The Convex Geometry of Backpropagation: Neural Network Gradient Flows Converge to Extreme Points of the Dual Convex Program

Yifei Wang[1] & Mert Pilanci[1]

[1]Stanford University, Electrical Engineering

Introduction

Problem setting

- Two-layer neural networks with ReLU activation, i.e.,
$$f(\boldsymbol{\theta}, \mathbf{X}) = (\mathbf{X}\mathbf{W}_1)_+\mathbf{w}_2,$$
where $\mathbf{W}_1 \in \mathbb{R}^{d\times m}$, $\mathbf{w}_2 \in \mathbb{R}^m$ and $\boldsymbol{\theta} = (\mathbf{W}_1, \mathbf{w}_2)$.
- Training problem
$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) =: \sum_{i=1}^n l(y_i f(\boldsymbol{\theta}; \mathbf{x}_i)),$$
where $l(q) = \log(1+\exp(-q))$ is the logistic loss.

Gradient descent and gradient flow

- The gradient descent update is
$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta(t)\mathbf{g}(t),$$
where $\mathbf{g}(t) \in \partial^\circ \mathcal{L}(\boldsymbol{\theta}(t))$ and ∂° represents Clarke subdifferential.
- For gradient flow, the trajectory of the parameter is an arc $\boldsymbol{\theta} : [0, +\infty) \to \Theta = \{(\mathbf{W}_1, \mathbf{w}_2) | \mathbf{W}_1 \in \mathbb{R}^{d\times m}, \mathbf{W}_2 \in \mathbb{R}^m\}$, which satisfies that for $t \geq 0$, a.e..
$$\frac{d}{dt}\boldsymbol{\theta}(t) \in -\partial^\circ \mathcal{L}(\boldsymbol{\theta}(t)).$$

Implicit regularization of two-layer ReLU networks

- Assume that there exists time t_0 such that $\mathcal{L}(\theta(t_0)) < 1$, i.e., the data is separated at time t_0.
- Lyu and Li[1] show that with $t \to \infty$, any limiting point of $\frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|_2}$ is along the direction to the KKT point of the max-margin problem
$$\min \frac{1}{2}\|\boldsymbol{\theta}\|_2^2, \text{ s.t. } y_i f(\boldsymbol{\theta}; \mathbf{x}_i) \geq 1, i \in [n].$$
where $\|\boldsymbol{\theta}\|_2^2 = \|\mathbf{W}_1\|_F^2 + \|\mathbf{w}_2\|_2^2$.
- This is a **nonconvex** optimization problem.
- Does gradient flow converge to a global minimizer?

Contributions

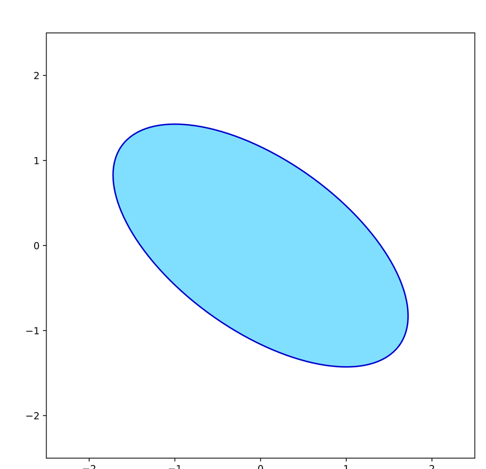
Theorem. Suppose that $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n\times d} \times \{-1, 1\}^n$ is orthogonally separable, i.e., for all $i, i' \in [n]$,
$$\mathbf{x}_i^T\mathbf{x}_{i'} > 0, \text{ if } y_i = y_{i'}, \quad \mathbf{x}_i^T\mathbf{x}_{i'} \leq 0, \text{ if } y_i \neq y_{i'}.$$
Consider the non-convex subgradient flow applied to the non-convex problem. Suppose that the initialization is sufficiently close to the origin. Then, the non-convex subgradient flow converges to the global optimum of the non-convex problem up to scaling, and equivalently to an optimal solution of the convex program P^*_{cvx} .

Convex max-margin problem: The non-convex max-margin problem is equivalent to the following convex program
$$P^*_{\text{cvx}} = \min \sum_{j=1}^p (\|\mathbf{u}_j\|_2 + \|\mathbf{u}'_j\|_2),$$
$$\text{s.t. } \mathbf{Y}\sum_{j=1}^p \mathbf{D}_j\mathbf{X}(\mathbf{u}'_j - \mathbf{u}_j) \geq \mathbf{1},$$
$$(2\mathbf{D}_j - I)\mathbf{X}\mathbf{u}_j \geq 0, (2\mathbf{D}_j - I)\mathbf{X}\mathbf{u}'_j \geq 0, \forall j \in [p].$$
Here $\mathbf{Y} = \text{diag}(\mathbf{y})$.

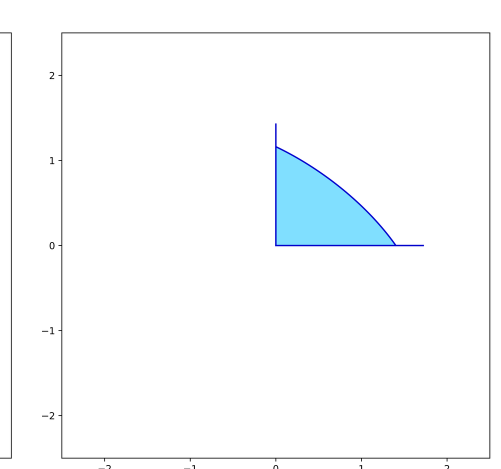
Theorem. The KKT point $(\mathbf{W}_1, \mathbf{w}_2, \boldsymbol{\lambda})$ of the non-convex max-margin problem corresponds to a KKT point of the convex max-margin problem if and only if $\boldsymbol{\lambda}$ satisfies
$$\max_{\mathbf{u}:\|\mathbf{u}\|_2\leq 1} |\boldsymbol{\lambda}^T(\mathbf{X}\mathbf{u})_+| \leq 1.$$

Dual problem

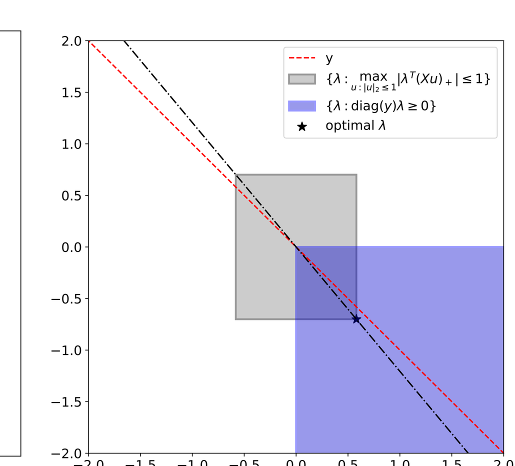
- The dual problem is given by
$$D^* = \max_{\boldsymbol{\lambda}} \mathbf{y}^T\boldsymbol{\lambda} \text{ s.t. } \mathbf{Y}\boldsymbol{\lambda} \succeq 0, \max_{\mathbf{u}:\|\mathbf{u}\|_2\leq 1} |\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{u})_+| \leq 1.$$
- Suppose that $\boldsymbol{\lambda}^*$ is the optimal dual variable. Then, any optimal primal variable $\mathbf{u}$ belongs to the set
$$\arg\max_{\mathbf{u}:\|\mathbf{u}\|_2\leq 1} |(\boldsymbol{\lambda}^*)^T(\mathbf{X}^T\mathbf{u})_+|.$$



Ellipsoid $= \{\mathbf{X}\mathbf{u} : \|\mathbf{u}\|_2 \leq 1\}$.



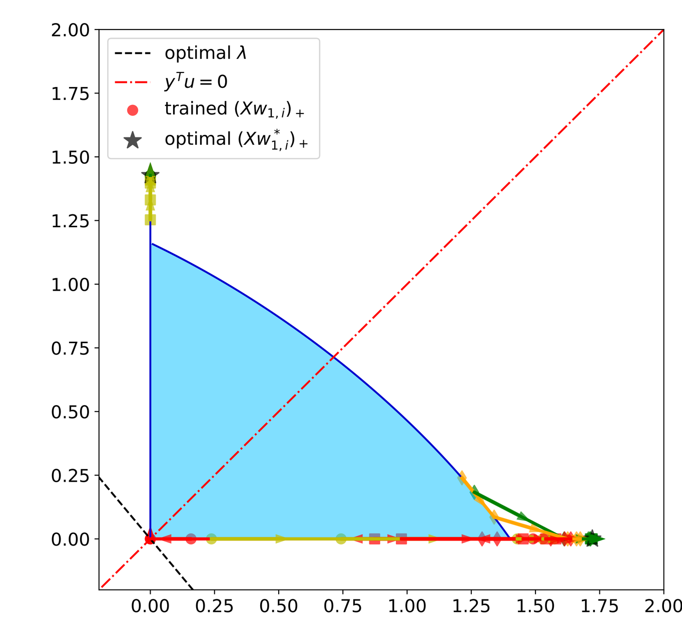
Rectified Ellipsoid $\mathcal{Q} := \{(\mathbf{X}\mathbf{u})_+ : \|\mathbf{u}\|_2 \leq 1\}$ and its extreme points (spikes).



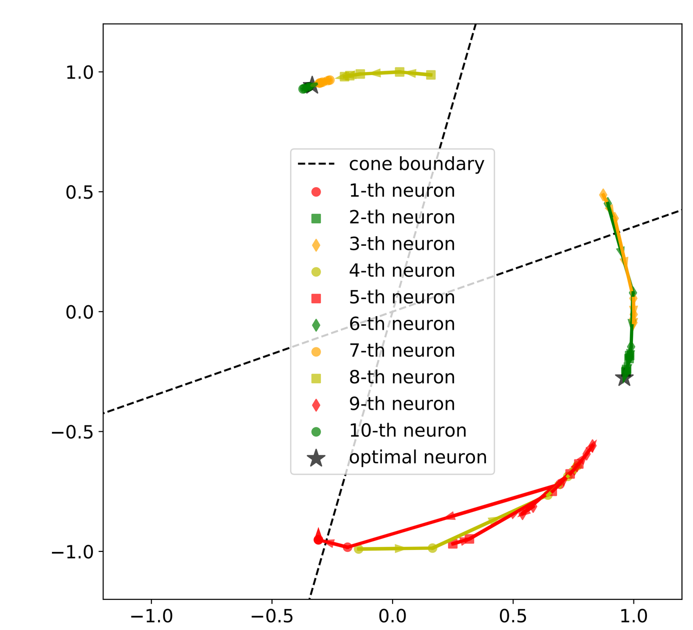
Polar set $\mathcal{Q}^*$ of the Rectified Ellipsoid: $\mathcal{Q}^* = \{\boldsymbol{\lambda} : \max_{\mathbf{z}\in\mathcal{Q}} |\boldsymbol{\lambda}^T\mathbf{z}| \leq 1\}$.

Proposition. Suppose that $(\mathbf{X}, \mathbf{y})$ is orthogonally separable. Suppose that the KKT point $(\mathbf{W}_1, \mathbf{w}_2, \boldsymbol{\lambda})$ of the non-convex problem includes two neurons $(\mathbf{w}_{1,i_+}, w_{2,i_+})$ and $(\mathbf{w}_{1,i_-}, w_{2,i_-})$ such that
$$\mathbb{I}(\mathbf{X}\mathbf{w}_{1,i_+} > 0) \geq \mathbb{I}(y = 1), \quad \mathbb{I}(\mathbf{X}\mathbf{w}_{1,i_-} > 0) \geq \mathbb{I}(y = -1).$$
Then, the dual variable $\boldsymbol{\lambda}$ satisfies
$$\max_{\mathbf{u}:\|\mathbf{u}\|_2\leq 1} |\boldsymbol{\lambda}^T(\mathbf{X}\mathbf{u})_+| \leq 1.$$
In other words, $(\mathbf{W}_1, \mathbf{w}_2)$ globally minimizes the non-convex max-margin problem.

Theorem. Consider the training problem for any dataset. Suppose that the neural network is scaled at initialization such that $\|\mathbf{w}_{1,i}\|_2 = |w_{2,i}|$ for $i \in [m]$. For random initialization, with high probability, there exists neurons $(\mathbf{w}_{1,i}, \mathbf{w}_{2,i})$ such that
$$\text{sign}(\mathbf{y}^T(\mathbf{X}\mathbf{w}_{1,i})_+) = \text{sign}(w_{2,i}) = s,$$
where $s \in \{1, -1\}$. Consider the subgradient flow applied to the non-convex problem. Let $\delta \in (0, 1)$. Suppose that the initialization is sufficiently close to the origin. Then, there exist $T = T(\delta)$ such that
$$\cos\angle\left(\mathbf{w}_{1,i}(T), s\mathbf{X}^T\mathbf{D}(\mathbf{w}_{1,i}(T))\mathbf{y}\right) \geq 1 - \delta.$$
Here $\mathbf{D}(\mathbf{u}) = \text{diag}(\mathbb{I}(\mathbf{X}\mathbf{u} > 0))$.



Trajectories of $(\mathbf{X}\hat{\mathbf{w}}_{1,i})_+$ along the training dynamics of gradient descent.



Trajectories of $\hat{\mathbf{w}}_{1,i} = \frac{\mathbf{w}_{1,i}}{\|\mathbf{w}_{1,i}\|_2}$ along the training dynamics of gradient descent.

Conclusion

- We provide a convex formulation of the non-convex max-margin problem for two-layer ReLU neural networks and uncover a primal-dual extreme point relation between non-convex subgradient flow
- Non-convex subgradient flow globally maximizes the margin of two-layer ReLU networks on orthogonally separable datasets.

Acknowledgements

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

ICLR