Stanford University

Fast Convex Optimization for Two-Layer ReLU Networks:

Equivalent Model Classes and Cone Decompositions

Aaron Mishkin
amishkin@cs.stanford.edu


Motivation: Neural Networks

Generated by DALL·E 2

A bowl of soup that is a portal to another dimension as digital art.

https://openai.com/dall-e-2/

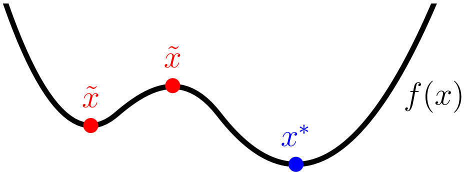
Motivation: Non-Convex Optimization

DALL·E 2 has 5.5 billion parameters and took billions of iterations to fit [1].

Motivation: Non-Convex Optimization

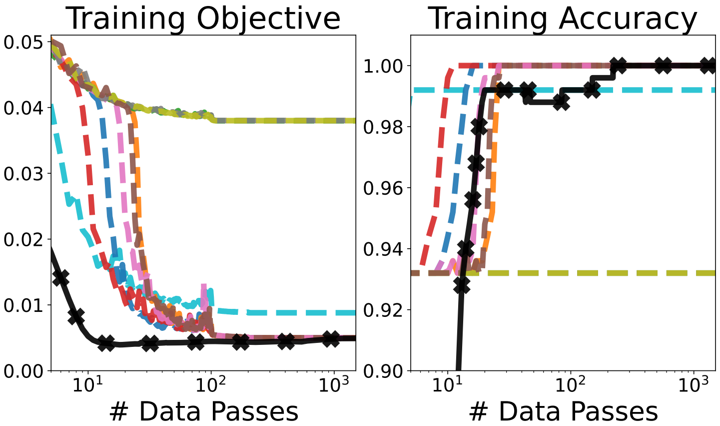
DALL·E 2 has 5.5 billion parameters and took billions of iterations to fit [1].

Main Challenge: neural networks are non-convex!



Motivation: Training Pathologies

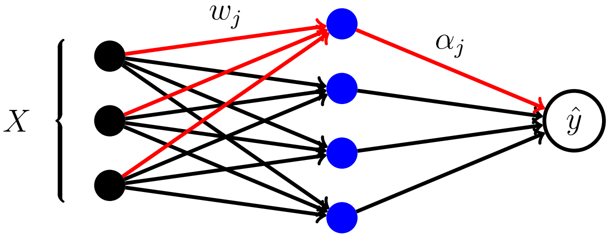
Optimizing neural networks with stochastic gradient descent is hard!



Convex Reformulations: Non-Convex Problem

Non-Convex Problem

$$\min_{w,\alpha} \|\sum_{j=1}^{m} (Xw_j)_+ \alpha_j - y\|_2^2 + \lambda \sum_{j=1}^{m} \|w_j\|_2^2 + |\alpha_j|^2$$

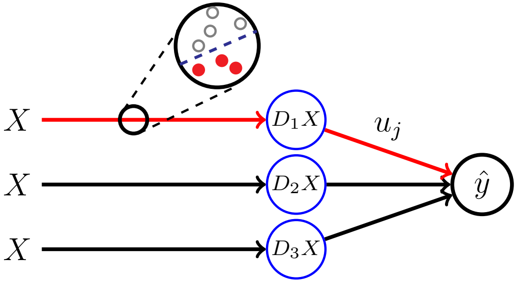


Convex Reformulations: Convex Problem

Convex Reformulation

$$\min_u \| \sum_{j=1}^{p} D_j X u_j - y\|_2^2 + \lambda \sum_{j=1}^{p} \|u_j\|_2,$$

$$\text{where } D_j = \mathrm{diag}[\mathbb{1}(X g_i \geq 0)]$$



Convex Reformulations: A Huge-Scale Linear Model

Convex Form : $$\min_u \|\sum_{j=1}^{p} D_j X u_j - y\|_2^2 + \lambda \sum_{j=1}^{p} \|u_j\|_2,$$

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

- Exponential in general: $p \in O(r \cdot (\frac{n}{r})^r)$, where $r = \text{rank}(X)$.
- But the problem is highly structured!

Convex Reformulations: Performance

Fast solvers use **numerical tricks** and **hardware acceleration**:

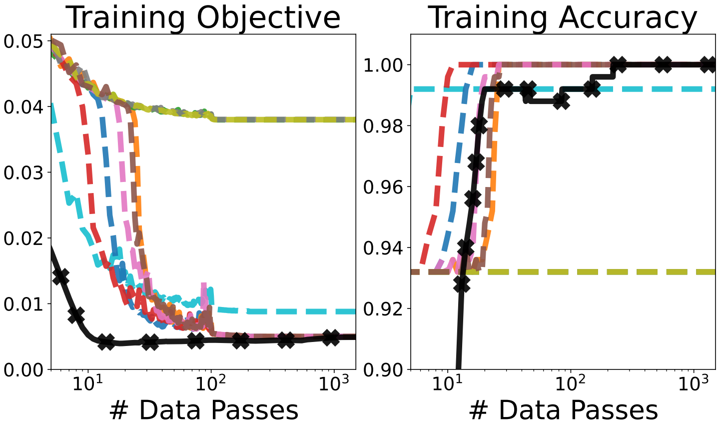
- Classic Tricks: faster convergence via line-search, restarts, ...
- CUDA GPUs: $70\times$ faster Mat-Vec operations using `float32`.

Scaling is a still a problem!

- Dense operations on GPUs are faster than sparse computations.
- GPU memory is limited — can we use multi-GPU programming models?

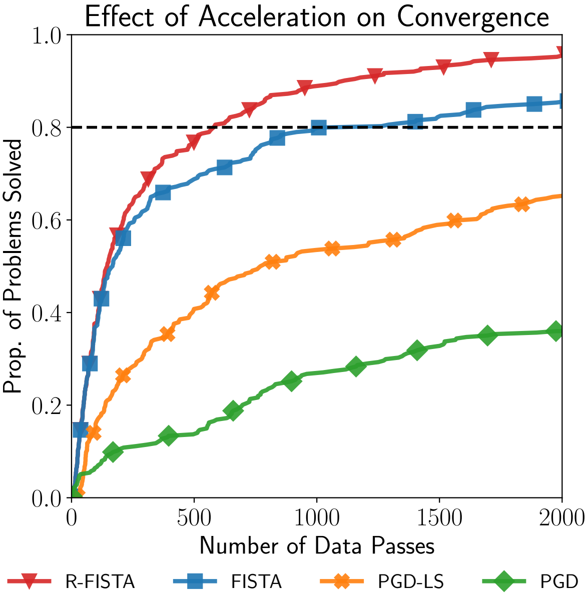
Results: Convex vs Non-Convex

The convex solver is stable!

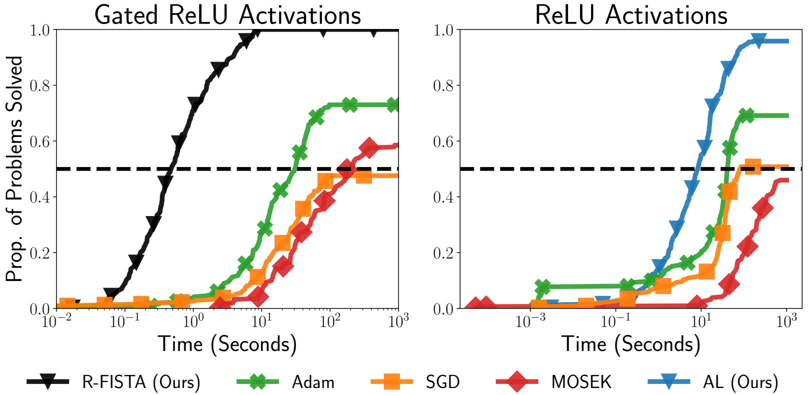


Thanks for Listening!

Bonus: Numerical Results (1)



Bonus: Numerical Results (2)



References I

Aditya Ramesh et al. “Hierarchical Text-Conditional Image Generation with CLIP Latents”. In: *CoRR* abs/2204.06125 (2022). arXiv: 2204.06125.