Stanford University

# Fast Convex Optimization for Two-Layer ReLU Networks:

## Equivalent Model Classes and Cone Decompositions

Aaron Mishkin    Arda Sahiner    Mert Pilanci

# Overview

Problem: Training shallow neural networks is challenging.

# Overview

Problem: Training shallow neural networks is challenging.

- **Tuning**: step-size and other hyper-parameters must be tuned.

# Overview

Problem: Training shallow neural networks is challenging.

- **Tuning**: step-size and other hyper-parameters must be tuned.
- **Model Churn**: models trained with different random seeds have different performance [Hen+18].

# Overview

Problem: Training shallow neural networks is challenging.

- **Tuning**: step-size and other hyper-parameters must be tuned.
- **Model Churn**: models trained with different random seeds have different performance [Hen+18].
- **Certificates**: final models have few guarantees.

# Overview

Problem: Training shallow neural networks is challenging.

- **Tuning**: step-size and other hyper-parameters must be tuned.
- **Model Churn**: models trained with different random seeds have different performance [Hen+18].
- **Certificates**: final models have few guarantees.

---

Our Contribution: robust training by convex reformulations.

# Overview

Problem: Training shallow neural networks is challenging.

- **Tuning**: step-size and other hyper-parameters must be tuned.
- **Model Churn**: models trained with different random seeds have different performance [Hen+18].
- **Certificates**: final models have few guarantees.

---

Our Contribution: robust training by convex reformulations.

- We develop new convex reformulations of two-layer neural networks with **gated ReLU** activations.

# Overview

Problem: Training shallow neural networks is challenging.

- **Tuning**: step-size and other hyper-parameters must be tuned.
- **Model Churn**: models trained with different random seeds have different performance [Hen+18].
- **Certificates**: final models have few guarantees.

---

Our Contribution: robust training by convex reformulations.

- We develop new convex reformulations of two-layer neural networks with **gated ReLU** activations.
- We show how to approximate the ReLU training problem by **unconstrained** convex optimization of a Gated ReLU network.

# Overview

Problem: Training shallow neural networks is challenging.

- **Tuning**: step-size and other hyper-parameters must be tuned.
- **Model Churn**: models trained with different random seeds have different performance [Hen+18].
- **Certificates**: final models have few guarantees.

---

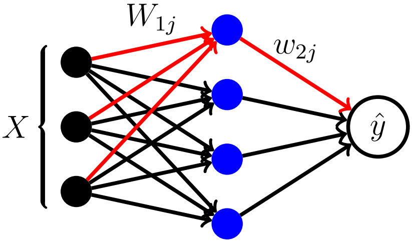
Our Contribution: robust training by convex reformulations.

- We develop new convex reformulations of two-layer neural networks with **gated ReLU** activations.
- We show how to approximate the ReLU training problem by **unconstrained** convex optimization of a Gated ReLU network.
- We propose and **exhaustively evaluate** algorithms for solving our convex reformulations.

# Background on Convex Reformulations

## Non-Convex Problem

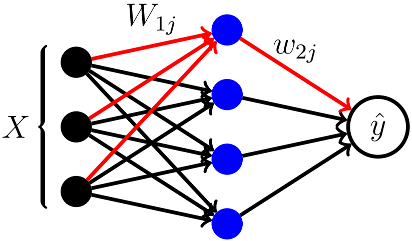
$$\min_{W} \| \sum_{j=1}^{m} (XW_{1j})_+ w_{2j} - y\|_2^2 + \lambda \sum_{j=1}^{m} \|W_{1j}\|_2^2 + \|w_{2j}\|^2$$



# Background on Convex Reformulations

## Non-Convex Problem

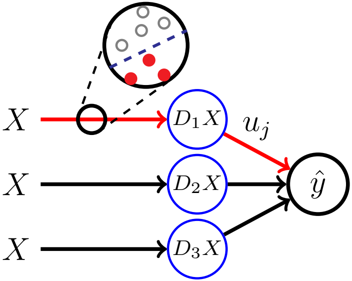
$$\min_{W} \|\sum_{j=1}^{m} (XW_{1j})_+ w_{2j} - y\|_2^2 + \lambda \sum_{j=1}^{m} \|W_{1j}\|_2^2 + \|w_{2j}\|^2$$




## Convex Reformulation [PE20]

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

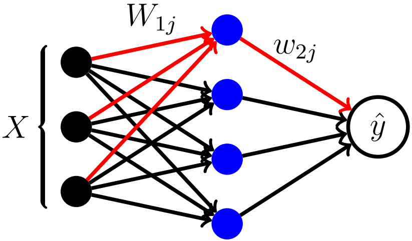
s.t. $v_j, w_j \in \mathcal{K}_j$ for $j = 1, \ldots, p$.



# Background on Convex Reformulations

## Non-Convex Problem

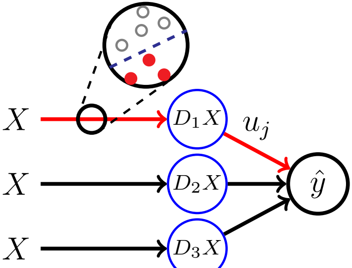
$$\min_{W} \|\sum_{j=1}^{m}(XW_{1j})_+ w_{2j} - y\|_2^2 + \lambda \sum_{j=1}^{m} \|W_{1j}\|_2^2 + \|w_{2j}\|^2$$




## Convex Reformulation [PE20]

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

s.t. $v_j, w_j \in \mathcal{K}_j$ for $j = 1, \ldots, p$.



# Gated ReLU Networks

$$\text{C-ReLU} : \min_{v,w} \| \sum_{j=1}^{p} D_j X (v_j - w_j) - y \|_2^2 + \lambda \sum_{j=1}^{p} \|v_j\|_2 + \|w_j\|_2,$$

$$\text{s.t. } v_j, w_j \in \mathcal{K}_j \text{ for } j = 1, \ldots, p.$$

# Gated ReLU Networks

$$\text{C-ReLU} : \min_{v,w} \|\sum_{j=1}^{p} D_j X(v_j - w_j) - y\|_2^2 + \lambda \sum_{j=1}^{p} \|v_j\|_2 + \|w_j\|_2,$$

$$\text{s.t. } v_j, w_j \in \mathcal{K}_j \text{ for } j = 1, \ldots, p.$$

---

$$\text{C-GReLU} : \min_{u} \|\sum_{j=1}^{p} D_j X u_j - y\|_2^2 + \lambda \sum_{j=1}^{p} \|u_j\|_2,$$

# Gated ReLU Networks

$$\text{C-ReLU} : \min_{v,w} \| \sum_{j=1}^{p} D_j X(v_j - w_j) - y\|_2^2 + \lambda \sum_{j=1}^{p} \|v_j\|_2 + \|w_j\|_2,$$

$$\text{s.t. } v_j, w_j \in \mathcal{K}_j \text{ for } j = 1, \ldots, p.$$

---

$$\text{C-GReLU} : \min_{u} \| \sum_{j=1}^{p} D_j X u_j - y\|_2^2 + \lambda \sum_{j=1}^{p} \|u_j\|_2,$$

**Prop.** (informal): C-GReLU is equivalent to a "gated ReLU" network [FMS19] with activation function

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

# Gated ReLU: Cone Decompositions

- We reparameterized as $u_j = v_j - w_j$.

# Gated ReLU: Cone Decompositions

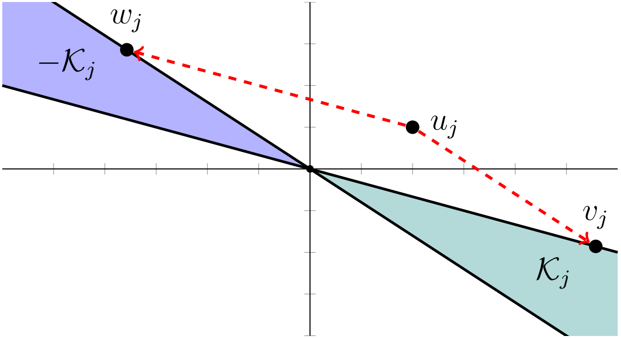
- We reparameterized as $u_j = v_j - w_j$.
- Given, $u_j$, can we go back to $v_j - w_j$?

# Gated ReLU: Cone Decompositions

- We reparameterized as $u_j = v_j - w_j$.
- Given, $u_j$, can we go back to $v_j - w_j$?
- That is, when does $\mathcal{K}_j - \mathcal{K}_j$ span $\mathbb{R}^d$?

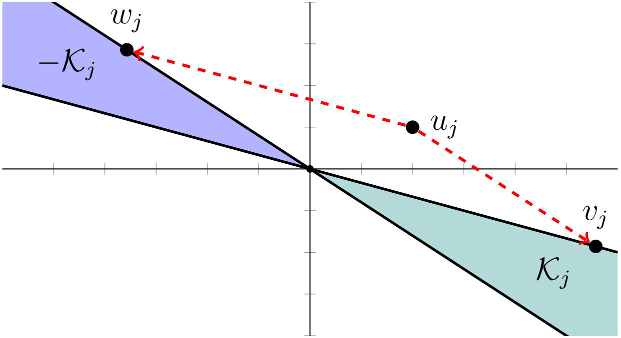
# Gated ReLU: Cone Decompositions

- We reparameterized as $u_j = v_j - w_j$.
- Given, $u_j$, can we go back to $v_j - w_j$?
- That is, when does $\mathcal{K}_j - \mathcal{K}_j$ span $\mathbb{R}^d$?



# Gated ReLU: Cone Decompositions

- We reparameterized as $u_j = v_j - w_j$.
- Given, $u_j$, can we go back to $v_j - w_j$?
- That is, when does $\mathcal{K}_j - \mathcal{K}_j$ span $\mathbb{R}^d$?




**Informal Result**: $\mathcal{K}_j - \mathcal{K}_j = \mathbb{R}^d$ or $\mathcal{K}_j$ is "unimportant".

# Main Approximation Result

We can decompose a Gated ReLU neuron into two ReLU neurons.

# Main Approximation Result

We can decompose a Gated ReLU neuron into two ReLU neurons.

Theorem (Approximation by Cone Decomposition)

*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,*

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

# Main Approximation Result

We can decompose a Gated ReLU neuron into two ReLU neurons.

Theorem (Approximation by Cone Decomposition)

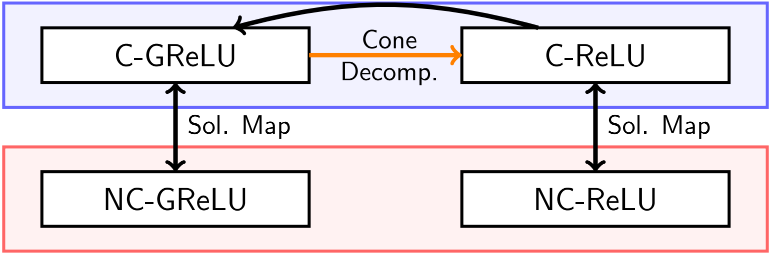
*Let $\lambda \geq 0$ and let $p^*$ be the optimal value of the ReLU problem. There exists a C-GReLU problem with minimizer $u^*$ and optimal value $d^*$ satisfying,*

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

Additional Consequences

- The approximation is exact for **unregularized** models!
- The Gated ReLU and ReLU models are formally **equivalent**!

# Solving the Convex Programs

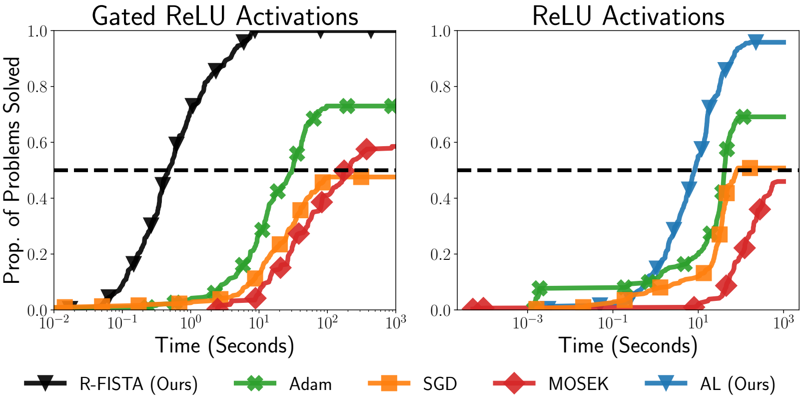



We develop two algorithms for solving the convex reformulations:

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

Our work exhaustively evaluates the performance of R-FISTA and AL.

# Numerical Results




- Generated by 438 training problems taken from UCI repo.
- R-FISTA/AL solve more, faster, than SGD and Adam.

Thanks for Listening!

# References I


Jonathan Fiat, Eran Malach, and Shai Shalev-Shwartz. "Decoupling gating from linearity". In: *arXiv preprint arXiv:1906.05032* (2019).

Peter Henderson et al. "Deep Reinforcement Learning That Matters". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 2018, pp. 3207–3214.

Mert Pilanci and Tolga Ergen. "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, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. 2020, pp. 7695–7705.