

# Optimal Sets and Solution Paths of ReLU Networks

Aaron Mishkin<sup>1</sup> Mert Pilanci<sup>2</sup>

<sup>1</sup>Department of Computer Science, Stanford University

<sup>2</sup>Department of Electrical Engineering, Stanford University



## Introduction

**Main Contribution:** characterize all optimal two-layer ReLU nets.

- **Critical Points:** we extend our expression to all Clarke stationary points.
- **Uniqueness:** we give conditions for optima to be permutation-unique.
- **Pruning:** we show how to compute the “narrowest” optimal ReLU networks.

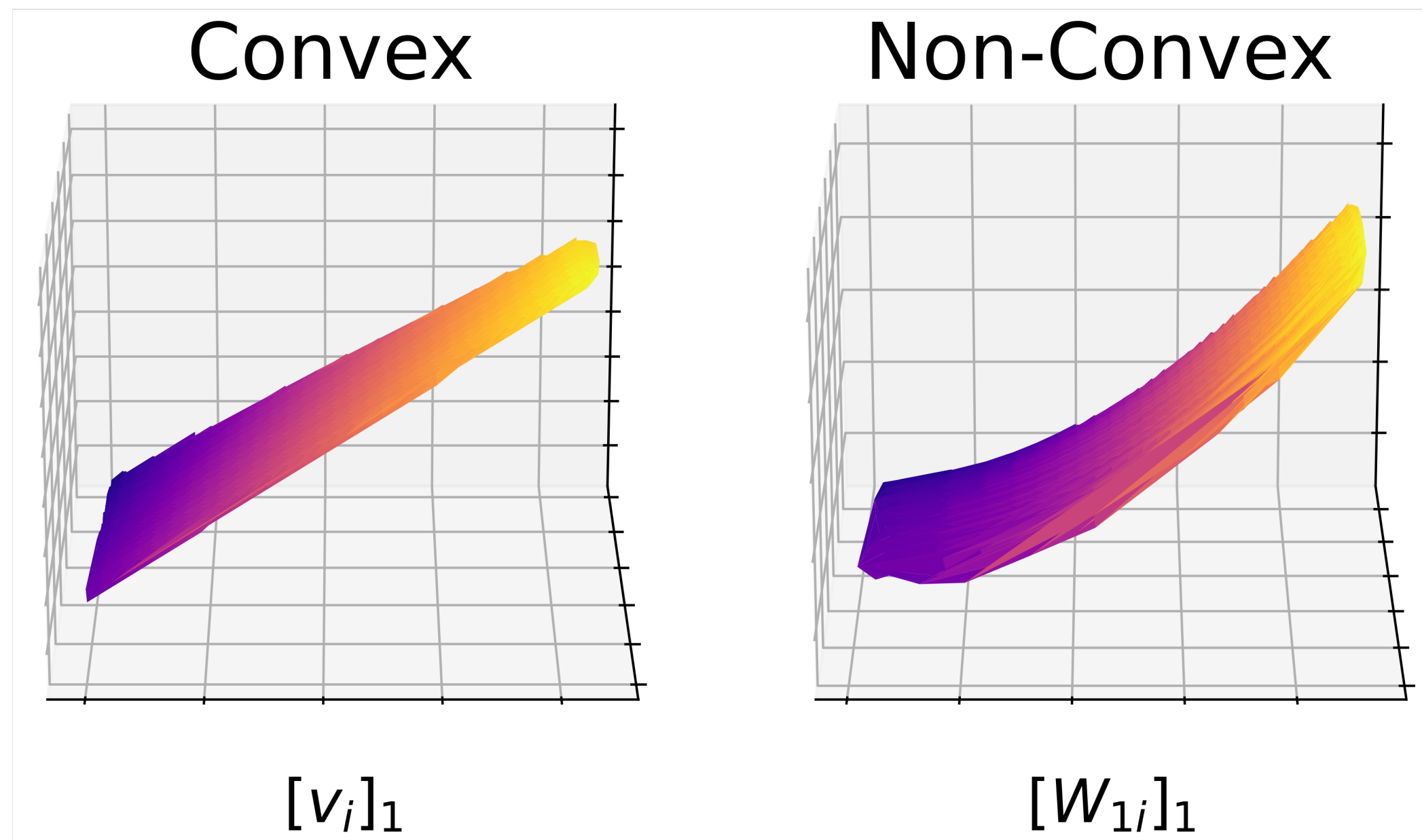


Figure 1. Optimal set for the first feature of three different neurons.

## Convex Reformulations: ReLU Networks

**Non-Convex Problem:**

$$\min_{\theta^1, \theta^2} \underbrace{\left\| \sum_{j=1}^m (X\theta_j^1)_+ \theta_j^2 - y \right\|_2^2}_{\text{Squared Error}} + \underbrace{\lambda \sum_{j=1}^m \|\theta_j^1\|_2^2 + \|\theta_j^2\|_2^2}_{\text{Weight Decay}},$$

where  $(x)_+ = \max\{x, 0\}$  is the ReLU activation.

**Convex Reformulation:** (Pilanci & Ergen, 2020)

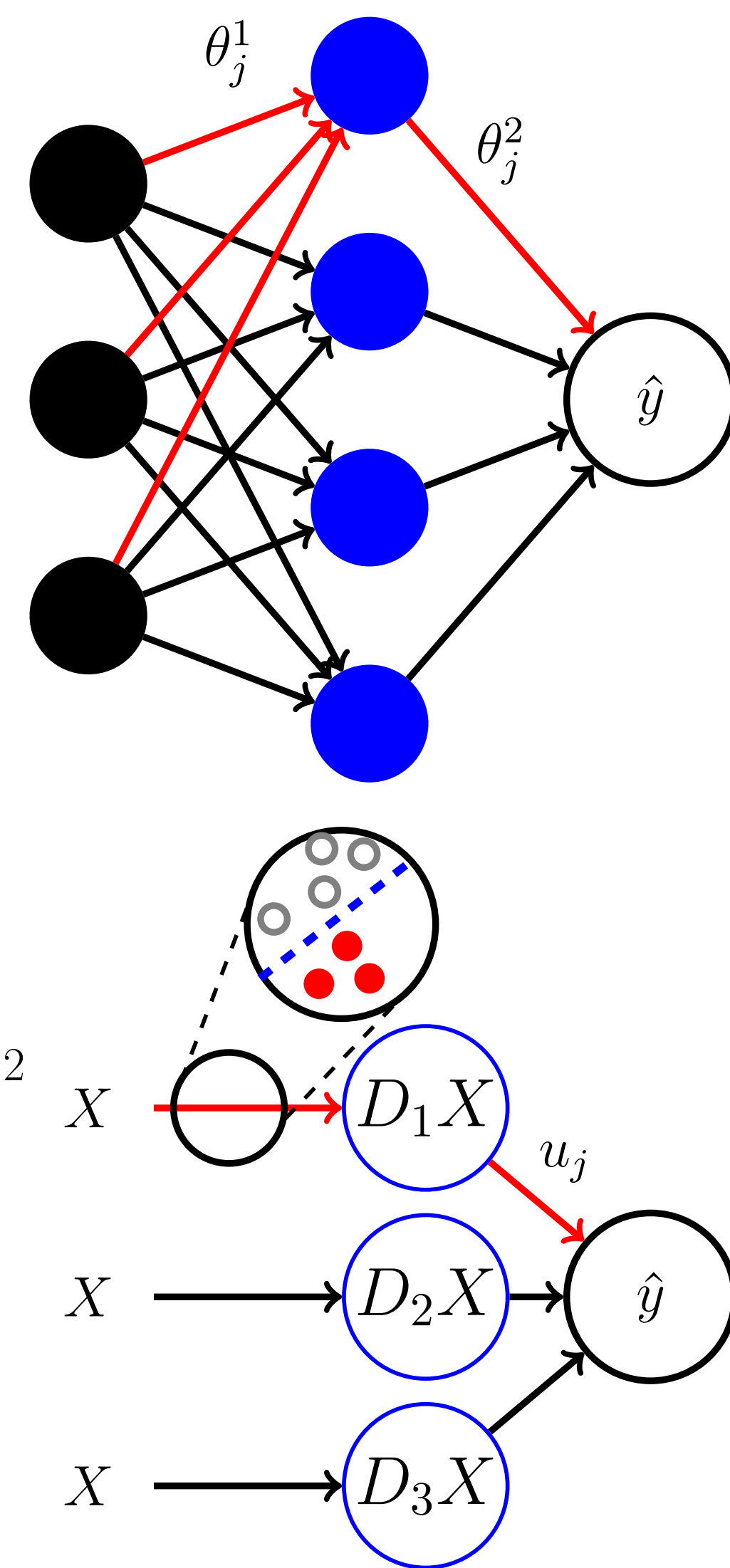
$$\min_{u, h} \left\| \sum_{j=1}^p D_j X(u_j - h_j) - y \right\|_2^2 + \lambda \sum_{j=1}^p \|u_j\|_2 + \|h_j\|_2$$

s.t.  $u_j, h_j \in \mathcal{K}_j := \{w : (2D_j - I)Xw \geq 0\},$

where  $D_j = \text{diag}[\mathbf{1}(Xg_j \geq 0)]$ .

**Equivalence:**

- If  $m \geq m^*$ , where  $m^* \leq n$ , then both programs have the same optimal value.
- The convex and non-convex solutions are related by a **solution mapping**.



## Proof Strategy

**Approach:** We use convex reformulations as an analytical tool!

1. Re-write convex reformulation as a **constrained group lasso (CGL)** problem,

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

s.t.  $K_{b_i}^\top w_{b_i} \leq 0$  for all  $b_i \in \mathcal{B}$ .

2. Derive optimal set for CGL using the **KKT conditions**: if  $w_{b_i} \neq 0$ ,

$$\underbrace{Z_{b_i}^\top (y - Zw)}_{v_{b_i}} - K_{b_i} \rho_{b_i} = \lambda \frac{w_{b_i}}{\|w_{b_i}\|_2}$$

3. Obtain ReLU optimal set using the **solution mapping**: for  $s_i \in \{+1, -1\}$ ,

$$\theta_i^1 = w_{b_i}^* / \sqrt{\|w_{b_i}^*\|}, \quad \theta_i^2 = s_i \cdot \sqrt{\|w_{b_i}^*\|}$$

## The ReLU Optimal Set

Define the set of blocks supported by a solution to the convex reformulation:

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

and let  $Zw^* = \hat{y}$  be the (unique) optimal model fit.

**Corollary 4.1** (informal): Suppose  $m \geq m^*$  and  $\lambda > 0$ . Then the set of optimal two-layer ReLU MLPs is,

$$\mathcal{O}_\lambda = \{(W_1, w_2) : f_{\theta^1, \theta^2}(X) = \hat{y},$$

$$\theta_i^1 = \left(\frac{\alpha_i}{\lambda}\right)^{\frac{1}{2}} v_i, \theta_i^2 = (\alpha_i \lambda)^{\frac{1}{2}},$$

$$\alpha_i \geq 0, i \in [2p] \setminus \mathcal{S}_\lambda \Rightarrow \alpha_i = 0\},$$

**Interpretation:** optimal neurons are scalings of the correlation vector  $v_{b_i}$  on the manifold of optimal predictors.

## Conditions for Uniqueness

**Q:** When are optimal networks unique up to permutations/splits?

**A:** When the convex reformulation has a unique solution!

**Proposition 4.3** (informal): Suppose  $m \geq m^*$  and  $\lambda > 0$ . If there does not exist  $\alpha \neq 0$  such that

$$\sum_{i \in \mathcal{S}_\lambda} \alpha_i (X\theta_i^1)_+ = 0,$$

then the non-convex solution is permutation/splitting unique (p-unique).

**Continuity:** p-uniqueness on a neighbourhood  $\mathcal{N}$  implies continuity of at least one regularization path  $\lambda \mapsto (\theta^1(\lambda), \theta^2(\lambda))$  on  $\mathcal{N}$ .

## Optimal Pruning

**Definition:** a ReLU network is **minimal** if there does not exist an optimal network with strictly fewer non-zero neurons.

**Proposition 3.6** (informal): Let  $m \geq m^*$  and  $\lambda > 0$ . A model is minimal if and only if the non-zero activations  $(X\theta_i^1)_+$  are linearly independent.

**Algorithm 1** Optimal Pruning

**Input:** data matrix  $X$ , solution  $w^0$ .  
**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**  
 $b_i^k \leftarrow \arg \max_{b_i} \{|\beta_{b_i}|\} : b_i \in \mathcal{A}_\lambda(w^k)\}$   
 $t^k \leftarrow 1/|\beta_{b_i^k}|$   
 $w^{k+1} \leftarrow w^k(1 - t^k \beta_{b_i^k})$   
**end while**  
**Output:** final weights  $w^k$

**Proposition** (informal): Algorithm 1 computes an optimal and minimal ReLU network with  $m^* \leq n$  non-zero neurons in  $O(n^3m + nd)$  time.

**Direct Pruning:** We give equivalent algorithm for non-convex params.

## Experiments

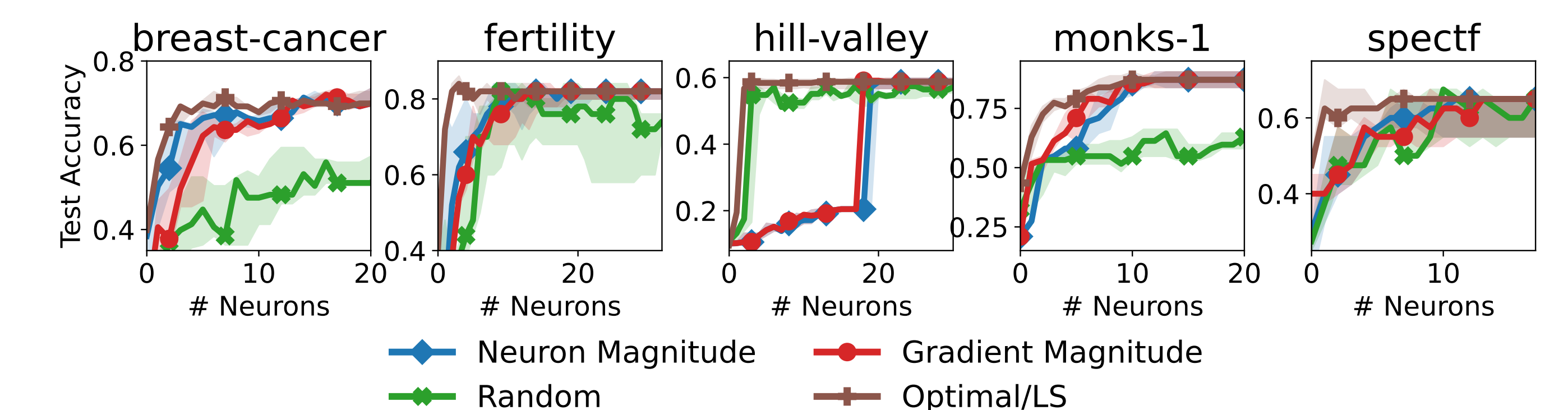


Figure 2. Test accuracy for theory-inspired pruning (Optimal/LS) and baseline methods.

- We consider pruning neurons on several UCI classification datasets.
- Our approach **dominates** every baseline considered.

Dataset	Min L <sub>2</sub>	EP	V-MSE	T-MSE	Max Diff.
mammogr.	0.77	0.77	0.57	0.78	<b>0.21</b>
horse-colic	0.75	0.59	0.74	0.85	<b>0.26</b>
ilpd-indian	0.59	0.59	0.53	0.72	<b>0.19</b>
parkinsons	0.74	0.74	0.65	0.88	<b>0.23</b>
pima	0.68	0.68	0.68	0.87	<b>0.2</b>

- We tune ReLU networks by direct optimization over the optimal set.
- Same training performance, but test accuracy differs by over **20 points!**

## References

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