The Hidden Convex Optimization Landscape of Two-Layer ReLU Networks

Victor Mercklé[1], Franck Iutzeler[2], Ievgen Redko[3]

[1] LJK, Univ. Grenoble Alpes [2] IMT, Univ. Toulouse, CNRS [3] Paris Noah's Ark lab

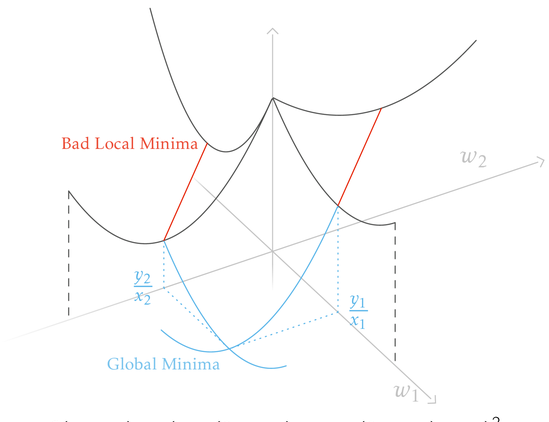
Online Blog Post

TL;DR

- Training a ReLU network is a **non-convex optimization** problem with many local minima which can be **hard** to escape in practice
- The **global minimum** is the unique solution of a derived **convex** optimization problem
- **Convex** problems have more predictable dynamics, and are **easier** to solve

Is this actually convex?

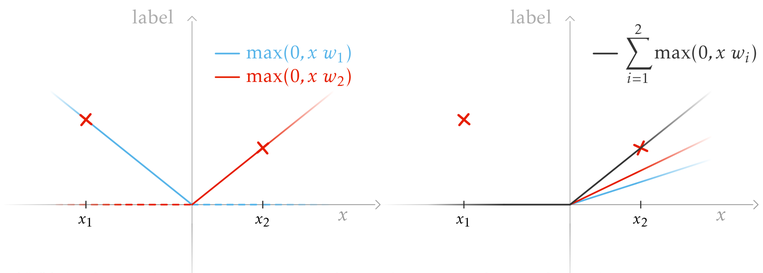
Loss landscape of a two neuron ReLU Network:



$$\mathcal{L}(w_1, w_2) = (\max(0, x_1 w_1) + \max(0, x_1 w_2) - y_1)^2 + (\max(0, x_2 w_1) + \max(0, x_2 w_2) - y_2)^2$$

- Two ✗ data points $x_1, x_2 \in \mathbb{R}$ with labels $y_1, y_2 \in \mathbb{R}$
- Two ReLU neurons w_1, $w_2 \in \mathbb{R}$, and second layer fixed to 1

Outputs of two networks found by gradient descent:

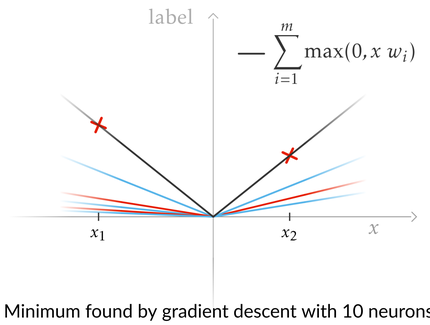


- $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ is the activation pattern of the red neuron that only activate x_2
- The blue neuron has a different activation pattern in the two local minima

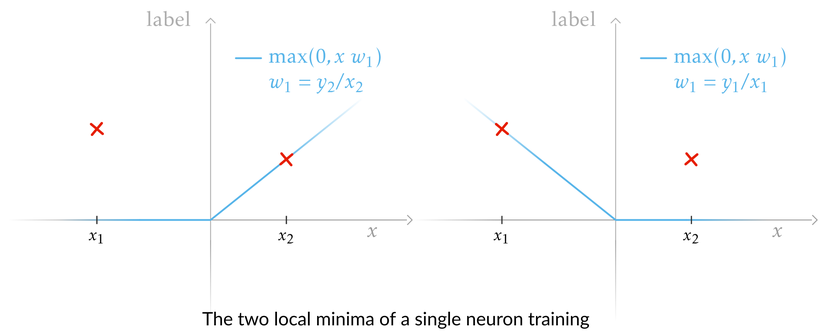
Equivalent convex problem:

$$\mathcal{L}(u_1, u_2) = \left\| \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} u_1 + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} u_2 - \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right\|_2^2$$

- This convex loss has the same optimal as the non convex loss for any $m \geq 2$
- We can get the two optimal neurons from the convex solution



Minimum found by gradient descent with 10 neurons



The two local minima of a single neuron training

General Case and Result

Setting

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times m}, \boldsymbol{\alpha} \in \mathbb{R}^m} \left\| \sum_{i=1}^{m} \max(0, \boldsymbol{X}\boldsymbol{w}_i)\alpha_i - \boldsymbol{y} \right\|_2^2 + \lambda \sum_{i=1}^{m} \|\boldsymbol{w}_i\|_2^2 + \alpha_i^2 \qquad (1)$$

- n samples: $\boldsymbol{x}_j \in \mathbb{R}^d$(rows of $\boldsymbol{X}$ with labels $y_j \in \mathbb{R}$, $j = 1, \dots n$
- m neurons: First layer $\boldsymbol{w}_i \in \mathbb{R}^d$, second layer $\alpha_i \in \mathbb{R}$, $i = 1, \dots m$
- $\lambda \geq 0$ regularization, $\gamma > 0$, step size

Theorem (simplified) (Wang et al., 2022b)

- $\boldsymbol{D}_i \in \{0, 1\}^{n \times n}$, *one activation pattern*
- $\mathcal{K}_i = \{\boldsymbol{v} \in \mathbb{R}^d, \mathbb{1}_{\boldsymbol{X}\boldsymbol{w}_i \geq 0} = \boldsymbol{D}_i\}$, *convex cones*

$$\min_{\boldsymbol{u}_i, \boldsymbol{v}_i \in \mathcal{K}_i} \left\| \sum_{i=1}^{m} \boldsymbol{D}_i \boldsymbol{X}(\boldsymbol{u}_i - \boldsymbol{v}_i) - \boldsymbol{y} \right\|_2^2 + \lambda \sum_{i=1}^{m} \|\boldsymbol{u}_i\|_2 + \|\boldsymbol{v}_i\|_2 \qquad (2)$$

Equivalence: The non convex (1) and the convex (2) problems have the same optimal loss when there are enough neurons ($m \geq n + 1$)

Characterization: We can find the optimal neurons for the non convex problem using the solution of the convex problem using a simple mapping:

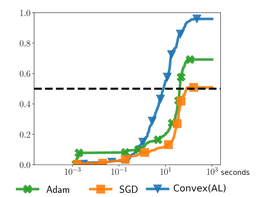
$$(\boldsymbol{w}_i, \alpha_i) = \left(\frac{\boldsymbol{u}_i}{\sqrt{\|\boldsymbol{u}_i\|_2}}, \sqrt{\|\boldsymbol{u}_i\|_2}\right) \quad \text{if } \boldsymbol{u}_i \text{ is non-zero}$$

$$(\boldsymbol{w}_i, \alpha_i) = \left(\frac{\boldsymbol{v}_i}{\sqrt{\|\boldsymbol{v}_i\|_2}}, -\sqrt{\|\boldsymbol{v}_i\|_2}\right) \quad \text{if } \boldsymbol{v}_i \text{ is non-zero}$$

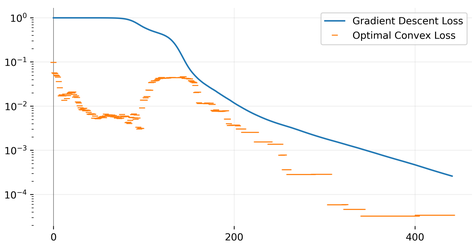
Results for other settings

- Additional layers, by using all possible combinations of activation matrices (Ergen and Pilanci, 2021)
- Vector output, the regularization changes to require a nuclear norm
- Batch Normalization by replacing $\boldsymbol{D}_i\boldsymbol{X}$ with the first matrix in its Singular Value Decomposition
- Wasserstein Generative Adversarial Network, as a convex-concave game
- Parallel networks (Wang et al., 2022a)

Applications

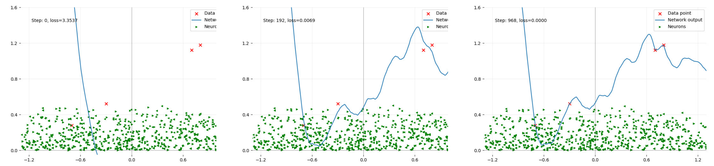


Solving the convex problem can be faster

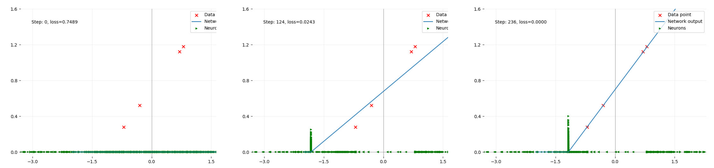


Gradient descent vs optimal convex loss

Large scale initialization dynamic



Small scale initialization dynamic



References

T. Ergen and M. Pilanci. Global optimality beyond two layers: Training deep relu networks via convex programs. In *International Conference on Machine Learning*, pages 2993–3003. PMLR, 2021.

Y. Wang, T. Ergen, and M. Pilanci. Parallel deep neural networks have zero duality gap. In *The Eleventh International Conference on Learning Representations*, Sept. 2022a.

Y. Wang, J. Lacotte, and M. Pilanci. The hidden convex optimization landscape of regularized two-layer relu networks: An exact characterization of optimal solutions. In *International Conference on Learning Representations, ICLR*, 2022b.