Fast Convex Optimization for Two-Layer ReLU Networks:
Equivalent Model Classes and Cone Decompositions

Aaron Mishkin [1] Arda Sahiner [2] Mert Pilanci [2]

[1]Department of Computer Science, Stanford University [2]Department of Electrical Engineering, Stanford University

Introduction

Problem: optimizing neural networks with stochastic gradient methods is hard.

- **Tuning**: good performance requires tuning the step-size, momentum, ...
- **Model Churn**: changing extrinsic parameters like random seed affects model performance (Henderson et al., 2018).
- **Certificates**: convergence to stationary point, but only with decreasing step-sizes.

Approach: train two-layer models by reformulating them as convex programs.

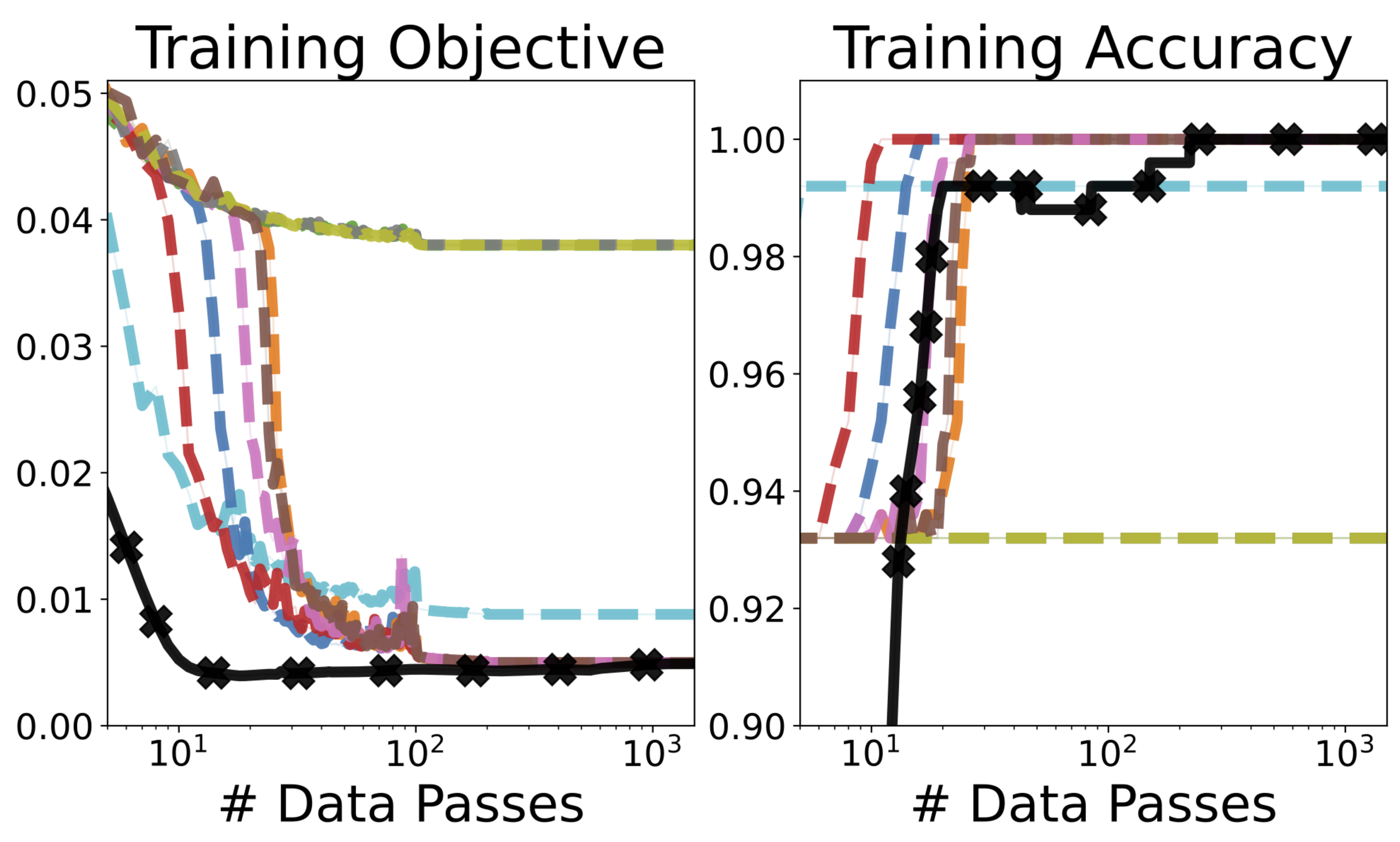


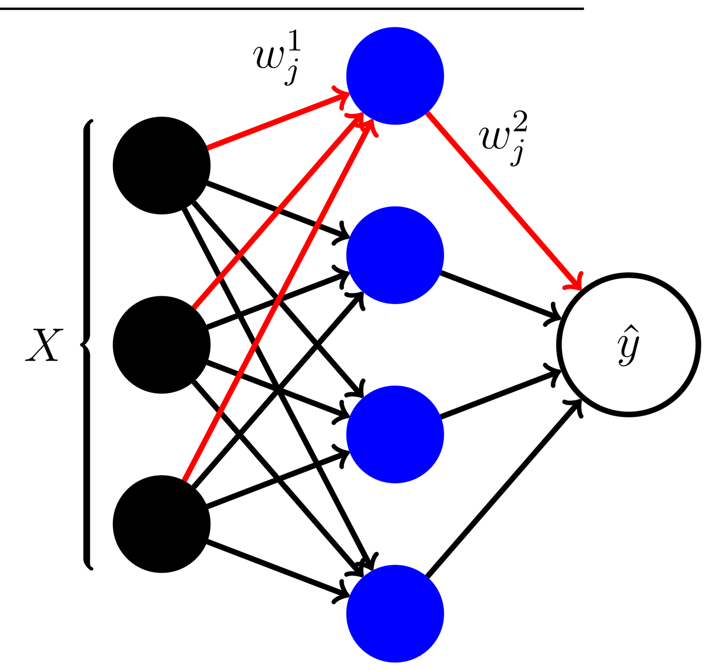
Figure 1. Effect of random seed on convergence of SGD for a realizable problem.

Convex Reformulations: ReLU Networks

Non-Convex Problem:

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

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

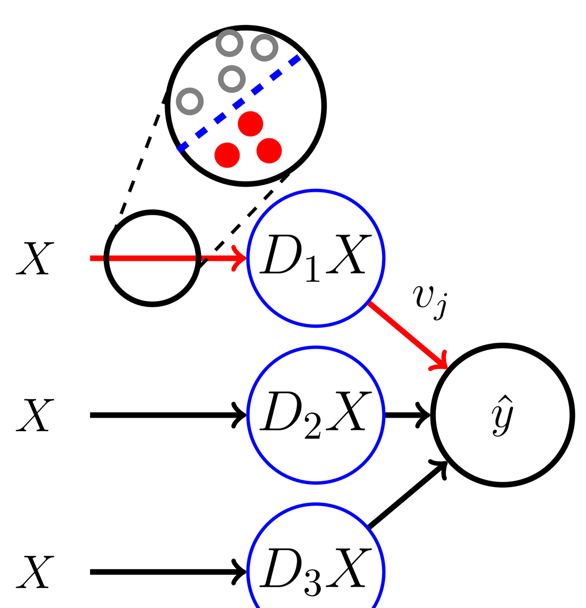


Convex Reformulation: (Pilanci & Ergen, 2020)

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

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

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



Problem Scale

The **convex program** enumerates all activation patterns,

$$p = \left| \left\{ D_j = \text{diag}[\mathbb{1}(Xg_j \geq 0)] : g_j \in \mathbb{R}^d \right\} \right|.$$

- Exponential in general: $p \in O(r \cdot (\frac{n}{r})^r)$, where $r = \text{rank}(X)$ (Winder, 1966).
  - But sub-sampling works well in practice.
- Highly structured — it's a huge-scale constrained generalized linear model!
- The convex reformulation exchanges one kind of hardness for another.

Gated ReLU Activations

Issue: convex problem is too large for IPMs, but projecting onto $\mathcal{K}_j$ is an LP.

Solution: consider unconstrained relaxation:

$$\textbf{C-GReLU}: \min_u \left\| \sum_{j=1}^p D_j X u_j - y \right\|_2^2 + \lambda \sum_{j=1}^p \|u_j\|_2$$

Theorem 2.2 (informal): C-GReLU is equivalent to solving

$$\textbf{NC-GReLU}: \min_{w^1,w^2} \frac{1}{2} \left\| \sum_{j=1}^p \phi_{g_j}(X, w_j^1) w^2 - y \right\|_2^2 + \frac{\lambda}{2} \sum_{j=1}^p \|w_j^1\|_2^2 + \|w_j^2\|_2^2,$$

with the "Gated ReLU" (Fiat et al., 2019) activation function

$$\phi_g(X, u) = \text{diag}(\mathbb{1}(Xg \geq 0))Xu,$$

and gate vectors g_j such that $D_j = \text{diag}[\mathbb{1}(Xg_j \geq 0)]$.

Interpretation: if $u_j \notin \mathcal{K}_j$, then activation must be decoupled from weights.

Cone Decompositions

Q: when are Gated ReLU and ReLU networks equivalent?

A: if we can decompose $u_j = v_j - h_j$ for some $v_j, h_j \in \mathcal{K}_j$.

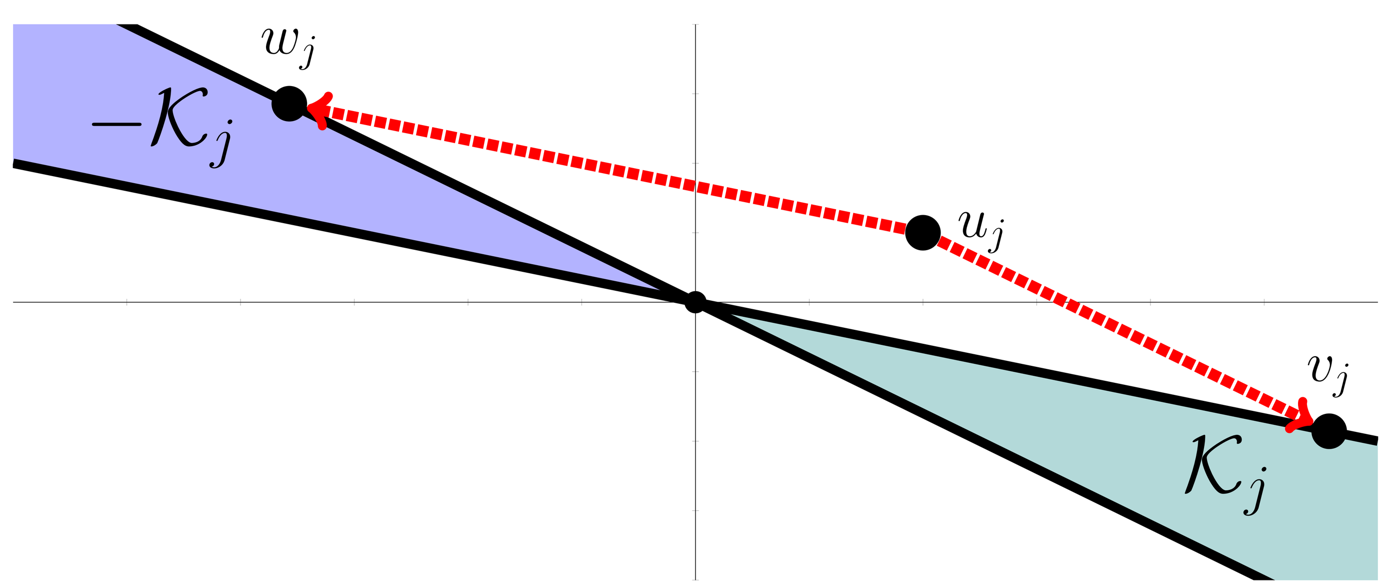


Figure 2. Illustration of cone decomposition procedure.

Existence of "cone decompositions":

- We can guarantee $\mathcal{K}_j - \mathcal{K}_j = \mathbb{R}^d$ if X is full row-rank.
- More generally, we show "flat" $\mathcal{K}_j$ can be safely merged into other neurons.

Main Approximation Result

Theorem 3.7 (informal): Let $\lambda \geq 0$ and let p^* be the optimal value of the ReLU problem. There exists a C-GReLU problem with minimizer u^* and optimal value d^* satisfying,

$$d^* \leq p^* \leq d^* + 2\lambda\kappa(\tilde{X}_{\mathcal{J}}) \sum_{D_i \in \tilde{\mathcal{D}}} \|u_i^*\|_2.$$

Consequence: ReLU and Gated ReLU model classes are equivalent!

Algorithms

We develop two algorithms for solving the convex reformulations:

- **R-FISTA**: a restarted FISTA variant for Gated ReLU.
- **AL**: an augmented Lagrangian method for the (constrained) ReLU Problem.

And we can use all the convex tricks!

- **Fast**: $O(1/t^2)$ convergence rate.
- **Tuning-free**: line-search, restarts, data normalization, ...
- **Certificates**: termination based on minimum-norm subgradient.

Experiments

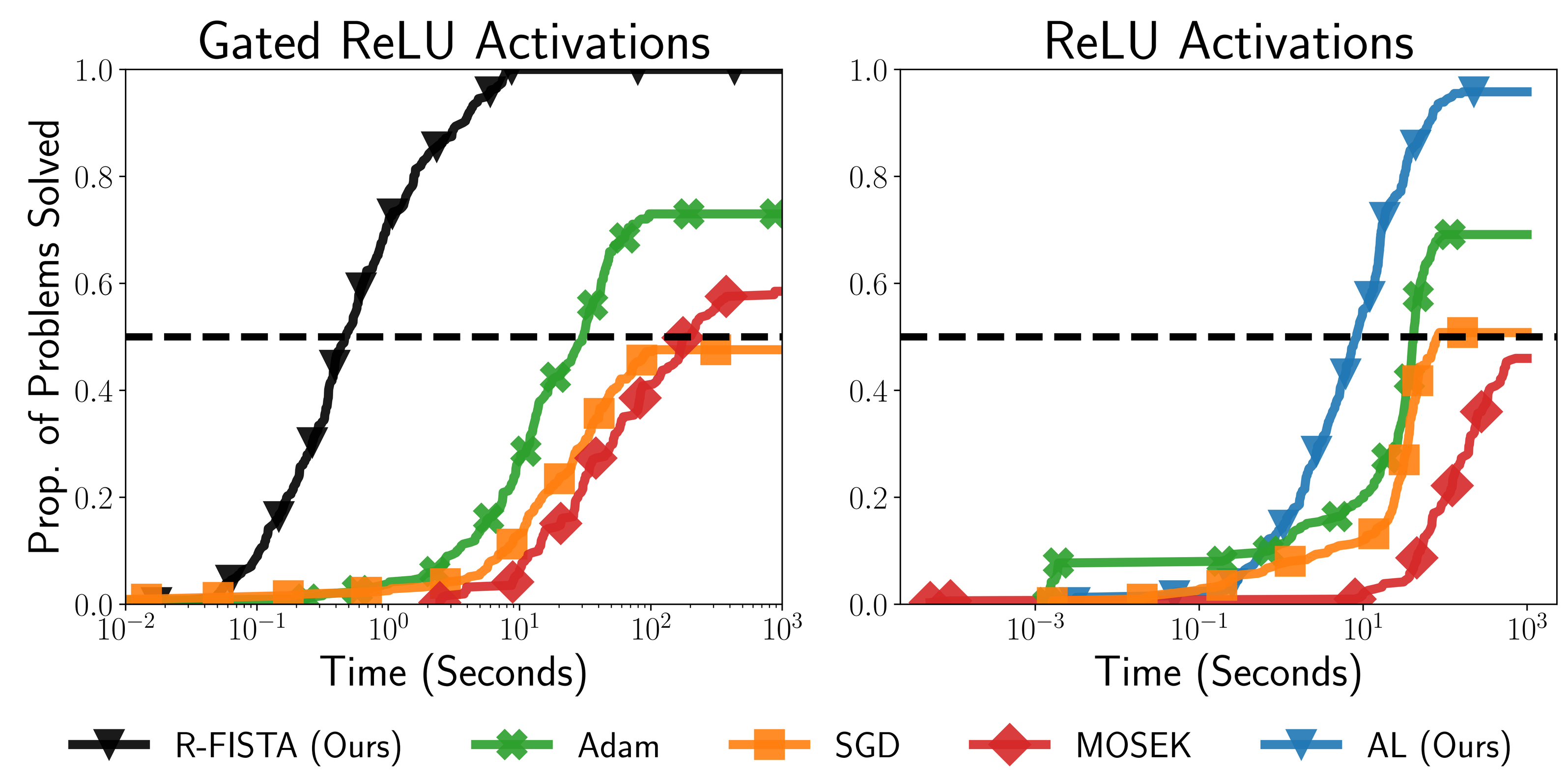


Figure 3. Performance profile comparing convex solvers to Adam and SGD.

- Performance on 438 training problems generated from the UCI repository.
- R-FISTA/AL solve **more problems, faster**, than SGD and Adam.

References

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

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *(AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3207–3214. AAAI Press, 2018.

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.

Winder, R. O. Partitions of n-space by hyperplanes. *SIAM Journal on Applied Mathematics*, 14(4):811–818, 1966.