The Hidden Convex Optimization Landscape of Regularized Two-layer ReLU Networks: an exact characterization of optimal solutions

Yifei Wang[1,*], Jonathan Lacotte[1,*] & Mert Pilanci[1]

[1]Stanford University, Electrical Engineering

*Equal contributions.

Introduction

Two layer neural network optimization problem with ReLU activations and m hidden neurons::

$$\mathcal{P}_m^* = \min_{\theta \in \Theta_m} \left\{ \mathcal{L}_\beta(\theta) := \ell\Big(\sum_{i=1}^m \sigma(Xu_i)\alpha_i\Big) + \frac{\beta}{2}\sum_{i=1}^m \left(\|u_i\|_2^2 + \alpha_i^2\right) \right\}.$$

with data matrix $X \in \mathbb{R}^{n \times d}$, $\sigma(z) = (z)_+ = \max\{z, 0\}$, and ℓ any convex loss function.

Equivalent convex formulation: Convex program with group-ℓ_2 regularization:

$$\mathcal{P}_c^* := \min_{W \in \mathcal{W}} \left\{ \mathcal{L}_\beta^c(W) := \ell\Big(\sum_{i=1}^{2p} D_i X w_i\Big) + \beta \cdot \sum_{i=1}^{2p} \|w_i\|_2 \right\},$$

where

- Diagonal matrices $D_1, \dots, D_p \in \mathbb{R}^{n \times n}$ = all possible values of $\mathrm{diag}(\mathbf{1}(Xu \geqslant 0))$ for $u \in \mathbb{R}^d$
- Convex cones partition $C_i = \{u \in \mathbb{R}^d | (2D_i - I)Xu \geqslant 0\}$
- Convex feasible set $\mathcal{W} := \{W = (w_1, \dots, w_{2p}) \mid w_i \in C_i\}$
- Convex set of all optimal solutions $\mathcal{W}^*$

Important result from recent literature:

$$\mathcal{P}_m^* = \mathcal{P}_c^* \text{ for } m \geqslant m^* \text{ where } m^* \leqslant n+1$$

(Q1) How to compute optimal set $\Theta_m^* = \{\theta \in \Theta_m \mid \mathcal{L}_\beta(\theta) = \mathcal{P}_m^*\}$?

(Q2) Can we map one-to-one $\mathcal{W}^*$ and Θ_m^*?

Regularization, sparsity and minimal neural nets

If a neural net $\theta \in \Theta_m$ is scaled (i.e., $\|u_i\| = |\alpha_i| \, \forall \, i$) and each convex cone C_j contains at most a single neuron (u_i, α_i) of θ, then the neural net θ has a **minimal** representation. If a neural net $\theta \in \Theta_m$ is scaled (i.e., $\|u_i\| = |\alpha_i| \, \forall \, i$) and if for each convex cone C_j, all neurons of θ within C_j are positively colinear, then the neural net is **nearly minimal**. Intuitively,

- ReLU activation partitions space into cones C_i
- loss function is locally linear over each cone
- then, regularization promotes sparsity, i.e., a single neuron per cone (minimal neural net) is good enough

Contributions

Lemma. Let $W = (w_1, \dots, w_{2p}) \in \mathcal{W}^*$, denote by $\mathcal{I} = \{i_1, \dots, i_{\|W\|_0}\} \subset [2p]$ the indices such that $w_{i_j}^* \neq 0$, and define for $i_j \in \mathcal{I}$:

$$(u_j, \alpha_j) = \Big(\frac{w_{i_j}}{\sqrt{\|w_{i_j}\|_2}}, \gamma_{i_j}\sqrt{\|w_{i_j}\|_2}\Big)$$

where $\gamma_i = 1$ if $i \leqslant p$ and $\gamma_i = -1$ if $i > p$. Then, the neural net $\theta = \{(u_i, \alpha_i)\}_{i=1}^{\|W\|_0}$ is **optimal and minimal**.

We denote the above mapping by ψ and define

$$\Theta_m^{\mathrm{cvx}} = \psi(\mathcal{W}_m^*)$$

where $\mathcal{W}_m^*$ = convex optimal solutions with cardinality less than m.

Given a neuron (u, α), a collection $\{(u_j, \alpha_j)\}_{j=1}^k$ is a **splitting** of (u, α) if $(u_j, \alpha_j) = (\sqrt{\gamma_j}u, \sqrt{\gamma_j}\alpha)$ for some $\gamma_j \geqslant 0$ and $\sum_{j=1}^k \gamma_j = 1$.

Let $\tilde{\Theta}_m^{\mathrm{cvx}}$ be the set of split neural nets generated from Θ_m^{cvx}.

Theorem.

- Let $m^* = \min_{W \in \mathcal{W}^*} \|W\|_0$. Then, we have $m^* \leqslant n+1$.
- For $m \geqslant m^*$, we have

$$\Theta_m^* = \tilde{\Theta}_m^{\mathrm{cvx}}$$

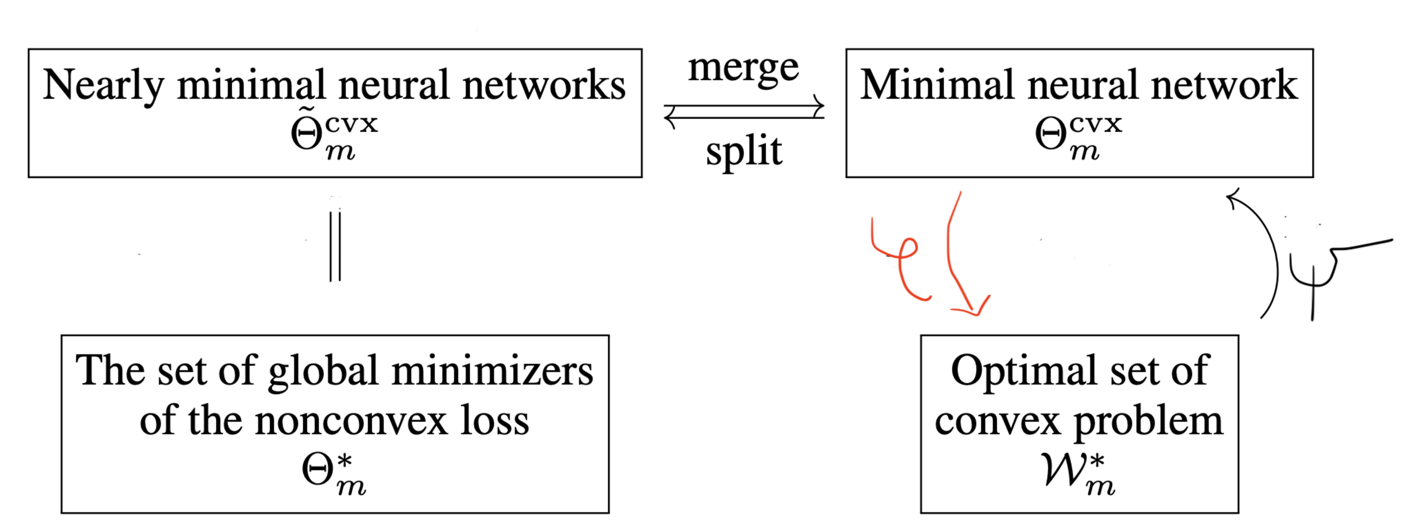

Figure 1:Diagram of relationships between Θ_m^*, $\tilde{\Theta}_m^{cvx}$, Θ_m^{cvx} and $\mathcal{W}_m^*$.

Theorem. Given any optimal neural net $\theta \in \Theta_m^*$, we can explicitly transform it into a minimal neural net $\theta^{min} \in \Theta_m^*$. Furthermore, there exists φ such that $\varphi(\theta^{min}) \in \mathcal{W}_m^*$ and $\varphi(\psi(\theta^{min})) = \theta^{min}$

Recall that limit points of SGD are almost surely Clarke stationary with respect to $\mathcal{L}_\beta$.

Theorem.

- **Any Clarke stationary neural net** θ of the non-convex loss function is a **nearly minimal neural net**. Consequently, **any local minimum of $\mathcal{L}_\beta$ is nearly minimal**.
- Let $\theta \in \Theta_m$ be any neural net. There exists a continuous path in Θ_m from θ to a **nearly minimal neural net** along which the **loss function is strictly decreasing**.

Conclusion

- Sets of convex and non-convex optimal solutions can be mapped one-to-one
- How to solve efficiently (approximately?) convex optimization program to construct good neural nets? Convex cones subsampling?
- How to relate solutions found specifically by SGD to convex optimal solutions?

Acknowledgements

This work was partially supported by the National Science Foundation under grants ECCS-2037304, DMS-2134248, and the Army Research Office.