,$$

which also implies the Lagrangian is stationary.

Now we show optimality of w' by checking feasibility and complementary slackness. If $w'_{b_i} \neq 0$ then $w'_{b_i} = \frac{\alpha'_{b_i}}{\alpha_{b_i}} w_{b_i}$ for some other optimal solution w with $\alpha_{b_i} > 0$. This follows since $w'_{b_i} \neq 0$ implies $b_i \in \mathcal{S}_\lambda$. Thus,

$$K_{b_i}^\top w'_{b_i} = \frac{\alpha'_{b_i}}{\alpha_{b_i}} w_{b_i} \leq 0,$$

by feasibility of w_{b_i}. Similarly, we find that complementary slackness is satisfied as follows:

$$[\rho_{b_i}]_j \cdot [K_{b_i}]_j^\top w'_{b_i} = \frac{\alpha'_{b_i}}{\alpha_{b_i}} [\rho_{b_i}]_j \cdot [K_{b_i}]_j^\top w_{b_i} = 0.$$

If $w_{b_i} = 0$, then both feasibility and complementary slackness are trivial. Since (w', ρ^*) are feasible and ρ^* is dual optimal, we conclude the KKT conditions are satisfied and thus $w' \in \mathcal{W}^*(\lambda)$. This completes the proof. $\square$

Proposition A.2. *Fix $\lambda > 0$. The optimal set for CGL problem is given by*

$$\begin{aligned} \mathcal{W}^*(\lambda) = \Big\{ w \in \mathbb{R}^d : &\forall\, b_i \in \mathcal{E}_\lambda,\, w_{b_i} = \alpha_{b_i} v_{b_i}, \alpha_{b_i} \geq 0, \\ &\forall\, b_j \in \mathcal{B} \setminus \mathcal{E}_\lambda, w_{b_j} = 0,\, Xw = \hat{y}, \\ &K^\top w \leq 0,\, \langle \rho^*, K^\top w \rangle = 0 \Big\} \end{aligned} \tag{20}$$

Proof. Fix $\lambda > 0$ and let $w \in \mathcal{W}^*(\lambda)$. If $w_{b_i} \neq 0$, then the KKT conditions require

$$\lambda \frac{w_{b_i}}{\|w_{b_i}\|_2} = v_{b_i} \implies w_{b_i} = \alpha_{b_i} v_{b_i},$$

for $\alpha > 0$. If $w_{b_i} = 0$, then $w_{b_i} = \alpha_{b_i} v_{b_i}$ holds trivially for $\alpha_{b_i} = 0$. Since w is optimal, it must satisfy

$$Xw = \hat{y},$$

by Lemma A.1. Finally, $\langle \rho^*, Kw \rangle = 0$ and is feasible by KKT conditions so that w satisfies the characterization.

For the reverse direction, we start by defining

$$\begin{aligned} \mathcal{X} = \Big\{ w \in \mathbb{R}^d : &\forall\, b_i \in \mathcal{E}_\lambda,\, w_{b_i} = \alpha_{b_i} v_{b_i}, \alpha_{b_i} \geq 0, \\ &\forall\, b_j \in \mathcal{B} \setminus \mathcal{E}_\lambda, w_{b_j} = 0,\, Xw = \hat{y}, \\ &K^\top w \leq 0,\, \langle \rho^*, K^\top w \rangle = 0 \Big\} \end{aligned}$$

Take any $w' \in \mathcal{X}$; If $w'_{b_i} \neq 0$, then

$$\begin{aligned} w'_{b_i} = \alpha v_{b_i} &\implies \frac{w'_{b_i}}{\|w'_{b_i}\|_2} = \frac{v_{b_i}}{\lambda} \\ &\implies \lambda \frac{w'_{b_i}}{\|w'_{b_i}\|_2} = v_{b_i}, \end{aligned}$$

where we have used the fact that $\|v_{b_i}\|_2 = \lambda$ for all $b_i \in \mathcal{E}_\lambda$. That is,

$$X_{b_i}^\top (Xw - y) + \lambda \frac{w'_{b_i}}{\|w'_{b_i}\|_2} + K_{b_i}\rho^*_{b_i} = 0,$$

which is exactly stationarity of the Lagrangian.

If $w'_{b_i} = 0$, then

$$Xw' = \hat{y} \implies \|X_{b_i}^\top (y - Xw') + K\rho^*_{b_i}\|_2 \le \lambda,$$

which also implies the Lagrangian is stationary.

Since $\langle \rho^*, K^\top w' \rangle = 0$ and $K^\top w \le 0$, i.e. w is feasible, it is clear that complementary slackness must also hold. We conclude that (w', ρ^*) are primal-dual optimal by the KKT conditions and the proof is complete. □

Lemma 3.3. *Fix $\lambda > 0$. The solution to CGL problem is unique if and only if $\{X_{b_i} v_{b_i}\}_{\mathcal{S}_\lambda}$ are linearly independent.*

Proof. Suppose by way of contradiction that $w, w' \in \mathcal{W}^*(\lambda)$ such that $w \neq w'$. By Proposition 3.1, we have

$$\begin{aligned} 0 = X(w - w') &= \sum_{b_i \in \mathcal{S}_\lambda} X_{b_i}(w_{b_i} - w'_{b_i}) \\ &= \sum_{b_i \in \mathcal{S}_\lambda} (\alpha_{b_i} - \alpha'_{b_i}) X_{b_i} v_{b_i}, \end{aligned}$$

which implies that the vectors $\{X_{b_i} v_{b_i}\}_{\mathcal{S}_\lambda}$ are linearly dependent.

Necessity follows from Algorithm 1, which shows that, given a solution $w \in \mathcal{W}^*(\lambda)$, linear dependence of $\{X_{b_i} w(\lambda) : b_i \in \mathcal{A}(w(\lambda))\}$ implies the exist of at least one additional solution. □

Proposition 3.5. *[Group General Position] Suppose for every $\mathcal{E} \subseteq \mathcal{B}$, $|\mathcal{E}| \le n+1$, there do not exist unit vectors $z_{b_i} \in \mathbb{R}^{|b_i|}$ such that for any $j \in \mathcal{E}$,*

$$X_{b_j} z_{b_j} \in \textit{affine}(\{X_{b_i} z_{b_i} : b_i \in \mathcal{E} \setminus b_j\}).$$

Then the group lasso solution is unique for all $\lambda > 0$.

Proof. Suppose the group Lasso solution is not unique. Then, Corollary 3.2 implies

$$\mathcal{N}_\lambda = \text{Null}(X_{\mathcal{E}_\lambda}) \bigcap \{z : z_{b_i} = \alpha_{b_i} c_{b_i}, b_i \in \mathcal{E}_\lambda\},$$

is non-empty. That is, there exist $\alpha_{b_i} \ge 0$ such that

$$\begin{aligned} X_{b_j} c_{b_j} &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i} X_{b_i} c_{b_i} \\ \implies X_{b_j} c_{b_j} &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i} X_{b_i} c_{b_i}. \end{aligned}$$

Taking inner-products on both sides with the residual r,

$$\begin{aligned} \implies \lambda^2 &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i} \lambda^2 \\ \implies 1 &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i}. \end{aligned}$$

Thus, we deduce that

$$X_{b_j} c_{b_j} = \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \beta_{b_i} X_{b_i} c_{b_i}, \tag{21}$$

where $\sum_{b_i \in \mathcal{E}_\lambda \setminus j} \beta_{b_i} = 1$. Now, suppose that $|\mathcal{E}_\lambda| > n + 1$. Then, $\{X_{b_i} c_{b_i} : b_i \in \mathcal{E}_\lambda \setminus j\}$ are linearly dependent and, by eliminating dependent vectors $X_{b_i} c_{b_i}$, we can repeat the above proof with a subset $\mathcal{E}'$ of at most $n+1$ blocks. Noting $\|c_{b_i}\|_2 = \lambda$ for each $b_i \in \mathcal{E}_\lambda$ and rescaling both sides of Equation (21) by λ implies the existence of unit vectors z_{b_i} which contradict GGP. This completes the proof. □

Proposition A.3. *Group general position does not imply the columns of X are in general position. Similarly, general position of the columns of X does not imply group general position.*

Proof. Consider the simple case where we have two groups: $b_1 = \{1\}$ and $b_2 = \{2, \ldots, d\}$. Group general position is violated if there exists a unit vector z_{b_2} such that

$$\begin{aligned} & x_1 = X_{b_2} z_{b_2}. \\ \iff & x_1 \in X_{b_2} B_{d-1}, \end{aligned}$$

where $B_{n-1} = \left\{ z \in \mathbb{R}^{d-1} : \|z\|_2 \leq 1 \right\}$. In contrast, general position is violated if

$$\begin{aligned} & x_1 \in \text{affine}(x_2, \ldots, x_d) \\ \iff & x_1 \in X \left\{ z : \langle z, 1 \rangle = 1 \right\}. \end{aligned}$$

Taking $X_{b_2} = I$, it is trivial to see that group general position can hold when general position is violated and vice-versa. □

A.2. Computing Dual Optimal Parameters

Lemma A.4. *One Lagrange dual of CGL is the following:*

$$\begin{aligned} \max_{\eta, \rho} \; & -\frac{1}{2}(\eta + K\rho - X^\top y)(X^\top X)^+(\eta + K\rho - X^\top y) + \frac{1}{2}\|y\|_2^2 \\ \textit{s.t.} \quad & \eta + K\rho \in \textit{Row}(X), \; \|\eta_{b_i}\|_2 \leq \lambda \; \forall \; b_i \in \mathcal{B}, \end{aligned} \tag{22}$$

where $\eta_{b_i} = c_{b_i} - K_{b_i}\rho_{b_i}$ shows that the vectors v_{b_i} are, in fact, dual variables. Moreover, if $K = 0$, then $\rho^* = 0$ and η has the unique solution $\eta_{b_i} = X_{b_i}^\top (y - Xw) = c_{b_i}$. That is, the dual parameters are the (unique) block correlation vectors.

Proof. We re-write the group Lasso problem as follows:

$$\min_w \frac{1}{2}\|Xz - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \quad \text{s.t. } z = w, \; K_{b_i}^\top z_{b_i} \leq 0.$$

The Lagrangian for this problem is

$$\begin{aligned} \mathcal{L}(w, z, \eta, \rho) &= \frac{1}{2}\|Xz - y\|_2^2 + \langle \eta, z - w \rangle + \left\langle \rho, K^\top z \right\rangle + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \\ &= \frac{1}{2}\|Xz - y\|_2^2 + \langle \eta + K\rho, z \rangle - \langle \eta, w \rangle + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2. \end{aligned}$$

Minimizing over z, we find that stationarity implies

$$X^\top (y - Xz) = \eta + K\rho,$$

so that $\eta + K\rho \in \text{Row}(X)$. Solving this system, we find

$$X^\top X z = X^\top y - \eta - K\rho \implies z = (X^\top X)^+ \left[X^\top y - \eta - K\rho \right] + c,$$

where $c \in \text{Null}(X)$. Let us minimize over w similarly. The Lagrangian decouples block-wise in w, so that we must solve

$$\min_{w_{b_i}} \lambda \|w_{b_i}\|_2 - \left\langle \eta_{b_i}, w_{b_i} \right\rangle,$$

for each $b_i \in \mathcal{B}$. The minimum value is achieved by the (negative) Fenchel conjugate of $\lambda \|w_{b_i}\|_2$ evaluated at η_{b_i}; that is,

$$\min_{w_{b_i}} \lambda \|w_{b_i}\|_2 - \langle \eta_{b_i}, w_{b_i} \rangle = -\mathbb{1}(\|\eta_{b_i}\|_2 \leq \lambda).$$

Combining this with the expression for z, we obtain

$$\begin{aligned}\mathcal{L}(c, \eta, \rho) &= -\frac{1}{2}(\eta + K\rho - X^\top y)(X^\top X)^+(\eta + K\rho - X^\top y) + \langle \eta + K\rho, c \rangle - \sum_{b_i \in \mathcal{B}} \mathbb{1}(\|\eta_{b_i}\|_2 \leq \lambda) \\ &= -\frac{1}{2}(\eta + K\rho - X^\top y)(X^\top X)^+(\eta + K\rho - X^\top y) - \sum_{b_i \in \mathcal{B}} \mathbb{1}(\|\eta_{b_i}\|_2 \leq \lambda),\end{aligned}$$

where the second equality follows since $\eta + K\rho \in \text{Row}(X)$ and $c \in \text{Null}(X)$ are orthogonal. (Alternatively, one can observe that the dual problem is unbounded below whenever $\langle c, \eta + K\rho \rangle \neq 0$.) Thus, the dual problem is equal to

$$\max_{\eta, \rho} -\frac{1}{2}(\eta + K\rho - X^\top y)(X^\top X)^+(\eta_K\rho - X^\top y) - \sum_{b_i \in \mathcal{B}} \mathbb{1}(\|\eta_{b_i}\|_2 \leq \lambda) - \mathbb{1}(\eta + K\rho \in \text{Row}(X)),$$

which completes the derivation.

Recalling $z = w$ for any primal-dual optimal pair and

$$X^\top (y - Xz) = \eta + K\rho,$$

shows that $\eta_{b_i} = c_{b_i} - K\rho$ as claimed. Moreover, if $K = 0$, then we may assume without loss of generality that he corresponding dual vectors ρ_{b_i} are zero. In this case, $\eta_{b_i} = c_{b_i}$ and, since c_{b_i} is unique, the dual solution must also be unique. □

Proposition A.5. *Let $\lambda > 0$ and $w \in \mathcal{W}^*(\lambda)$. If $w_{b_i} = 0$, then any solution to Equation (15) is dual optimal for block b_i.*

Proof. Let ρ_{b_i} be a solution to Equation (15). Since w is optimal and strong duality holds, there exists some min-norm dual optimal vector ρ^*. Moreover ρ^* satisfies $\rho^*_{b_i} \geq 0$ and

$$\|K_{b_i}\rho_{b_i} - c_{b_i}\|_2^2 \leq \|K_{b_i}\rho^*_{b_i} - c_{b_i}\|_2^2 \leq \lambda^2,$$

so w is both feasible satisfies stationarity of the Lagrangian, Finally, because $w_{b_i} = 0$, complementary slackness,

$$[\rho_{b_i}]_j \cdot [K_{b_i}]_j^\top w_{b_i} = 0,$$

is verified for every $j \in [a_{b_i}]$. Since the KKT conditions are sufficient for primal-dual optimality, we conclude that ρ_{b_i} is dual optimal. This completes the proof. □

A.3. Minimal Solutions and Optimal Pruning

Proposition 3.6. *For $\lambda > 0$, $w \in \mathcal{W}^*(\lambda)$ is minimal if and only if the vectors $\{X_{b_i} w_{b_i}\}_{\mathcal{A}(w)}$ are linearly independent.*

Proof. Let $w \in \mathcal{W}^*(\lambda)$ and assume that the vectors $\{X_{b_i} w_{b_i}\}_{\mathcal{A}(w)}$ are linearly independent. By way of contradiction, assume there exists $w' \in \mathcal{W}^*(\lambda)$ with strictly smaller support. By Proposition 3.1, we have

$$w'_{b_i} = \beta_{b_i} w_{b_i},$$

for some $\beta_{b_i} \geq 0$. This holds for each $b_i \in \mathcal{A}_\lambda(w)$ (with $\beta_{b_i} = 0$ when $b_i \in \mathcal{A}_\lambda(w) \setminus \mathcal{A}_\lambda(w')$) so that

$$Xw = Xw' \implies \sum_{b_i \in \mathcal{A}_\lambda(w)} (1 - \beta_{b_i}) X_{b_i} w_{b_i} = 0,$$

which is a contradiction.

For the reverse direction, assume that w is minimal, but that $\{X_{b_i} w_{b_i}\}_{\mathcal{A}(w)}$ are not linearly independent. Then the correctness of Algorithm 1 (see Proposition A.6) implies w is not minimal. □

Proposition A.6. *Algorithm 1 returns a minimal solution to the constrained group lasso problem in at most $O(n^3 l + nd)$ time, where l is the number of non-zero groups in the initial solution.*

Proof. **Correctness**: Let $w \in \mathcal{W}^*(\lambda)$ and $\mathcal{A}$ be the associated active set. If $w^0 = w$ is minimal, then Proposition 3.6 implies $\{X_{b_i} w_{b_i}\}_{\mathcal{A}}$ are linearly independent the algorithm returns a minimal solution.

Let $k \geq 0$ and suppose w^k is not minimal. Then there exist weights β_{b_i} such that

$$\sum_{b_i \in \mathcal{A}} \beta_{b_i} X_{b_i} w^k = 0.$$

Let $w^t_{b_i} = (1 - t\beta_{b_i}) w_{b_i}$ and let t^k be as defined in the algorithm. By construction, t^k is the smallest magnitude t such that $(1 - t\beta_{b_i}) = 0$ for some $b_i \in \mathcal{A}$. We assume without loss of generality that $t^k > 0$.

Fix $0 < t < t^k$. Let's show that w^t is a solution to the constrained problem. Firstly, we have

$$Xw^t = Xw^k - t \sum_{b_i \in \mathcal{A}} \beta_{b_i} X_{b_i} w^k_{b_i} = Xw,$$

showing that the model fit preserved. Moreover,

$$w^t_{b_i} = (1 - t\beta_{b_i}) w^k_{b_i} = (1 - t\beta_{b_i}) \alpha_{b_i} v_{b_i},$$

where $(1 - t\beta_{b_i})\alpha_{b_i} > 0$ by the choice of t. We conclude that w^t is optimal by Proposition 3.1.

By construction,

$$\lim_{t \uparrow t^k} w^t = w^{k+1}.$$

Since w^t is an optimal solution, it has the (unique) optimal squared error and sum of group norms. Taking limits as $t \uparrow t^k$, we see that

$$Xw^{k+1} = Xw^k, \quad \sum_{b_i \in \mathcal{B}} \|w^{k+1}_{b_i}\|_2 = \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2, \quad K^\top w^{k+1} \leq 0,$$

which implies that w^{k+1} obtains the optimal objective value and is feasible. Thus w^{k+1} is also a solution. Finally, $\mathcal{A}(w^{k+1})$ is strictly smaller than $\mathcal{A}$, as required.

Arguing by induction now implies that Algorithm 1 returns an minimal solution in a finite number of steps.

Complexity: First, observe that we can pre-compute the block-wise model fits $q_{b_i} = X_{b_i} w_{b_i}$ before running the algorithm. The complexity is at most $O(nd)$. At iteration iteration of the algorithm, we must do two things: (i) compute a non-trivial solution to a homogeneous equation and (ii) update the weights of the model fits. For (i), it is clear that any set of $n + 1$ q_{b_i} vectors will be linearly dependent, so that we compute a non-trivial solution to the homogeneous equation using the SVD in an at most $O(n^3)$ operations. For (ii), updating at most n of the β_{b_i}'s requires $O(n)$ time. Since the algorithm runs at most l iterations, we obtain a final complexity o f $O(n^3 l + nd)$, as claimed. □

Lemma A.7. *Each step of the Algorithm 1 preserves the span of $\left\{X_{b_i} w^k_{b_i}\right\}$. That is, only linearly dependent vectors are removed.*

Proof. Let $\mathcal{V} = \text{Span}\left\{X_{b_i} w_{b_i}\right\}$ be the span of the initial solution. Since $w^0 = w$ by definition, the base case holds trivially.

Suppose $\text{Span}(X_{b_i} w_{b_i}^k) = \mathcal{V}$. Let $v \in \mathcal{V}$ and observe that

$$\begin{aligned} v &= \sum_{b_i \in \mathcal{A}_\lambda(w^k)} \alpha_{b_i} X_{b_i} w_{b_i}^k \\ &= \sum_{b_i \in \mathcal{A}_\lambda(w^{k+1})} \left(\frac{\alpha_{b_i}}{1 - t^k \beta_{b_i}} \right) X_{b_i} w_{b_i}^{k+1} + \sum_{b_i \in \mathcal{A}_\lambda(w^k) \setminus \mathcal{A}_\lambda(w^{k+1})} \alpha_{b_i} X_{b_i} w_{b_i}^k \\ &= \sum_{b_i \in \mathcal{A}_\lambda(w^{k+1})} \left(\frac{\alpha_{b_i}}{1 - t^k \beta_{b_i}} \right) X_{b_i} w_{b_i}^{k+1} + \frac{1}{\beta_{{b_i}^k}} \sum_{b_i \in \mathcal{A}_\lambda(w^{k+1})} \alpha_{b_i} \beta_{b_i} X_{b_i} w_{b_i}^{k+1} \\ &= \sum_{b_i \in \mathcal{A}_\lambda(w^{k+1})} \alpha_{b_i} \left(\frac{1}{1 - t^k \beta_{b_i}} + \frac{\beta_{b_i}}{\beta_{{b_i}^k}} \right) X_{b_i} w_{b_i}^{k+1}, \end{aligned}$$

where we have used the fact that $\beta_{b_i} = \beta_{{b_i}^k}$ for every block that is pruned at iteration k. Thus, $\text{Span}(X_{b_i} w_{b_i}^{k+1}) = \mathcal{V}$. Arguing by induction completes the proof. $\square$

Lemma A.8. *Let $\mathcal{X} = \{x_1, \ldots, x_k\}$ be a set of linearly dependent vectors. Every linearly independent subset of $\mathcal{X}$ obtained by iteratively removing linearly dependent vectors has the same cardinality.*

Proof. Let $\mathcal{Y}, \mathcal{Y}' \subset \mathcal{X}$ be linearly independent subsets obtained by pruning linearly dependent vectors from $\mathcal{X}$ and assume $|\mathcal{Y}'| < |\mathcal{Y}|$. Since $\mathcal{Y}$ and $\mathcal{Y}'$ are obtained by pruning only linearly dependent vectors, it must be that

$$\text{Span}\left(\mathcal{Y}'\right) = \text{Span}\left(\mathcal{Y}\right) = \text{Span}\left(\mathcal{X}\right).$$

Let $c = \dim\left(\text{Span}\left(\mathcal{X}\right)\right)$. Only a set of c linearly independent vectors can span $\text{Span}\left(\mathcal{X}\right)$; thus, $|\mathcal{Y}| = c$ must hold and $\mathcal{Y}'$ cannot span $\text{Span}\left(\mathcal{X}\right)$. This is a contradiction. We conclude $|\mathcal{Y}'| = |\mathcal{Y}|$ as claimed. $\square$

Lemma A.9. *There exists a solution to CGL with support exactly $\mathcal{S}_\lambda$.*

Proof. By definition, there exists $w \in \mathcal{W}^*(\lambda)$ such that $w_{b_i} \neq 0$ for every $b_i \in \mathcal{S}_\lambda$. Taking convex combination of these solutions yields w' with support exactly $\mathcal{S}_\lambda$. Since $\mathcal{W}^*(\lambda)$ is convex, w' is also a solution. This completes the proof. $\square$

Lemma A.10. *Let w be a minimal solution and $\bar{w}$ be a solution with support $\mathcal{S}_\lambda$, which exists by Lemma A.9. Then, $\bar{w}$ can be pruned step to obtain w.*

Proof. Suppose $\bar{w}$ is minimal. Then $\{X_{b_i} \bar{w}_{b_i}\}_{\mathcal{S}_\lambda}$ are linearly independent, which implies $\{X_{b_i} v_{b_i}\}_{\mathcal{S}_\lambda}$ are also linearly independent. We conclude $\bar{w}$ is the unique solution to CGL by Lemma 3.3 and the claim holds trivially.

Suppose $\bar{w}$ is not minimal. Since $\mathcal{A}_\lambda(w) \subset \mathcal{S}_\lambda$, $\text{Span}\left\{X_{b_i} w_{b_i}\right\} \subset \text{Span}\left\{X_{b_i} \bar{w}_{b_i}\right\}$ so that every vector $X_{b_i} w_{b_i}$ can be written as a linear combination of vectors in $\{X_{b_i} \bar{w}_{b_i}\}$. Thus, we find that

$$\begin{aligned} \sum_{b_i \in \mathcal{A}_\lambda(\bar{w})} X_{b_i} \bar{w}_{b_i} &= \hat{y} = \sum_{b_i \in \mathcal{A}_\lambda(w)} X_{b_i} w_{b_i} \\ \implies \sum_{b_i \in \mathcal{S}_\lambda \setminus \mathcal{A}_\lambda(w)} X_{b_i} \bar{w}_{b_i} + \sum_{b_i \in \mathcal{A}_\lambda(w)} \beta_{b_i} X_{b_i} \bar{w}_{b_i} &= 0, \end{aligned}$$

where $\beta_{b_i} = 1 - \frac{\alpha_{b_i}}{\bar{\alpha}_{b_i}}$ for $w_{b_i} = \alpha_{b_i} v_{b_i}$, $\bar{w}_{b_i} = \bar{\alpha}_{b_i} v_{b_i}$. Thus, $\{X_{b_i} \bar{w}_{b_i}\}$ are linearly dependent and it is possible to prune the solution.

Now we show that we can, in fact, prune all vectors in $\mathcal{S}_\lambda \setminus \mathcal{A}_\lambda(w)$ in one pruning step. First, observe that $\frac{\alpha_{b_i}}{\bar{\alpha}_{b_i}} > 0$ so that $\beta_{b_i} < 1$ for every $b_i \in \mathcal{A}_\lambda(w)$. Following the proof of Proposition A.6, define

$$w_{b_i}^t = (1 - t\beta_{b_i}) \bar{w}_{b_i} = (1 - t\beta_{b_i}) \bar{\alpha}_{b_i} v_{b_i},$$

for $0 < t < 1$. Since $\beta_{b_i} < 1$ for every $b_i \in \mathcal{S}_\lambda$, it is straightforward to deduce that $w_{b_i}^t$ is optimal by Proposition 3.1. Arguing as in Proposition A.6, we can show w^1 is also optimal. It remains only to notice that $w_{b_i}^1 = 0$ for $b_i \in \mathcal{S}_\lambda \setminus \mathcal{A}_\lambda(w)$ and $w_{b_i}^1 = w_{b_i}$ for $b_i \in \mathcal{A}_\lambda(w)$. Thus, the pruning algorithm can move from $\bar{w}$ to w in one step. $\square$

Proposition 3.8. *Let $\mathcal{V}$=Span($\{X_{b_i} \bar{w_{b_i}}\}$) for $\bar{w} \in \mathcal{W}^*(\lambda)$. Every minimal solution has $c = dim(\mathcal{V})$ active blocks.*

Proof. Suppose w and w' are two minimal solutions. Both can be obtained by pruning the maximal solution with support $\mathcal{S}_\lambda$ (Lemma A.10). Thus, w and w' both span Span($\{X_{b_i} w_{b_i}\}$) by Lemma A.7. Lemma A.8 now implies w and w' have the same number of active blocks. This number must be c, otherwise there would be a linearly dependent vector in $\{X_{b_i} w_{b_i}\}_{(} \mathcal{A}_\lambda(w))$ and w would not be minimal. This completes the proof. ☐

A.4. Continuity of the Solution Path

Lemma A.11. *Every solution to the constrained group lasso problem is bounded by an absolute constant independent of λ. Specifically, every $w \in \mathcal{W}^*(\lambda)$ satisfies*

$$\sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \leq \sum_{b_i \in \mathcal{B}} \|\bar{w}_{b_i}\|_2.$$

where $\bar{w} \in \mathcal{W}^*(0)$ is the least-squares solution with minimum ℓ_2-norm.

Proof. Let $h(w) = \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2$, and define W_g to be the set of least squares solutions with minimum group norm. That is,

$$W_g = \arg\min \{h(w) : w \in \mathcal{W}^*(0)\}.$$

Let $w_g \in W_g$, and suppose that $h(w_g) < h(w(\lambda))$ for some $\lambda > 0$, $w \in \mathcal{W}^*(\lambda)$. Since

$$\frac{1}{2}\|Xw_g - y\|_2^2 \leq \frac{1}{2}\|Xw(\lambda) - y\|_2^2$$

we deduce

$$\frac{1}{2}\|Xw_g - y\|_2^2 + \lambda h(w_g) < \frac{1}{2}\|Xw(\lambda) - y\|_2^2 + \lambda h(w(\lambda)),$$

which is a contradiction. So $h(w(\lambda)) \leq h(w_g)$ for all $\lambda > 0$. Observing $h(w_g) \leq h(\bar{w})$ since $\bar{w}$ may not be in W_g gives the result. Since $h(\bar{w})$ is independent of λ, we conclude that $\mathcal{W}^*(\lambda)$ is bounded independent of λ. ☐

Proposition 3.12. *$\lambda \mapsto p^*(\lambda)$ is continuous for all $\lambda \geq 0$.*

Proof. Define the joint objective function

$$f(w, \lambda) = \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2.$$

Clearly $f(w, \lambda)$ is jointly continuous in w and λ. By Lemma A.11, minimization of $f(w, \lambda)$ subject to $K_{b_i}^\top w_{b_i} \leq 0$ is equivalent to the constrained minimization problem,

$$p^*(\lambda) = \min_w f(w, \lambda) \quad \text{s.t.} \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \leq C,\ K_{b_i}^\top w_{b_i} \leq 0,$$

where C is a finite absolute constant. Note that this expression is also valid when $\lambda = 0$ as the min-norm solution to the unregularized least squares problem obeys the constraint.

Thus, $\mathcal{W}^*$ is a continuous optimization problem over a continuous (constant in this case) compact constraint set and the classical result of Berge (1997) (see also Hogan (1973)[Theorem 7]) implies p^* is continuous. ☐

Proposition 3.13. *While $\mathcal{W}^*$ is closed on $\mathbb{R}_+$, it is open 0 if only if X is full column rank. However, if the solution is unique on $\Lambda \subset \mathbb{R}^+$, then $\mathcal{W}^*$ is open at every $\lambda \in \Lambda$.*

Proof. Joint continuity of the objective

$$f(w, \lambda) = \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2,$$

combined with continuity of the (constant) constraint allows us to use Robinson & Day (1974, Theorem 1) to obtain that that $\mathcal{W}^*$ is upper semi-continuous. Since $\mathcal{W}^*$ is convex and bounded, it is compact. It is thus also uniformly compact and lower semi-continuity is equivalent to closedness (Hogan, 1973)[Theorem 3]. We conclude that $\mathcal{W}^*$ is closed as claimed.

If X is full column rank, then the constrained group lasso solution is unique for all $\lambda \geq 0$. The solution map is a singleton on $\mathbb{R}^+$, and closedness and openness are equivalent properties for singleton maps. Since we have already shown it is closed, the solution map must also be open. An identical argument shows that the solution map is open at on any interval over which the solution is unique.

Now we show the reverse implication by proving the contrapositive. Assume X is not full column-rank and suppose $K_{b_i} = 0$ for each $b_i \in \mathcal{B}$. The solution map at $\lambda = 0$ is the solution set to the least squares problem,

$$\min_w \frac{1}{2}\|Xw - y\|_2^2,$$

which is known to be $\mathcal{W}^*(0) = \{w^*(0) + z : z \in \text{Null}(X)\}$. While $\mathcal{W}^*(0)$ is unbounded, it holds that $\mathcal{W}^*(\lambda) \subset C$ for some bounded C for every $\lambda > 0$ (Lemma A.11). As a result, there exist uncountably many solutions in $\mathcal{W}^*(0)$ which are not limit points of solutions in $\mathcal{W}^*(\lambda_k)$ as $\lambda_k \to 0$. In other words, $\mathcal{W}^*(0)$ is not open at 0. □

Proposition 3.15. *If $K = 0$, then $\hat{y}(\lambda)$ is continuous on $\mathbb{R}_+$ and the penalty $\sum_{b_i \in \mathcal{B}} \|w_{b_i}(\lambda)\|_2$ is continuous for $\lambda > 0$.*

Proof. Consider the dual problem from Lemma A.4,

$$\begin{aligned} \max_\eta \; & -\frac{1}{2}(\eta - X^\top y)(X^\top X)^+(\eta - X^\top y) + \frac{1}{2}\|y\|_2^2 \\ \text{s.t.} \; & \eta \in \text{Row}(X), \; \|\eta_{b_i}\|_2 \leq \lambda \; \forall \; b_i \in \mathcal{B}. \end{aligned}$$

The objective function is a convex quadratic and continuous in η. The constraint set is

$$\mathcal{C}(\lambda) = \{\eta : \eta \in \text{Row}(X), : \|\eta_{b_i}\|_2 \leq \lambda \; \forall \; b_i \in \mathcal{B}\}.$$

Let's show that $\mathcal{C}$ is continuous.

Let $\lambda_k \geq 0$, $\lambda_k \to \bar{\lambda}$ and $\eta_k \in \mathcal{C}(\lambda_k)$ such that $\eta_k \to \bar{\eta}$. Since $\eta_k \in \text{Row}(X)$, $\bar{\eta} \in \text{Row}(X)$. Moreover,

$$\|[\eta_k]_{b_i}\|_2 \leq \lambda_k,$$

so that taking limits on both sides implies $\|\bar{\eta}_k\|_2 \leq \bar{\lambda}$. Thus, $\bar{\eta} \in \mathcal{C}(\bar{\lambda})$, showing that $\mathcal{C}$ is closed.

Hogan (1973, Theorem 12) states that $\mathcal{C}(\lambda)$ is open at $\bar{\lambda}$ if for each $b_i \in \mathcal{B}$, $g_{b_i}(\lambda, \eta) = \|\eta_{b_i}\|_2 - \lambda$ is continuous on $\bar{\lambda} \times \mathcal{C}(\lambda)$, convex in η, and for fixed λ, and there exists $\bar{\eta}$ such that such that $g(\lambda, \bar{\eta}) < 0$.

Let us check these conditions. First, observe that g_{b_i} is continuous and convex in η for any choice of λ. Taking $\bar{\eta} = 0$, we find

$$g_{b_i}(\lambda, 0) = -\lambda < 0,$$

as long as $\lambda > 0$. Thus, $\mathcal{C}(\lambda)$ is open at each $\lambda > 0$.

At $\lambda = 0$, $\mathcal{C}(\lambda) = \{0\}$. Since $\mathcal{C}$ is closed everywhere, we conclude it is open at $\lambda = 0$. Putting these results together proves that $\mathcal{C}$ is continuous.

Recall that the dual solution satisfies $\eta^*_{b_i} = c_{b_i}$, so that it is always unique. Combining this fact with Robinson & Day (1974, Theorem 1) implies that $\lambda \mapsto \eta^*(\lambda)$ is a continuous function. Since we have

$$\eta^*(\lambda) = c(\lambda) = X^\top(y - Xw^*(\lambda)),$$

it must be that the model fit $\hat{y}(\lambda) = Xw^*(\lambda)$ is continuous as well (any discontinuities must be in Null(X), but $\hat{y}$ is orthogonal to Null(X)).

Finally, using Proposition 3.12, we have

$$p^*(\lambda) - \frac{1}{2}\|\hat{y}(\lambda) - y\|_2^2 = \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2,$$

is a sum of continuous functions and thus continuous. Writing

$$g(\lambda) = \sum_{b_i \in \mathcal{B}} \|b_i\|_2 = \left[\lambda \sum_{b_i \in \mathcal{B}} \|b_i\|_2\right] \left[\frac{1}{\lambda}\right],$$

as the product of continuous functions shows $g(\lambda)$ is continuous at every $\lambda > 0$. $\square$

A.5. The Min-Norm Path

Proposition A.12. *Consider the min group-norm interpolation problem,*

$$\min_{w} \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \quad \text{s.t. } Xw = y.$$

There exist X, y such that the minimum ℓ_2 norm solution to this problem is not in Row$(X_{\mathcal{A}^*_\lambda})$.

Proof. We provide a counter-example where the solution is not the row space of the active set. Consider the problem given by

$$X = \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 1 & 0 & 2 \end{array}\right], \quad y = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where the vertical line indicates the block structure, i.e., $b_1 = \{1, 2\}$ and $b_2 = \{3\}$.

Clearly a solution using only b_2 cannot interpolate the data, so the active set must be $\{b_1, b_2\}$ or $\{b_1\}$. If the active set is b_1, then the minimum norm interpolating solution can only be $w = [1\ 0\ 0]^\top$, which has group norm 1.

Now, consider when the active set is $\{b_1, b_2\}$. The interpolating solution in Row(X) satisfies the following system

$$\begin{aligned} \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 1 & 0 & 2 \end{array}\right] \left(\alpha \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ \implies \begin{bmatrix} 5\alpha + \beta \\ \alpha + 5\beta \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \end{aligned}$$

Solving for α and β yields $\alpha = 1 - 5\beta$ and $24\beta = 4$, which implies $\beta = 1/6$ and $\alpha = 1/6$. The optimal w^* is thus

$$w^* = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix},$$

and group norm of w^* is

$$\sum_{b_i \in \mathcal{B}} \|w^*_{b_i}\|_2 = \sqrt{2/9} + 1/3 = (1 + \sqrt{2})/3.$$

Now let's see if we can reduce the norm by including directions in Null(X). The Null space is orthogonal to both rows of X, from which we conclude

$$\text{Null}(X) = \left\{ \gamma z : z = \begin{bmatrix} -2/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \gamma \in \mathbb{R} \right\}.$$

Any vector $w' = w^* + \gamma z$ is an interpolating solution, so it only remains to check if there is a choice of γ that decreases the group norm. Assuming $\gamma > -1$,

$$\begin{aligned}
&\sum_{b_i \in \mathcal{B}} \|w^*_{b_i}\|_2 \leq \sum_{b_i \in \mathcal{B}} \|w'_{b_i}\|_2 \\
\iff & \frac{1+\sqrt{2}}{3} = \sqrt{(\frac{1}{3} - \frac{2\gamma}{3})^2 + (\frac{1}{3} + \frac{\gamma}{3})^2} + \left|\frac{1+\gamma}{3}\right| \\
\iff & \frac{1+\sqrt{2}}{3} - \frac{1+\gamma}{3} \leq \sqrt{(\frac{1}{3} - \frac{2\gamma}{3})^2 + (\frac{1}{3} + \frac{\gamma}{3})^2} && \text{(since } \gamma > -1\text{)} \\
\iff & \left(\frac{1+\sqrt{2}}{3} - \frac{1+\gamma}{3}\right)^2 \leq (\frac{1}{3} - \frac{2\gamma}{3})^2 + (\frac{1}{3} + \frac{\gamma}{3})^2 \\
\iff & \left(\sqrt{2} - \gamma\right)^2 \leq (1-2\gamma)^2 + (1+\gamma)^2
\end{aligned}$$

The left-hand side satisfies

$$\left(\sqrt{2} - \gamma\right)^2 = 2 - 2\sqrt{2}\gamma + \gamma^2,$$

while the right-hand side is

$$(1-2\gamma)^2 + (1+\gamma)^2 = 1 - 4\gamma + 4\gamma^2 + 1 + 2\gamma + \gamma^2 = 2 - 2\gamma + 5\gamma^2.$$

As a result,

$$\begin{aligned}
&\sum_{b_i \in \mathcal{B}} \|w^*_{b_i}\|_2 \leq \sum_{b_i \in \mathcal{B}} \|w'_{b_i}\|_2 \\
\iff & 2 - 2\gamma + 5\gamma^2 - 2 - \gamma^2 + 2\sqrt{2}\gamma \geq 0 \\
\iff & 2(\sqrt{2}-1)\gamma + 4\gamma^2 \geq 0.
\end{aligned}$$

However, it is easy to check that this fails for $\gamma \in ((1-\sqrt{2})/2, 0)$. So the minimum ℓ_2-norm interpolating solution is not in $\text{Row}(X_{\mathcal{A}^*_\lambda})$. ☐

Lemma A.13. *Let W_g be the set of least squares solution with minimum group norm. That is,*

$$W_g = \arg\min \left\{ \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 : X^\top X w = X^\top y \right\}.$$

Then every limit point of the min-norm group lasso solution lies in W_g.

Proof. Let $\lambda_k \to 0$ and observe that $w^*(\lambda_k)$ has at least one limit point since it is bounded (Lemma A.11). Since $\|c_{b_i}(\lambda_k)\|_2 \leq \lambda_k$, we see that $\lim_k \|c_{b_i}(\lambda_k)\|_2 = 0$ and thus $\lim_k c_{b_i}(\lambda_k) = 0$.

FO optimality conditions imply

$$(X^\top X) w^*(\lambda) = X^\top y - c(\lambda), \tag{23}$$

which, taking limits on both sides, gives

$$\lim_k (X^\top X) w^*(\lambda_k) = X^\top y. \tag{24}$$

That is, every limit point $\bar{w}$ of $w^*(\lambda_k)$ is a least squares solution satisfying $h(\bar{w}) \leq h(w_g)$. We conclude that $\bar{w} \in W_g$ as claimed. ☐

Proposition 3.16. *Suppose $K_{b_i} = 0$. There exists (X, y) and $\lambda > 0$ such that $w^*_{\mathcal{A}^*_\lambda}(\lambda) \notin$ Row$(X_{\mathcal{A}^*_\lambda})$.*

Proof. Consider the setting of Proposition A.12, with

$$X = \left[\begin{array}{cc|c} 1 & 2 & 0 \\ 1 & 0 & 2 \end{array}\right], \quad y = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where the vertical line indicates the block structure, i.e., $b_1 = \{1, 2\}$ and $b_2 = \{3\}$. We have shown that the min group norm interpolant is unique, is supported on b_1 and b_2, and does not lie in $Row(X)$. Let w_g be this solution.

Let $\lambda_k \downarrow 0$. By Lemma A.13, every limit point of $w^*(\lambda_k) = w_g$. Thus, $\lim_k w^*(\lambda_k)$ exists and is exactly w_g. Moreover, $w^*(\lambda_k)$ must be supported on b_1 and b_2 for all k sufficiently large.

Decomposing $w_g = a + b$ and $w^*(\lambda_k) = r_k + n_k$ where $a, r_k \in \text{Row}(X)$ and $b, n_k \in \text{Null}(X)$, we see that

$$\|w_g - w^*(\lambda_k)\|_2^2 = \|a - r_k\|_2^2 + \|b - n_k\|_2^2 \to 0,$$

implying that $n_k \neq 0$ for sufficiently large k. In other words, the min-norm solution to the group lasso problem fails to fall in $\text{Row}(X)$ for some $\lambda > 0$. □

Proposition 3.17. *Let $\lambda > 0$ and consider the program:*

$$\alpha^* = \arg\min_{\alpha \geq 0} \|\alpha\|_2^2 \;\; \textit{s.t.} \sum_{b_i \in \mathcal{S}_\lambda} \alpha_{b_i} X_{b_i} v_{b_i} = \hat{y}. \tag{16}$$

Then the min-norm solution is given by $w^*_{b_i} = \alpha^*_{b_i} v_{b_i}$.

Proof. Let $w \in \mathcal{W}^*(\lambda)$. By Proposition 3.1, $w_{b_i} = \alpha_{b_i} v_{b_i}$ where $\alpha_{b_i} \geq 0$. Moreover, $\alpha_{b_i} = 0$ for every $b_i \in \mathcal{B} \setminus \mathcal{S}_\lambda$. As a result,

$$\begin{aligned} \|w\|_2^2 &= \|w_{\mathcal{E}_\lambda}\|_2^2 \\ &= \sum_{b_i \in \mathcal{S}_\lambda} \|\alpha_{b_i} v_{b_i}\|_2^2 \\ &= \lambda \|\alpha\|_2^2, \end{aligned}$$

where the last equality follows from $b_i \in \mathcal{S}_\lambda \implies b_i \in \mathcal{E}_\lambda$, which implies $\|v_{b_i}\|_2 = \lambda$.

Now suppose α^* is optimal for the cone program and let $w \in \mathbb{R}^d$ such that $w_{b_i} = \alpha^*_{b_i} v_{b_i}$. By construction, $\alpha^*_{b_i} = 0$ for all $b_i \in \mathcal{B} \setminus \mathcal{S}_\lambda$ (or it could not be optimal) so that $w_{b_i} = 0$ for all $b_i \in \mathcal{B} \setminus \mathcal{S}_\lambda$. Moreover,

$$Xw = \sum_{b_i \in \mathcal{S}_\lambda} \alpha^*_{b_i} X_{b_i} v_{b_i} = \hat{y},$$

Thus, w solves

$$\begin{aligned} \arg\min_w \|w\|_2^2 \;\; \text{s.t.} \; &\forall\, b_i \in \mathcal{S}_\lambda,\; w_{b_i} = \alpha_{b_i} v_{b_i}, \alpha_{b_i} \geq 0, \\ &\forall\, b_j \in \mathcal{B} \setminus \mathcal{S}_\lambda, w_{b_j} = 0,\; Xw = \hat{y}. \end{aligned}$$

Invoking Proposition 3.1 now proves that w is the min-norm solution. □

Lemma A.14. *The ℓ_2-penalized group lasso problem in Equation (17) is equivalent to the following group lasso problem:*

$$\begin{aligned} \min_w \frac{1}{2}\|\tilde{X}w - \tilde{y}\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \\ \textit{s.t.} \quad K_{b_i}^\top w_{b_i} \leq 0 \textit{ for all } b_i \in \mathcal{B}. \end{aligned} \tag{25}$$

where we have defined the extended data matrix and targets

$$\tilde{X} = \begin{bmatrix} X \\ \sqrt{\delta} I \end{bmatrix} \qquad \tilde{y} = \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

Moreover, $\tilde{X}$ is full column-rank and thus the group lasso solution is unique.

Proof. It is straightforward to show the equivalence by direct calculation. For any $w \in \mathbb{R}^d$,

$$\begin{aligned}\frac{1}{2}\|\tilde{X}w - \tilde{y}\|_2^2 &= \frac{1}{2}\|Xw - y\|_2^2 + \frac{1}{2}\|\sqrt{\delta}Iw - 0\|_2^2 \\ &= \frac{1}{2}\|Xw - y\|_2^2 + \frac{\delta}{2}\|w\|_2^2,\end{aligned}$$

Substituting this identity into Equation (25) establishes the equivalence.

It is clear by inspection that $\tilde{X}$ is full column rank. Then $\text{Null}(X_{\mathcal{C}}) = \emptyset$ for all $\mathcal{C} \subset \mathcal{B}$ and the solution is unique by Proposition 3.1. ☐

Proposition 3.18. *The solution to the ℓ_2-penalized problem converges to the min-norm solution as $\delta \to 0$. That is,*

$$\lim_{\delta \to 0} w^\delta(\lambda) = w^*(\lambda).$$

Proof. First we show that $\|w^\delta(\lambda)\|_2 \leq \|w^*(\lambda)\|_2$. Suppose by way of contradiction that $\|w^\delta(\lambda)\|_2 > \|w^*(\lambda)\|_2$ for some $\delta > 0$. Since

$$\begin{aligned}\frac{1}{2}\|Xw^*(\lambda) - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w^*_{b_i}(\lambda)\|_2 &= \min_{w : K_{b_i} w_{b_i} \leq 0} \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \\ &\leq \frac{1}{2}\|Xw^\delta(\lambda) - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w^\delta_{b_i}(\lambda)\|_2,\end{aligned}$$

we deduce

$$\frac{1}{2}\|Xw^*(\lambda) - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w^*_{b_i}(\lambda)\|_2 + \frac{\delta}{2}\|w^*(\lambda)\|_2^2 < \frac{1}{2}\|Xw^\delta(\lambda) - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|[w^\delta(\lambda)]_{b_i}\|_2 + \frac{\delta}{2}\|w^\delta(\lambda)\|_2^2,$$

which contradicts optimality of $w^\delta(\lambda)$. So $\|w^\delta(\lambda)\|_2 \leq \|w^*(\lambda)\|_2$ for all $\delta > 0$. As a result, the sequence $\left\{w^{\delta_k}(\lambda)\right\}_{\delta_k}$, where $\delta_k \downarrow 0$, is bounded and admits at least one convergent subsequence. Let $\bar{w}(\lambda)$ be the limit point associated with one such subsequence; clearly $\|\bar{w}(\lambda)\|_2 \leq \|w(\lambda)^*\|_2$.

Let's show that $\bar{w}(\lambda)$ is a solution to the group lasso problem by checking the KKT conditions. Suppose $\lambda > 0$. Stationarity of the Lagrangian is

$$X^\top(Xw^{\delta_k}(\lambda) - y) + K\rho^{\delta_k}(\lambda) + s^{\delta_k}(\lambda) + \delta_k w^{\delta_k}(\lambda) = 0,$$

where $s^{\delta_k}_{b_i}(\lambda) \in \partial\lambda\|w^{\delta_k}_{b_i}\|_2$. Since $\|s^{\delta_k}_{b_i}(\lambda)\|_2 \leq \lambda$ and $w^{\delta_k}(\lambda)$ is bounded, clearly $K\rho^{\delta_k}(\lambda)$ is also bounded.

Dropping to a subsequence if necessary, let $\lim_k w^{\delta_k}(\lambda) = \bar{w}$ and $\lim_k K_{b_i}\rho^{\delta_k}_{b_i} = \bar{z}_{b_i}$. Define

$$R_{1/n} = \left\{\rho_{b_i} : \|K_{b_i}\rho_{b_i} - \bar{z}_{b_i}\|_\infty \leq \frac{1}{n}, \rho_{b_i} \geq 0\right\}.$$

The sequence of sets $R_{1/n}$ is polyhedral and thus retractive. Moreover, for each $n \in \mathbb{N}$, there exists k such that

$$\|K_{b_i}\rho^{\delta_k}_{b_i} - \bar{z}_{b_i}\|_\infty \leq 1/n,$$

since $K_{b_i}\rho^{\delta_k}_{b_i} \to \bar{z}_{b_i}$. Recalling $\rho^{\delta_k}_{b_i} \geq 0$ shows that $\mathbb{R}_{1/n}$ is non-empty. The limit of a sequence of nested, non-empty, retractive sets is also non-empty (Bertsekas, 2009, Proposition 1.4.10). Moreover, since the limit is exactly

$$\bar{R} = \left\{\rho_{b_i} : K_{b_i}\rho_{b_i} = \bar{z}_{b_i}, \rho_{b_i} \geq 0\right\},$$

we deduce that there exists $\bar{\rho} \geq 0$ such that $K_{b_i}\bar{\rho_{b_i}} = \bar{z}_{b_i}$.

Taking limits on either side of the stationarity condition, we find

$$X_{b_i}^\top(y - X\bar{w}(\lambda)) - K_{b_i}\bar{\rho}(\lambda) = \bar{s}_{b_i},$$

where the limit point $\bar{s}_{b_i}$ satisfies $\|\bar{s}_{b_i}\| \leq \lambda$. If $b_i \in \mathcal{B} \setminus \mathcal{A}_\lambda(\bar{w}(\lambda))$, then $\bar{w}_{b_i}$ satisfies stationarity.

Let $b_i \in \mathcal{A}_\lambda(\bar{w}(\lambda))$. Since $w^{\delta_k}(\lambda) \to \bar{w}(\lambda)$, $\|w_{b_i}^{\delta_k} - w_{b_i}\|_2 \to 0$ and it must happen that $w_{b_i}^{\delta_k} > 0$ for all k sufficiently large. That is, $\mathcal{A}(w^{\delta_k}(\lambda)) \supseteq \mathcal{A}(\bar{w})$ for all $k \geq k'$. Using $b_i \in \mathcal{A}(w^{\delta_k}(\lambda))$ provides a closed-form expression for $s_{b_i}^{\delta_k}$:

$$\begin{aligned}\lim_k s^{\delta_k}(\lambda) &= \lambda \lim_k \frac{w_{b_i}^{\delta_k}(\lambda)}{\|w_{b_i}^{\delta_k}(\lambda)\|_2} \\ &= \lambda \frac{\bar{w}_{b_i}(\lambda)}{\|\bar{w_{b_i}}(\lambda)\|_2},\end{aligned}$$

which shows that $\bar{s}_{b_i}$ is a subgradient of $\lambda\|w_{b_i}\|_2$. We conclude that the Lagrangian is stationary in $\bar{w}_{b_i}$ as well.

Let us check the remainder of the KKT conditions. For feasibility, it is straightforward to observe that

$$K_{b_i}^\top w_{b_i}^{\delta_k}(\lambda) \leq 0 \quad \forall\, k \implies K_{b_i}^\top \bar{w}_{b_i}(\lambda) \leq 0.$$

Similarly,

$$\left\langle \rho_{b_i}^{\delta_k}, K_{b_i}^\top w_{b_i}^{\delta_k} \right\rangle = 0 \quad \forall\, k \implies \left\langle \bar{\rho_{b_i}}, K_{b_i}^\top \bar{w}_{b_i} \right\rangle \leq 0,$$

which, combined with $\bar{\rho} \geq 0$, is sufficient to establish complementary slackness. We have shown the subsequential limits $(\bar{w}, \bar{\rho})$ satisfies the KKT conditions and thus $\bar{w}$ is a solution to the constrained group lasso problem.

Since the min-norm solution is unique and $\|\bar{w}(\lambda)\|_2 \leq \|w^*(\lambda)\|_2$, it must be that $\bar{w}(\lambda) = w^*(\lambda)$. Noting that this holds for every limit point implies $\lim_{\delta \downarrow 0} w^\delta(\lambda)$ exists and is $w^*(\lambda)$. This completes the proof for $\lambda > 0$.

If $\lambda = 0$, then the proposition follows similarly with the additional observation that $s^{\delta_k}(0) = 0$ for all k. □

A.6. Sensitivity

Proposition 3.19. *Let $w \in \mathcal{W}^*(\lambda, y)$ be minimal. The active blocks $w_{\mathcal{A}_\lambda}$ are the unique solution to Problem* (18).

Proof. Let w be as in the theorem statement. We starting by showing that w obtains the optimal objective value for the reduced problem:

$$\begin{aligned}\frac{1}{2}\|X_{\mathcal{A}_\lambda} w_{\mathcal{A}_\lambda} - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{A}_\lambda} \|w_{b_i}\|_2 &= \min_{w : K_{b_i} w_{b_i} \leq 0} \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \\ &\leq \min_{w_{\mathcal{A}_\lambda} : K_{\mathcal{A}_\lambda} w_{\mathcal{A}_\lambda} \leq 0} \frac{1}{2}\|X_{\mathcal{A}_\lambda} w_{\mathcal{A}_\lambda} - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{A}_\lambda} \|w_{b_i}\|_2 \\ &\leq \frac{1}{2}\|X_{\mathcal{A}_\lambda} w_{\mathcal{A}_\lambda} - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{A}_\lambda} \|w_{b_i}\|_2,\end{aligned}$$

where the last inequality makes explicit use of feasibility of $w_{\mathcal{A}_\lambda}$. Since $w_{\mathcal{A}_\lambda}$ is feasible for the reduced problem and attains the minimum objective value, it must be optimal. Note that it is straightforward to check that the active blocks of the min-norm dual parameter $\rho^*_{\mathcal{A}_\lambda}$ are dual optimal for the reduced problem.

Now, let $w'_{\mathcal{A}_\lambda}$ be an optimal solution to the reduced problem. Since $\rho^*_{\mathcal{A}_\lambda}$ is dual optimal for the reduced problem, Proposition 3.1 implies

$$w'_{b_i} = \alpha'_{b_i} v_{b_i},$$

for every $b_i \in \mathcal{A}_\lambda$, with $\alpha'_{b_i} \geq 0$. Since $X_{\mathcal{A}_\lambda} w'_{\mathcal{A}_\lambda} = X_{\mathcal{A}_\lambda} w_{\mathcal{A}_\lambda}$, we deduce

$$\sum_{b_i \in \mathcal{A}_\lambda} (\alpha_{b_i} - \alpha'_{b_i}) X_{b_i} b_i = 0,$$

which contradicts minimality of w unless $\alpha_{b_i} = \alpha'_{b_i}$. That is, $w'_{\mathcal{A}_\lambda} = w_{\mathcal{A}_\lambda}$. We conclude that the reduced problem provides the unique minimal solution with support $\mathcal{A}_\lambda$. □

Lemma A.15. *Let $w \in \mathcal{W}^*(\lambda)$ be minimal. Then w is a second-order stationary point of the reduced problem (Equation (18)).*

Proof. Define $M(w)$ to be the block-diagonal projection matrix given by

$$M(\bar{w})_{b_i} = \frac{1}{\|\bar{w}_{b_i}\|_2} \left(I - \frac{\bar{w}_{b_i}}{\|\bar{w}_{b_i}\|_2} \frac{\bar{w}_{b_i}^\top}{\|\bar{w}_{b_i}\|_2} \right). \tag{26}$$

The Hessian of the Lagrangian of the reduced problem with respect to w is exactly

$$\nabla_w^2 \mathcal{L}(w_{\mathcal{A}_\lambda}, \rho_{\mathcal{A}_\lambda}) = X_{\mathcal{A}_\lambda}^\top X_{\mathcal{A}_\lambda} + \lambda M_{\mathcal{A}_\lambda}(w).$$

A sufficient condition for w to be second-order stationary is that this Hessian is positive-definite. We now shows this fact holds.

Clearly $\nabla_w^2 \mathcal{L}(w_{\mathcal{A}_\lambda}, \rho_{\mathcal{A}_\lambda})$ is positive semi-definite as it is the sum of a PSD projection matrix and a Gram matrix, which is always PSD. Let $\bar{w} \in \mathbb{R}^{|\mathcal{A}_\lambda|}$ such that $\bar{w} \neq 0$. Suppose that

$$\begin{aligned} 0 &= \bar{w}^\top \nabla_w^2 \mathcal{L}(w_{\mathcal{A}_\lambda}, \rho_{\mathcal{A}_\lambda}) \bar{w} \\ &= \bar{w}^\top X_{\mathcal{A}_\lambda}^\top X_{\mathcal{A}_\lambda} \bar{w} + \bar{w}^\top \lambda M(\bar{w}) \bar{w} \\ &= \|X_{\mathcal{A}_\lambda} w\|_2^2 + \lambda w^\top M(\bar{w}) w. \end{aligned}$$

Since $M(\bar{w})$ is PSD, it must hold that

$$\bar{w}^\top M(\bar{w}) \bar{w} = 0,$$

which is true if and only if $\bar{w}_{b_i} = \beta_{b_i} \bar{w}_{b_i}$, $\beta_{b_i} \in \mathbb{R}$, for each $b_i \in \mathcal{A}_\lambda$. As a result, we find that

$$X_{\mathcal{A}_\lambda} w = \sum_{b_i \in \mathcal{A}_\lambda} \beta_{b_i} X_{b_i} \bar{w}_{b_i} = 0,$$

which is a contradiction with the fact that $\bar{w}$ is a minimal solution. We conclude that the Hessian is positive-definite as desired. ☐

Proposition 3.22. *Let $w \in \mathcal{W}^*(\bar{\lambda}, \bar{y})$ be minimal and suppose w satisfies LICQ on the active set $\mathcal{A}_\lambda$ and SCS on the equicorrelation set $\mathcal{E}_\lambda$. Then w has a locally continuous solution function $(\lambda, y) \mapsto w(\lambda, y)$. Moreover, if*

$$D = \begin{bmatrix} X_{\mathcal{A}_\lambda}^\top X_{\mathcal{A}_\lambda} + M(\bar{w}) & K_{\mathcal{A}_\lambda} \\ \rho_{\bar{\mathcal{A}}_\lambda} \odot K_{\mathcal{A}_\lambda} & diag(K_{\mathcal{A}_\lambda}^\top w_{\bar{\mathcal{A}}_\lambda}) \end{bmatrix},$$

where $\odot$ is the element-wise product, $u_{b_i} = \frac{w_{b_i}}{\|w_{b_i}\|_2}$, u is the concatenation of these vectors, and M is block-diagonal projection matrix in Equation (26), then the Jacobians of $w(\bar{\lambda}, \bar{y})$ with respect to λ and y are given as follows:

$$\nabla_\lambda w(\bar{\lambda}, \bar{y}) = -[D^{-1}]_{\mathcal{A}_\lambda} u_{\mathcal{A}_\lambda} \;\; \nabla_y w(\bar{\lambda}, \bar{y}) = [D^{-1}]_{\mathcal{A}_\lambda} X_{\mathcal{A}_\lambda}^\top,$$

where $[D_{\mathcal{A}_\lambda}^{-1}]_{\mathcal{A}_\lambda}$ is the $|\mathcal{A}_\lambda| \times |\mathcal{A}_\lambda|$ dimensional leading principle submatrix of D.

Proof. Recall from Proposition 3.19 that $w_{\mathcal{A}_\lambda}$ is the unique solution to the reduced group lasso problem. In fact, as we show in Lemma A.15, $w_{\mathcal{A}_\lambda}$ is a second order stationary point for the reduced problem. Now, combining this fact with LICQ and SCS and using standard results on differential sensitivity from optimization theory (see, e.g. Fiacco & Ishizuka (1990, Theorem 5.1) and the references therein) we obtain the following:

For (λ, y) in a neighborhood of $\bar{\lambda}, \bar{y}$, there exists a unique once continuously differentiable function

$$\tilde{l}(\lambda, y) = \begin{bmatrix} \tilde{h}(\lambda, y) \\ \tilde{g}(\lambda, y), \end{bmatrix}$$

such that $\tilde{h}(\bar{\lambda}, \bar{y}) = w_{\mathcal{A}_\lambda}$, $\tilde{g}(\bar{\lambda}, \bar{y}) = \rho_{\mathcal{A}_\lambda}^*$, and $\tilde{l}(\lambda, y)$ is a primal-dual solution to the reduced problem.

Now we show that $\tilde{l}$ can be extended from the reduced problem to obtain a local solution function for the constrained group lasso. Define $h(\lambda, y)$ such that $h_{\mathcal{A}_\lambda}(\lambda, y) = \tilde{h}(\lambda, y)$ and $h_{\mathcal{B}\setminus\mathcal{A}_\lambda}(\lambda, y) = 0$. We shall show how to extend g shortly. For $b_i \in \mathcal{A}_\lambda$, the pair $h_{b_i}(\lambda, y), g_{b_i}(\lambda, y)$ verifies the KKT conditions (which are separable over block) since it verifies them for the reduced problem. So, we need only consider $b_i \in \mathcal{B} \setminus \mathcal{A}_\lambda$.

First, consider $b_i \in \mathcal{B} \setminus \mathcal{E}_\lambda$. In this case, we have

$$\|X_{b_i}^\top(\bar{y} - Xh(\bar{\lambda}, \bar{y})) + K_{b_i}\rho_{b_i}(\bar{\lambda}, \bar{y})\|_2 < \bar{\lambda},$$

Since this inequality is strict and

$$z(\bar{\lambda}, \bar{y}) = X_{b_i}^\top(\bar{y} - Xh(\bar{\lambda}, \bar{y}),$$

is continuous in $\bar{\lambda}, \bar{y}$, there exists a neighborhood of $\bar{\lambda}, \bar{y}$ on which

$$\|z(\lambda, y) + K_{b_i}\rho_{b_i}(\bar{\lambda}, \bar{y})\|_2 \leq \lambda.$$

Since $\rho_{b_i}(\bar{\lambda}, \bar{y}) \geq 0$ and $w_{b_i} = 0$, dual feasibility and complementary slackness hold. We conclude that the extension $g_{b_i}(\lambda, y) = \rho_{b_i}(\bar{\lambda}, \bar{y})$ satisfies KKT conditions on this neighbourhood.

Now suppose $b_i \in \mathcal{E}_\lambda(\bar{\lambda}, \bar{y}) \setminus \mathcal{A}_\lambda$. If

$$\|X_{b_i}^\top(y - Xh(\lambda, y))\|_2 = \lambda,$$

then taking $g_{b_i}(y, \lambda) = 0$ satisfies KKT conditions. Otherwise, observe that

$$\|X_{b_i}^\top(\bar{y} - Xh(\bar{\lambda}, \bar{y})) + K_{b_i}\rho_{b_i}(\bar{\lambda}, \bar{y})\|_2 = \bar{\lambda},$$

must hold for some dual parameter $\rho_{b_i}(\bar{\lambda}, \bar{y})$ by KKT conditions. Moreover, SCS implies that we can choose the dual parameter to satisfy,

$$\rho_{b_i}(\bar{\lambda}, \bar{y}) > 0,$$

since $K_{b_i}^\top w_{b_i}(\bar{\lambda}, \bar{y}) = 0$. Finally, because

$$X_{b_i}^\top(y - Xh(\lambda, y))$$

is a continuous function of (y, λ), taking λ, y sufficiently close to $\bar{\lambda}, \bar{y}$ implies there exists $\rho_{b_i}(\lambda, y) \geq 0$ such that

$$\|X_{b_i}^\top(y - Xh(\lambda, y)) + K_{b_i}\rho_{b_i}(\lambda, y)\|_2 \leq \lambda.$$

Now we choose our extension to be $g_{b_i}(\lambda, y) = \rho_{b_i}(\lambda, y)$ so that (h_{b_i}, g_{b_i}) satisfies stationarity of the Lagrangian as well. Since g_{b_i} is feasible and h_{b_i} is the zero function, primal feasibility, dual feasibility, and complementary slackness also hold.

Since $l = (g, h)$ satisfies the KKT conditions in a local neighborhood of $\bar{\lambda}, \bar{y}$, it is exactly a local solution function. Moreover, since $g_{\mathcal{B}\setminus\mathcal{A}}(\lambda, y) = 0$ over this neighborhood, it is easy to see that the gradient for parameter blocks in $\mathcal{B} \setminus \mathcal{A}$ is 0. For $g_{\mathcal{A}_\lambda}$, Fiacco & Ishizuka (1990, Theorem 5.1) implies that the gradients are given as follows:

Recall from Lemma A.15, that $M_{\mathcal{A}_\lambda}$ is a block-diagonal projection matrix with blocks given by

$$M(\bar{w})_{b_i} = \frac{1}{\|\bar{w}_{b_i}\|_2}\left(I - \frac{\bar{w}_{b_i}}{\|\bar{w}_{b_i}\|_2}\frac{\bar{w}_{b_i}^\top}{\|\bar{w}_{b_i}\|_2}\right).$$

Then, the Jacobian of $\nabla_w \mathcal{L}(w_{\bar{\mathcal{A}}_\lambda}, \rho_{\bar{\mathcal{A}}_\lambda})$ for the reduced problem with respect to the primal-dual parameters is given by

$$D = \begin{bmatrix} X_{\mathcal{A}_\lambda}^\top X_{\mathcal{A}_\lambda} + M(\bar{w}) & K_{\mathcal{A}_\lambda} \\ \rho_{\bar{\mathcal{A}}_\lambda} \odot K_{\mathcal{A}_\lambda} & \text{diag}(K_{\mathcal{A}_\lambda}^\top w_{\bar{\mathcal{A}}_\lambda}) \end{bmatrix}.$$

It also holds that D is invertible. Finally, let $u_i = \frac{w_i}{\|w_{b_i}\|_2}$ and u the concatenation of these vectors. We are now able to write the Jacobians of $w(y, \lambda)$ with respect to y and λ as follows:

$$\nabla_\lambda w(\bar{\lambda}, \bar{y}) = -[D^{-1}]_{\mathcal{A}_\lambda} u_{\mathcal{A}_\lambda} \qquad \nabla_y w(\bar{\lambda}, \bar{y}) = [D^{-1}]_{\mathcal{A}_\lambda} X_{\mathcal{A}_\lambda}^\top,$$

where $[D_{\mathcal{A}_\lambda}^{-1}]_{\mathcal{A}_\lambda}$ is the $|\mathcal{A}_\lambda| \times |\mathcal{A}_\lambda|$ dimensional leading principle submatrix of D. □

B. Specialization: Proofs

Proposition 4.2. *The set of stationary points of two-layer ReLU networks is given by*

$$\begin{aligned}\mathcal{C}_\lambda = \big\{(W_1,w_2) : \tilde{\mathcal{D}} \subseteq \mathcal{D}_Z,\ f_{W_1,w_2}(Z) = \hat{y}_{\tilde{\mathcal{D}}},\\ W_{1i} = (\alpha_i/\lambda)^{1/2} v_i(\tilde{\mathcal{D}}),\ w_{2i} = (\alpha_i \lambda)^{1/2},\\ \alpha_i \geq 0,\ i \in [2|\tilde{\mathcal{D}}|] \setminus \mathcal{S}_\lambda \implies \alpha_i = 0\big\},\end{aligned}$$

where $\tilde{\mathcal{D}}$ are sub-sampled activation patterns, $\hat{y}_{\tilde{\mathcal{D}}}$ is the optimal model fit using those patterns, and $v_i(\tilde{\mathcal{D}}) = c_{b_i}(\tilde{\mathcal{D}}) - K_{b_i}\rho_{b_i}(\tilde{\mathcal{D}})$ is determined by the fit and the dual parameters.

Proof. The proof is almost immediate.

Given any sub-sampled set of activation patterns $\tilde{\mathcal{D}} \subset \mathcal{D}_Z$, Wang et al. (2021, Theorem 3) prove that the solutions to the sub-sampled convex program are Clarke stationary points (Clarke, 1990) of the non-convex ReLU optimization problem in Equation (1), and vice-versa. Using the expression for the CGL solution set in Proposition 3.1, which applies to sub-sampled convex reformulations as well as the full program, we obtain a version of Corollary 4.1 for stationary points. That is, every model (W_1, w_2) in

$$\begin{aligned}\mathcal{C}_\lambda(\tilde{\mathcal{D}}) = \big\{(W_1,w_2) :\ f_{W_1,w_2}(Z) = \hat{y}_{\tilde{\mathcal{D}}}, W_{1i} = (\alpha_i/\lambda)^{1/2} v_i(\tilde{\mathcal{D}}),\ w_{2i} = (\alpha_i \lambda)^{1/2},\\ \alpha_i \geq 0,\ i \in [m] \setminus \mathcal{S}_\lambda \implies \alpha_i = 0\big\},\end{aligned}$$

is a stationary point of the non-convex ReLU program. Taking the union over all sub-sampled sets of activation patterns gives $\mathcal{C}_\lambda$, which is guaranteed to contain every stationary point of the non-convex objective. This completes the proof. ☐

Lemma B.1. *Let (W_1, w_2) and (W_1', w_2') be two solutions to the non-convex ReLU training problem. If for every $i \in [m]$, it holds that*

$$W_{1i} w_{2i} = W_{1i}' w_{2i}',$$

and sign(w_{2i}) = sign(w_{2i}'), then $W_1 = W_1'$ and $w_2 = w_2'$. That is, the solutions are the same.

Proof. The ReLU prediction function f_{W_1,w_2} is invariant to scalings of the form

$$\bar{W}_{1i} = \alpha W_{1i} \quad \bar{w}_{2i} = w_{2i}/\alpha,$$

where $\alpha > 0$. Using this, we deduce that both solutions must satisfy the following equations:

$$\begin{aligned}1 &= \arg\min_\alpha \alpha^2 \|W_{1i}\|_2^2 + \|w_{2i}\|_2^2/\alpha^2\\ 1 &= \arg\min_\alpha \alpha^2 \|W_{1i}'\|_2^2 + \|w_{2i}'\|_2^2/\alpha^2,\end{aligned}$$

which in turn implies that

$$\|W_{1i}\|_2 = \|w_{2i}\|_2.$$

We deduce

$$\begin{aligned}\|W_{1i}\|_2^2 &= \|W_{1i} \cdot \|w_{2i}\|_2\|_2\\ &= \|W_{1i}' \cdot \|w_{2i}'\|_2\|_2\\ &= \|W_{1i}'\|_2^2,\end{aligned}$$

where we have used the fact that $\|w_{2i}\|_2 = |w_{2i}|$. But this implies $W_{1i} = W_{1i}'$ and $w_{2i} = w_{2i}'$, completing the proof. ☐

Lemma B.2. *Let (W_1, w_2) and (W_1', w_2') be two solutions to the non-convex ReLU training problem. If (W_1, w_2) and (W_1', w_2') map to the same solution in the convex reformulation, then they are equal up to permutations of the neurons.*

Proof. We recall from Pilanci & Ergen (2020) that the mapping from solutions to the non-convex training problem to solutions of the convex reformulation is given by

$$v_i = \begin{cases} W_{1i} w_{2i} & \text{if } w_{2i} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$
$$u_i = \begin{cases} -W_{1i} w_{2i} & \text{if } w_{2i} < 0 \\ 0 & \text{otherwise,} \end{cases}$$

Since (W_1, w_2) and (W_1', w_2') map to the same solution, the following must hold (up to orderings of the neurons):

$$W_{1i} w_{2i} = W_{1i}' w_{2i}'.$$

Lemma B.1 now implies the two solutions are the same up to permutations. □

Proposition 4.3. *Let $\lambda > 0$ and suppose that the convex ReLU problem has a unique solution. Then the ReLU model solution is p-unique. In particular, if $\{D_i Z v_{b_i}\}_{\mathcal{E}_\lambda}$ are linearly independent, then the non-convex solution is p-unique.*

Proof. Since is only one solution to the convex reformulation, all solutions to the non-convex training problem must map to that solution. Lemma B.2 now implies that the solution map for the non-convex problem is p-unique. □

Proposition 4.4. *Let $\lambda > 0$ and $p = |\tilde{\mathcal{D}}|$. Suppose Z follows a continuous probability distribution and nnz$(D_i) \geq p \cdot d$ for every $D_i \in \tilde{\mathcal{D}}$. If $\mathcal{E}_\lambda$ does not contain two blocks with the same activation pattern, then the sub-sampled convex ReLU program has a unique solution almost surely.*

Proof. We assume without loss of generality that only indices from 1 to p are in $\mathcal{E}_\lambda$. By Lemma 3.3, the constrained group lasso admits a unique solution if and only if

$$\bigcup_{i \in \mathcal{E}_\lambda} \{D_i Z v_{b_i}\},$$

are linearly independent. We now show that this fact holds under the proposed sufficient condition by proving the stronger fact that $\bigcup_{i \in \mathcal{E}_\lambda} \{[D_i Z]_j : j \in [d]\}$ are linearly independent with probability one, where $[D_i Z]_j$ is the j^{th} column of $D_i Z$. Since nnz$(D_i) \geq d * p$ and Z has a continuous probability distribution, it holds that $[D_i Z]_j$ has at least $d * p$ non-zero entries with probability 1. Let

$$\mathcal{S}_{ij} = \text{Span}\left(\bigcup_{i \in \mathcal{E}_\lambda} \{[D_i Z]_j : j \in [d]\} \setminus [D_i Z]_j \right),$$

and observe that dim$(\mathcal{S}) \leq d * p - 1$. As a result, the conditional probability $[D_i Z]_j$ falls in this subspace satisfies

$$\Pr([D_i Z]_j \in \mathcal{S}_{ij} | \bigcup_{i \in \mathcal{E}_\lambda} \{[D_i Z]_j : j \in [d]\} \setminus [D_i Z]_j) = 0.$$

Taking expectations over the remaining vectors in Z implies

$$\Pr([D_i Z]_j \in \mathcal{S}_{ij}) = 0.$$

Finally, using a union bound over i, j implies that $\bigcup_{i \in \mathcal{E}_\lambda} \{[D_i Z]_j : j \in [d]\}$ are linearly independent almost surely. □

Proposition 4.5. *Suppose $r = rank(X)$. Then an optimal and minimal ReLU network with at most $m^* \leq n$ non-zero neurons can be computed in $O\left(d^3 r^3 (n/r)^{3r}\right)$ time.*

Proof. The proof follows directly from existing results.

Recall from Pilanci & Ergen (2020) that there are at most

$$p \in O\left(r \frac{n}{r}^{3r} \right),$$

activation patterns in the convex reformulation and that the complexity of computing an optimal ReLU model using a standard interior-point solver is $O(d^3 r^3 (n/r)^{3r})$.

We know from Proposition A.6 that the complexity of pruning an optimal neural network with at most $2p$ neuron is $O(n^3 p + nd)$. Combining these complexities, we find that the cost of optimization dominates and overall complexity of computing an optimal and minimal neural network grows as $O(d^3 r^3 (n/r)^{3r})$.

Finally, the bound on the number of active neurons follows from the fact that $\dim \text{Span}(\{(XW_{1i}^*)_+\}_i) \leq n$. This completes the proof. □

Lemma 4.8. *The minimum ℓ_2-norm solution to the convex reformulation of a (gated) ReLU model corresponds to the p-unique optimal neural network which minimizes*

$$r(W_1, w_2) = \sum_{i=1}^{m} \|W_{1i}\|_2^4 + \|w_{2i}\|_2^4.$$

Proof. Let (u, v) be an optimal solution to the convex reformulation. Pilanci & Ergen (2020) show that an optimal solution to the original two-layer ReLU optimization problem is given by setting

$$W_{1i} = \frac{u_i}{\sqrt{\|u_i\|_2}}, w_{2i} = \sqrt{\|u_i\|_2},$$

and

$$W_{1j} = \frac{v_i}{\sqrt{\|v_i\|_2}}, w_{2j} = -\sqrt{\|v_i\|_2},$$

where we define $\frac{0}{0} = 0$. Then, the r-value of any such solution can be calculated as

$$\begin{aligned} r(W_1, w_2) &= \sum_{i=1}^{m} \|W_{1i}\|_2^4 + \|w_{2i}\|_2^4 \\ &= \sum_{u_i \neq 0} \|\frac{u_i}{\sqrt{\|u_i\|_2}}\|_2^4 + \|\sqrt{\|u_i\|_2}\|_2^4 + \sum_{v_i \neq 0} \|\frac{v_i}{\sqrt{\|v_i\|_2}}\|_2^4 + \|\sqrt{\|v_i\|_2}\|_2^4 \\ &= \sum_{u_i \neq 0} \|u_i\|_2^2 + \|u_i\|_2^2 + \sum_{v_i \neq 0} \|v_i\|_2^2 + \|v_i\|_2^2 \\ &= 2\|u\|_2^2 + 2\|v\|_2^2. \end{aligned}$$

That is, $r(W_1, w_2)$ is a monotone transformation of the Euclidean norm of (u, v). Moreover, since every optimal ReLU network can be obtained as the solution to a convex reformulation, the minimum r-valued optimal ReLU network is given by the minimum ℓ_2-norm solution to the convex reformulation.

Finally, let's show that this solution is unique up to permutations. Suppose (W_1', w_2') is a optimal ReLU model which also minimizes r. We know from the theory of convex reformulations that W_{1i}', w_{2i}' corresponds to an optimal solution of the convex program; by reversing our calculations above, we deduce that this convex parameterization must also minimize the ℓ_2-norm. Such the minimum $\ell - 2$-norm solution to the convex reformulation is unique, we have

$$W_{1i}' = \alpha_i' u_i^* = \frac{\alpha_i'}{\alpha_i} W_{1i},$$

so that each neuron in the two solutions is related by a strictly positive scaling. Lemma B.2 now implies the optimal neural network which minimizes r is p-unique. □

Proposition 4.6. *There exists (Z, y) for which $\mathcal{O}^*$ is not open nor is the model fit $f_{W_1, w_2}(Z)$ continuous in λ.*

Proof. Mishkin et al. (2022) show that Equation 1 has the same global optimal values as

$$\min_{v, \gamma \in \{-1, 1\}} \frac{1}{2} \| \sum_{i=1}^{p} (Xv_i)_+ \gamma - y \| + \lambda \sum_{i=1}^{p} \|w_i\|_2. \tag{27}$$

Moreover, the objective-preserving mapping $W_{1i} = v_i / \sqrt{\|v_i\|_2}$, $w_{2i} = \gamma_i \sqrt{\|v_i\|_2}$ can be used to obtain an optimal ReLU network from a solution to Equation (27). We proceed by analyzing this equivalent problem and then use the mapping to return to the original non-convex formulation.

Consider the case $p = 1$. Let X, y consist of two training points, $(x_1, y_1) = (-100, 1)$ and $(x_2, y_2) = (1, 10)$. In what follows, we drop the subscript for v and γ since $p = 1$ and we consider a one-dimensional. The optimization problem of interest is

$$\min_{v,\gamma} \frac{1}{2} \left((x_1 v)_+ \gamma - y_1\right)^2 + \left((x_2 v)_+ \gamma - y_2\right)^2 + \lambda \left|v\right|.$$

Since $x_1 < 0$ and $x_2 > 0$, we can re-write this optimization problem as

$$\min_{v,\gamma} \frac{1}{2} \mathbb{1}_{v \leq 0} \left((x_1 v \gamma - y_1)^2 + y_2^2 \right) + \mathbb{1}_{v > 0} \left((x_2 v \gamma - y_2)^2 + y_1^2 \right) + \lambda \left|v\right|.$$

By inspection, we see that $\gamma = +1$ is optimal in both cases, leading to the following simplified expression:

$$\min_{v} \frac{1}{2} \mathbb{1}_{v \leq 0} \left((x_1 v - y_1)^2 + y_2^2 \right) + \mathbb{1}_{v > 0} \left((x_2 v - y_2)^2 + y_1^2 \right) + \lambda \left|v\right|.$$

This is a piece-wise continuous (but non-smooth) quadratic with a breakpoint at $v^* = 0$. We determine the solution to this minimization problem via a case analysis.

Case 1: $v^* = 0$. Then, the optimal objective is trivially $f(v^*) = y_1^2 + y_2^2 = 101$.

Case 2: $v^* < 0$. First order optimality conditions are

$$x_1 \left(x_1 v_-^* - y_1 \right) - \lambda = 0 \implies v_-^* = \frac{y_1 \cdot x_1 + \lambda}{x_1^2},$$

which is valid only if $\lambda < |y_1 \cdot x_1| = 100$. The minimum objective value is then

$$\begin{aligned}
\frac{1}{2} \left(\left(x_1 v_-^* - y_1 \right)^2 + y_2^2 \right) - \lambda v_-^* &= \left(\left(x_1 \frac{y_1 \cdot x_1 + \lambda}{x_1^2} - y_1 \right)^2 + y_2^2 \right) - \lambda \left(\frac{y_1 \cdot x_1 + \lambda}{x_1^2} \right) \\
&= \frac{\lambda^2}{x_1^2} + y_2^2 - \lambda \left(\frac{y_1 \cdot x_1 + \lambda}{x_1^2} \right) \\
&= -\frac{\lambda y_1}{x_1} + y_2^2 \\
&= \frac{\lambda}{100} + 100.
\end{aligned}$$

Case 3: $v_+^* > 0$. Similarly to the previous case, we obtain

$$x_2 \left(x_2 v_+^* - y_2 \right) + \lambda = 0 \implies v_+^* = \frac{y_2 \cdot x_2 - \lambda}{x_2^2},$$

which is valid only if $\lambda < |y_2 \cdot x_2| = 10$. In this case, the minimum objective is

$$\begin{aligned}
\frac{1}{2} \left(\left(x_2 v_+^* - y_2 \right)^2 + y_1^2 \right) + \lambda v_+^* &= \left(\left(x_2 \frac{y_2 \cdot x_2 + \lambda}{x_2^2} - y_2 \right)^2 + y_1^2 \right) - \lambda \left(\frac{y_2 \cdot x_2 + \lambda}{x_2^2} \right) \\
&= \frac{\lambda^2}{x_2^2} + y_2^2 + \lambda \left(\frac{y_2 \cdot x_2 - \lambda}{x_2^2} \right) \\
&= \frac{\lambda y_2}{x_2} + y_1^2 \\
&= 10\lambda + 1.
\end{aligned}$$

To combine the cases, observe that

$$\begin{aligned} f(v_+^*) - f(v_-^*) &= \frac{\lambda y_2}{x_2} + y_1^2 - \left[-\frac{\lambda y_1}{x_1} + y_2^2 \right] \\ &= \lambda \left(\frac{x_1 y_2 + x_2 y_1}{x_1 x_2} \right) + y_1^2 - y_2^2 \\ &= \lambda \left(\frac{-100(10) + 1(1)}{-100(1)} \right) + 1 - 100 \\ &= 9.99\lambda - 99. \end{aligned}$$

We deduce that $f(v_+^*) - f(v_-^*) > 0$ (and thus $v^* \leq 0$) whenever $\lambda > \frac{99}{9.99} \approx 10$ and $f(v_+^*) - f(v_-^*) > 0 < 0$ otherwise. In this latter case, $v^* \geq 0$.

Taking $\lambda = 10$, we find

$$f(v_-^*) = 100 - 0.1 = 99.99 < 101 = y_1^2 + y_2^2 = f(0),$$

so that v_-^* is optimal and $v_-^* < 0$. Moreover, v_-^* is strictly increasing as a function of λ, for all $\lambda > \frac{99}{9.99}$ so that v_-^* is optimal and strictly negative on the interval $[\frac{99}{9.99}, 10]$.

Now, consider $\lambda = \frac{99}{9.99} - \epsilon$ to see that

$$f(v_+^*) = \frac{990}{9.99} + 1 - 10\epsilon < 101 = y_1^2 + y_2^2 = f(0),$$

for all $\epsilon > 0$. We deduce that v_+^* is optimal for all $\epsilon > 0$ and thus v_+^* is optimal and strictly positive on the interval $[0, \frac{99}{9.99}]$.

To summarize, the solution function for this problem is as follows:

$$\mathcal{W}^*(\lambda) = \begin{cases} \frac{\lambda}{100^2} - 0.01 & \text{if } \lambda > \frac{99}{9.99} \\ \left\{ \frac{\lambda}{100^2} - 0.01, 0.1 - \frac{\lambda}{100} \right\} & \text{if } \lambda = \frac{99}{9.99} \\ 0.1 - \frac{\lambda}{100} & \text{if } \lambda < \frac{99}{9.99}. \end{cases}$$

This point-to-set map is clearly not open: for every sequence $\lambda_k \uparrow \frac{99}{9.99}$, $v_k \in \mathcal{W}^*(\lambda_k)$ implies $\lim_k v_k \neq \frac{\lambda}{100^2} - 0.01$. Moreover, a similar result holds for limits from above. Finally, it is clear by inspection that the optimal model fit is not unique at $\lambda = \frac{99}{9.99}$, cannot be continuous in the functional sense, and, since it is not open, also fails to be continuous in the sense of point-to-set maps. □

C. Extension to General Losses

In this section, we briefly discuss how to extend our results to general loss functions. Although we use the least-squares error throughout our derivations, this can be generalized to a smooth and strictly convex loss function $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ without difficulty. To do so, consider the more general problem,

$$\begin{aligned} p^*(\lambda) = \min_w F_\lambda(w) := \frac{1}{2} L(Xw, y) + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2 \\ \text{s.t.} \quad K_{b_i}^\top w_{b_i} \leq 0 \text{ for all } b_i \in \mathcal{B}. \end{aligned} \tag{28}$$

If L is strictly convex, then uniqueness of the optimal model fit $\hat{y}(\lambda) = Xw$ follows from straightforward adaption of Lemma A.1. Indeed, the only property of the squared-error used in this lemma is strict convexity.

Since the model fit is constant in $\mathcal{W}^*$ and L is both smooth and strictly convex, the gradient $\nabla_w L(Xw, y)$, which is given by

$$X^\top \nabla_{\hat{y}} L(\hat{y}, y),$$

must also be constant over $\mathcal{W}^*$. Thus, it is straightforward to replace the correlation vector $c_{b_i} = X_{b_i}^\top (y - Xw)$ with $X_{b_i}^\top \nabla_{\hat{y}} L(\hat{y}, y)$ throughout our derivations.

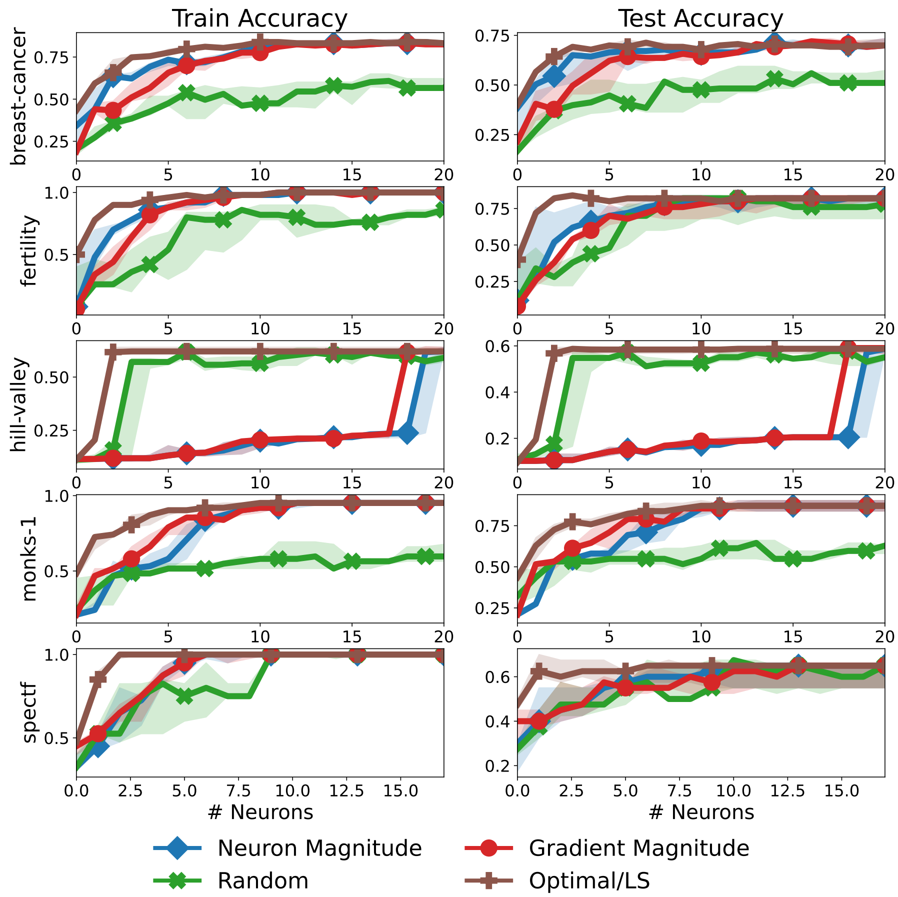

Figure 4. Pruning neurons on five datasets from the UCI repository. This figure extends Figure 2 with training accuracy in addition to the test accuracies shown in the main paper.

We form a Lagrange dual problem for CGL for one continuity-type result. Proposition 3.15 uses the Lagrange dual to show that the correlation vector c_{b_i} is the unique solution to a convex optimization program and applies standard sensitivity results to obtain continuity of $\hat{y}$. In this same fashion, $X_{b_i}^\top \nabla_{\hat{y}} L(\hat{y}, y)$ is the unique solution to a Lagrange dual problem where the dual objective uses the convex conjugate L^*, rather than the dual of the quadratic penalty. If $\nabla_{\hat{y}} L(\hat{y}, y)$ is continuous in $\hat{y}$, then this is sufficient to deduce continuity of the model fit using the same argument.

D. Additional Experiments

In this section we provide additional experimental results as well as the necessary details to replicate our pruning experiments. Code to replicate all of our experiments is provided at `https://github.com/pilancilab/relu_optimal_sets`.

D.1. Additional Results

Tuning Table 2 shows results for our tuning task on an additional 7 datasets, as well as the 10 given in Table 1. We report the interquartile range as well as median test accuracies for each method. We observe similar results as presented in the main text. Only one dataset (`tic-tac-toe`) shows no variation in test accuracy as we explore the optimal set.

Pruning: Figure 4 shows train and test accuracy for our optimal/least-squares pruning method as well as magnitude/gradient-based pruning and random pruning on the same five datasets from the UCI repository as presented in Figure 2. Our approach shows significantly less decay in train accuracy as neurons are pruned; this matches the intuition of the least-squares heuristic for pruning, which selects the coefficients β to best preserve the model fit.

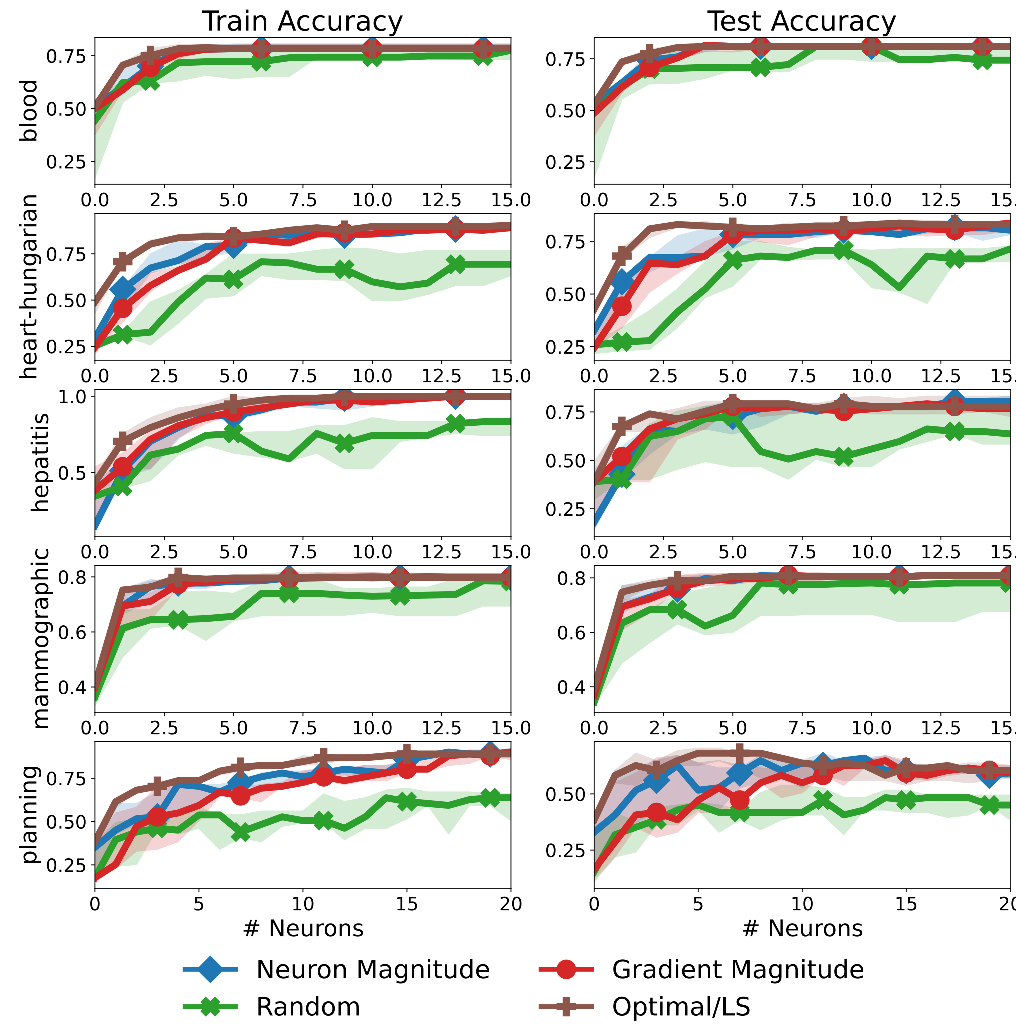

Figure 5. Pruning neurons on five additional datasets from the UCI repository. See Figure 2 for details. Our method (**Optimal/LS**) preservers test accuracy for longer than the baseline methods, leading to compact models with better generalization.

We observe that both our pruning method and pruning by neuron/gradient norm show very similar training accuracy until most of the neurons have been pruned. While this behavior is expected from our theory-based approach, it is somewhat surprising that pruning by neuron/gradient-norm also maintains train accuracy nearly as well. This behavior suggests that there are many neurons with very small norm which can be eliminated without significantly affecting the model prediction.

Figure 5 presents results for neuron pruning on five additional datasets from the UCI repository, while Figure 7 shows results for three binary classification tasks taken from the MNIST dataset. The trends are generally the same as in Figure 4, with our approach (**Theory/LS**) outperforming the baselines. Finally Figure 6 extends the results on CIFAR-10 given in Figure 3 with one additional task and with training accuracies.

D.2. Experimental Details

Now we give the details necessary to reproduce our experiments. Our experiments use the pre-processed versions of UCI datasets provided by Delgado et al. (2014), but do we do not use their evaluation procedure as it is known to have data leakage.

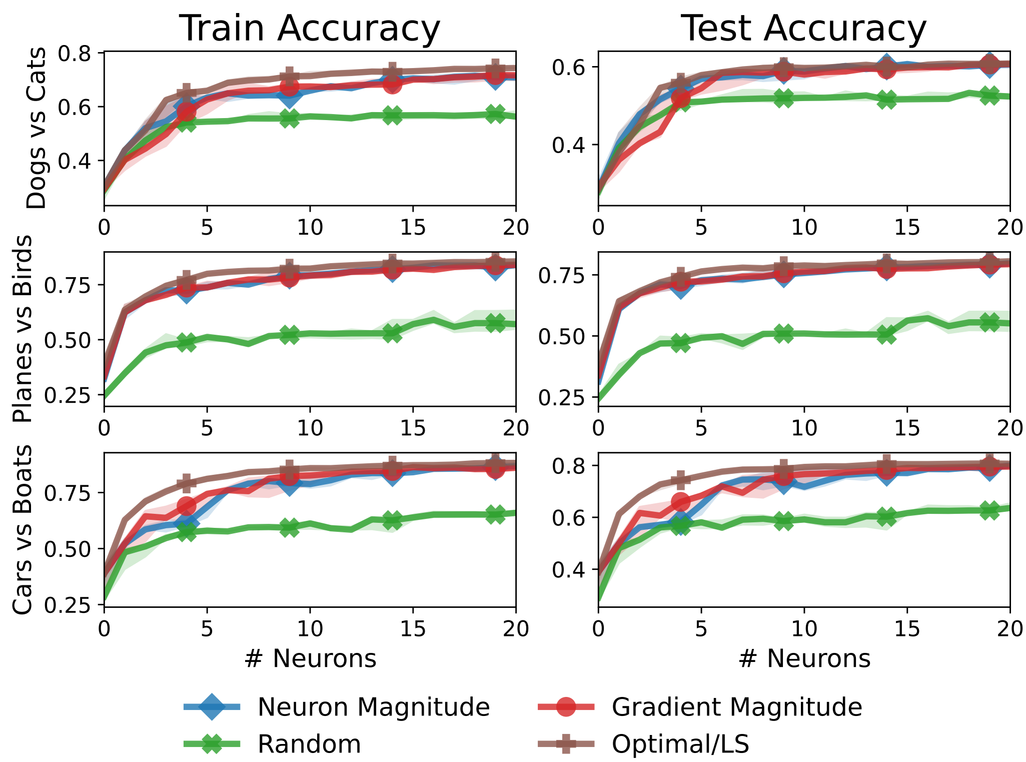

Figure 6. Pruning experiments on binary classification tasks from the CIFAR-10 dataset. This figure reproduces results from Figure 3 with training accuracies added and also includes results for an additional task, `cats vs dogs`, not presented in the main paper.

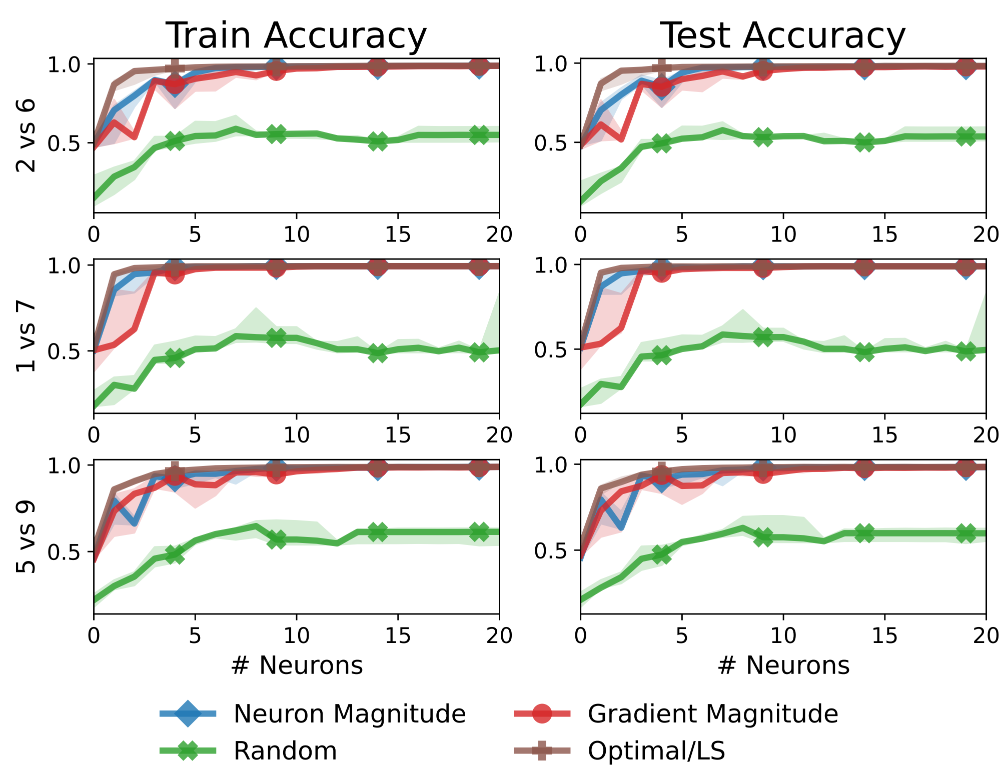

Figure 7. Pruning experiments on three binary classification tasks taken from MNIST.

Table 2. Tuning neural networks by searching over the optimal set. We fit two-layer ReLU networks on the training set and compute the minimum ℓ_2 norm solution (Min L_2). Then we tune by finding an extreme point approximating the maximum ℓ_2-norm solution (EP), minimizing validation MSE over the optimal set (V-MSE), and minimizing test MSE over the optimal set (T-MSE). Max Diff. reports the difference between the best and worse models found. For each method we give the median and interquartile range as `median (lower/upper)`.

Dataset	Min L_2	EP	V-MSE	T-MSE	Max Diff.
blood	0.72 (0.72/0.74)	0.72 (0.72/0.74)	0.62 (0.61/0.62)	0.7 (0.68/0.71)	0.1 (0.11/0.12)
breast-cancer	0.64 (0.61/0.65)	0.64 (0.61/0.65)	0.61 (0.6/0.64)	0.71 (0.66/0.71)	0.1 (0.06/0.08)
fertility	0.66 (0.62/0.7)	0.69 (0.62/0.69)	0.65 (0.64/0.7)	0.64 (0.57/0.64)	0.05 (0.06/0.06)
heart-hungarian	0.75 (0.7/0.77)	0.75 (0.7/0.77)	0.71 (0.56/0.72)	0.85 (0.82/0.86)	0.14 (0.26/0.14)
hepatitis	0.75 (0.74/0.78)	0.75 (0.74/0.78)	0.73 (0.69/0.75)	0.77 (0.77/0.9)	0.05 (0.08/0.15)
hill-valley	0.64 (0.64/0.65)	0.65 (0.64/0.65)	0.64 (0.64/0.67)	0.64 (0.64/0.65)	0.0 (0.0/0.01)
mammographic	0.77 (0.77/0.77)	0.77 (0.77/0.77)	0.57 (0.56/0.62)	0.78 (0.78/0.8)	0.21 (0.22/0.18)
monks-1	0.67 (0.64/0.71)	0.66 (0.64/0.71)	0.49 (0.48/0.51)	0.57 (0.51/0.61)	0.17 (0.15/0.2)
planning	0.53 (0.51/0.61)	0.52 (0.51/0.61)	0.53 (0.52/0.53)	0.7 (0.68/0.74)	0.17 (0.17/0.21)
spectf	0.64 (0.62/0.7)	0.64 (0.62/0.7)	0.56 (0.53/0.56)	0.58 (0.56/0.66)	0.08 (0.09/0.14)
horse-colic	0.75 (0.75/0.76)	0.59 (0.57/0.61)	0.74 (0.73/0.75)	0.85 (0.85/0.85)	0.26 (0.27/0.24)
ilpd-indian-liver	0.59 (0.57/0.6)	0.59 (0.57/0.6)	0.53 (0.53/0.57)	0.72 (0.7/0.73)	0.19 (0.17/0.17)
parkinsons	0.74 (0.72/0.74)	0.74 (0.72/0.74)	0.65 (0.65/0.74)	0.88 (0.86/0.9)	0.23 (0.21/0.16)
pima	0.68 (0.66/0.68)	0.68 (0.66/0.68)	0.68 (0.64/0.7)	0.87 (0.86/0.88)	0.2 (0.22/0.19)
tic-tac-toe	0.98 (0.98/0.98)	0.76 (0.69/0.8)	0.98 (0.98/0.99)	1.0 (1.0/1.0)	0.24 (0.31/0.2)
statlog-heart	0.71 (0.7/0.73)	0.71 (0.7/0.73)	0.7 (0.67/0.73)	0.84 (0.83/0.86)	0.14 (0.17/0.13)
ionosphere	0.85 (0.83/0.86)	0.76 (0.73/0.76)	0.84 (0.84/0.84)	0.88 (0.88/0.89)	0.12 (0.15/0.12)

D.2.1. TUNING

We select 17 binary classification datasets from the UCI repository. For each dataset we use a random 60/20/20 split of the data into train, validation, and test sets. We use the commercial interior point method MOSEK (ApS, 2022) through the interface provided by CVXPY (Diamond & Boyd, 2016) to compute the initial model which is then tuned. We modify the tolerances of this method to use $\tau = 10^{-8}$ for measuring both primal convergence and violation of the constraints. For each dataset, we use fixed $\lambda = 0.001$ and a maximum of 100 neurons. To compute the min ℓ_2-norm optimal model, we use the MOSEK and the optimization problem given in Proposition 3.17.

To approximate the maximum ℓ_2-norm model, we solve the following program:

$$\alpha^* = \arg\max_{\alpha \geq 0} \sum_{b_i \in \mathcal{S}_\lambda} \alpha_{b_i} \quad \text{s.t.} \sum_{b_i \in \mathcal{S}_\lambda} \alpha_{b_i} X_{b_i} v_{b_i} = \hat{y}.$$

This is a linear program which is straightforward to solve using interior point methods. Moreover, we have

$$\|w^*\|_2^2 = \lambda^2 \sum_{b_i} (\alpha^*_{b_i})^2,$$

so that $\sum_{b_i} \alpha_{b_i}$ acts as an approximation, where we recall $\alpha_{b_i} \geq 0$ necessarily.

To tune each model with respect to the validation/test MSE, we solve the following optimization problem:

$$\min_{w} \left\{ \frac{1}{2} \|\tilde{X} w - \tilde{y}\|_2^2 : w \in \mathcal{W}^*(\lambda) \right\},$$

with respect to the parameters of the convex formulation. Here, $(\tilde{X}, \tilde{y})$ is either the validation or test set. We repeat each experiment five times with different random splits of the data and random resamplings of 500 activation patterns D_i. This guarantees that each non-convex network has at most 1000 neurons after optimization, although it may have less due to the sparsity inducing penalty.

D.2.2. PRUNING

Methods: We use the augmented Lagrangian method of Mishkin et al. (2022) to compute the starting model which is then pruned. We modify the tolerances of this method to use $\tau = 10^{-8}$ for measuring both primal convergence and violation of the constraints.

Pruning by neuron magnitude is straightforward: we sort the neurons by $\|W_{1i} w_{2i}\|_2$, which measures the total magnitude of the neuron, and then drop the smallest one. For pruning by gradient norm, we compute $G_{1i} = \nabla_{W_{1i}} L(f_{W_1, w_2}(Z), y)$, $g_{2i} = \nabla_{W_{2i}} L(f_{W_1, w_2}(Z), y)$ and then score each neuron as

$$s_i = \|W_{1i} \cdot G_{1i} w_{2i} g_{2i}\|_2,$$

where $\cdot$ indicates the element-wise product. The neuron with smallest score is zeroed. This is consistent the existing implementations of pruning by gradient norm (Blalock et al., 2020) and attempts to measure the variation of a linearization of the loss in neuron i. For Random, we simply select a neuron from a uniform random distribution.

For Optimal/LS, we start by using Algorithm 1 to prune until no linear dependence exists amongst the neuron fits $D_i Z W_{1i}$. At this point, we choose β to minimize the squared-error in the training fit. We choose the neuron to prune by selecting the index that minimizing the residual in the least-squares fit,

$$i_k = \arg\min_j \min_\beta \frac{1}{2} \| \sum_{i \neq j} \beta_i D_i Z w_i - D_j Z w_j \|.$$

This produced the best result in all of our experiments, although you can also select i_k using neuron magnitude or any other rule in the literature on structured pruning.

UCI Datasets: We select 10 moderately-sized binary classification datasets from the UCI repository. For each dataset we use a random 50/50 split of the data into train and test sets, fixed $\lambda = 0.01$, and sample 25 activation patterns D_i. This results in a maximum of 50 neurons in each final model; since the datasets are low dimensional, randomly sampling activation patterns typically results in fewer than 50 neurons. All results are repeated for five different random splits and we plot the median and interquartile ranges of the results.

MNIST and CIFAR-10: We select three binary classification tasks from each dataset such that no task shares a target with another task. For each dataset we use a random 50/50 split of the data into train and test sets. For MNIST, we use $\lambda - 0.01$, while we used $\lambda = 0.05$ for CIFAR-10. We sample 50 activation patterns D_i for each tasking, which produces a maximum of 100 neurons in each final model. All results are repeated for five different random splits and we plot the median and interquartile ranges of the results.