Exploring the Loss Landscape of Regularized Neural Networks via Convex Duality

Sungyoon Kim[1], Aaron Mishkin[2], Mert Pilanci[1]

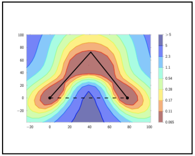
[1]Electrical Engineering Department
[2]Computer Science Department
Stanford University

ICLR 2025
April 24th, 2025

Motivation: Loss Landscape and Global Minima

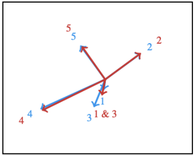
Loss Landscape and Global Minima

Mode Connectivity



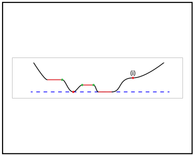
Garipov et al. (2018)

Permutation Symmetry



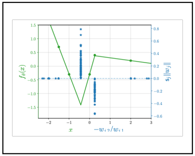
Brea et al. (2019)

Benign Landscape



Vidal et al. (2022)

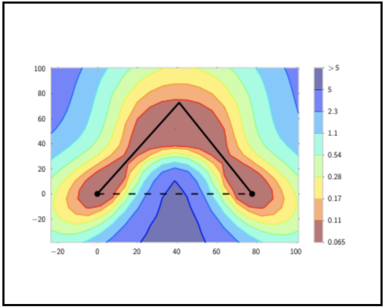
Unique Interpolator



Boursier and Flammarion (2023)

Motivation: Loss Landscape and Global Minima

Mode Connectivity

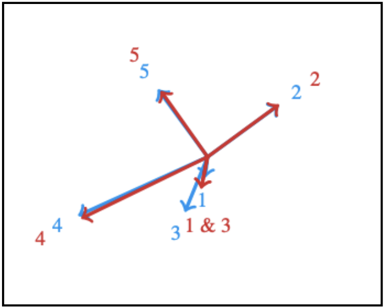


Garipov et al. (2018)

- Two different solutions are connected by a very simple curve.

Motivation: Loss Landscape and Global Minima

Permutation Symmetry

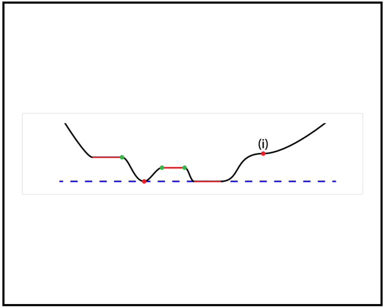


Brea et al. (2019)

- Permutations of an optimal neural network is still optimal.
- They are connected with a smooth path with low training loss.

Motivation: Loss Landscape and Global Minima

Benign Landscape

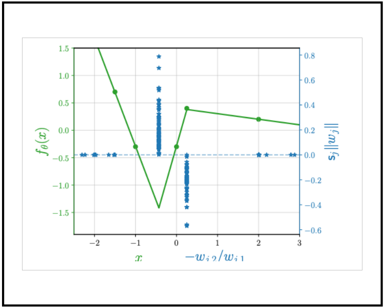


Vidal et al. (2022)

- For sufficiently wide neural networks, there is always a decreasing path to a global optimum.

Motivation: Loss Landscape and Global Minima

Unique Interpolator



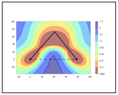
Boursier and Flammarion (2023)

- Penalizing the bias and using free skip connections (e.g. an unregularized linear neuron).

Our Contributions

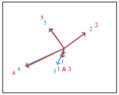
Existing Work

Mode Connectivity



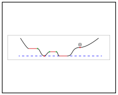
Garipov et al. (2018)

Permutation Symmetry



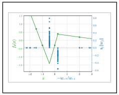
Brea et al. (2019)

Benign Landscape



Vidal et al. (2022)

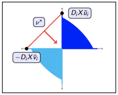
Unique Interpolator



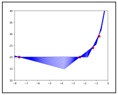
Boursier and Flammarion (2023)

"Our Contributions"

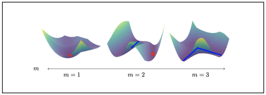
Dual Geometry



Continuum of Interpolators



Staircase of Connectivity



We extend our knowledge on the loss landscape and global minima of neural networks via **the convex optimization perspective**.

For two-layer scalar neural networks with weight decay, we use an equivalent convex program and its dual to

- discuss **novel geometric insights** on how the nonconvex problem's global minima ands stationary points are related to a polytope.
- show how the change in width may lead to **phase transition in the connectivity** of the set of global minima.
- construct examples where a **continuum of optimal interpolators exist** with regularized bias.

We extend some of our results to cases where equivalent convex programs exist, e.g. vector-valued networks and parallel three-layer neural networks.

In this talk...

For two-layer scalar neural networks with weight decay, we use an equivalent convex program and its dual to

- discuss **novel geometric insights** how the nonconvex problem's global minima ands stationary points are related to a polytope.
- show how the change in width may lead to **phase transition in the connectivity** of the set of global minima.
- construct examples where a continuum of optimal interpolators exist with regularized bias.

We extend some of our results to cases where equivalent convex programs exist, e.g. vector-valued networks and parallel three-layer neural networks.

Background: Convex Neural Networks

- Let $X \in \mathbb{R}^{n \times d}$ be the data matrix, $y \in \mathbb{R}^n$ be the labels, $u_j \in \mathbb{R}^d$, $\alpha_j \in \mathbb{R}$ for $j = 1, 2, \cdots m$, and $\beta > 0$.
- Also, $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a convex loss function.

Background: Convex Neural Networks

The training objective that we are interested can be written as the following optimization problem,

$$\min_{u_j \in \mathbb{R}^d, \alpha_j \in \mathbb{R}} \; L\Big(\sum_{j=1}^{m} (Xu_j)_+ \alpha_j, y\Big) + \frac{\beta}{2} \sum_{j=1}^{m} (\|u_j\|_2^2 + \|\alpha_j\|_2^2).$$

Notation

We will use the term "nonconvex objective" to refer to the above problem. Also, the optimal objective will be noted as p^*_{noncvx} and the set of optimal parameters will be noted as Θ^*_m.

The Convex Reformulation

Pilanci and Ergen (2020) shows that if $m \geq m^*$ for some critical width m^*, there exists an **equivalent convex optimization problem**.

- When we denote p^*_{cvx} as the optimal objective of the convex problem,

$$p^*_{\text{cvx}} = p^*_{\text{noncvx}}$$

- There exists a mapping between the optimal solution of the nonconvex objective and the equivalent convex problem (Pilanci and Ergen (2020), Wang et al. (2021)).

The Convex Reformulation

The equivalent convex optimization problem is written as

$$\min_{u_i,v_i} \frac{1}{2}\|\sum_{i=1}^{P} D_i X(u_i - v_i) - y\|_2^2 + \beta \sum_{i=1}^{P}(\|u_i\|_2 + \|v_i\|_2),$$

subject to constraints $(2D_i - I)Xu_i \geq 0,\ (2D_i - I)Xv_i \geq 0.$

- Here, $D_i = Diag(1(Xh \geq 0))$ denote all possible "hyperplane arrangement patterns"
- Intuition: in the constraint set, ReLU becomes linear.

The Convex Reformulation

The equivalent convex optimization problem is written as

$$\min_{u_i, v_i} \frac{1}{2}\|\sum_{i=1}^{P} D_i X(u_i - v_i) - y\|_2^2 + \beta \sum_{i=1}^{P}(\|u_i\|_2 + \|v_i\|_2),$$

subject to constraints $(2D_i - I)Xu_i \geq 0, \ (2D_i - I)Xv_i \geq 0.$

Notation

We will use the term "convex reformulation" to refer to the above problem. Also, the optimal objective will be noted as p^*_{cvx} and the set of optimal parameters will be noted as $\mathcal{P}^*$.

The Dual Problem

The dual of the convex reformulation can be written as

$$\max_{\nu \in \mathbb{R}^n} -L^*(\nu) \quad \text{subject to} \quad |\nu^T (Xu)_+| \leq \beta \quad \forall \|u\|_2 \leq 1.$$

Here, L^* is the Fenchel conjugate of $L(\cdot, y)$.

- Denote the optimal objective of the dual problem as d^*_{cvx}.
- If $m \geq m^*$,

$$p^*_{\text{cvx}} = d^*_{\text{cvx}} = p^*_{\text{noncvx}}.$$

Optimal Set Characterization

Theorem (Mishkin and Pilanci (2023))

$\mathcal{P}^*$ is a polyhedral set. Moreover, with the solution mapping, Θ_m^* is a curved image of the polyhedral set $\mathcal{P}^*$ when m is sufficiently large.

- The optimal set Θ_m^* is related with a simple geometric object (polytope)!
- Our work largely builds upon this characterization, and many concepts needed for proof were adapted.

Optimal Polytope and the Dual Optimum

Theorem (The Optimal Polytope, informal)

Suppose L is a strictly convex loss function. The directions of optimal parameters of the convex problem, noted as $\bar{u}_i, \bar{v}_i$, are uniquely determined from the dual optimum ν^*. Moreover, the solution set is the polytope,

$$\mathcal{P}^* = \left\{ (c_i \bar{u}_i, d_i \bar{v}_i)_{i=1}^P \mid c_i, d_i \geq 0 \quad \forall i \in [P], \quad \sum_{i=1}^P D_i X \bar{u}_i c_i - D_i X \bar{v}_i d_i = y^* \right\}$$

for the unique optimal model fit y^*.

- The optimal directions $\bar{u}_i, \bar{v}_i$ are solutions of

$$\max_{\|u\|_2 \leq 1} |(\nu^*)^T (Xu)_+|.$$

Geometric Intuition: The Rectified Ellipsoid

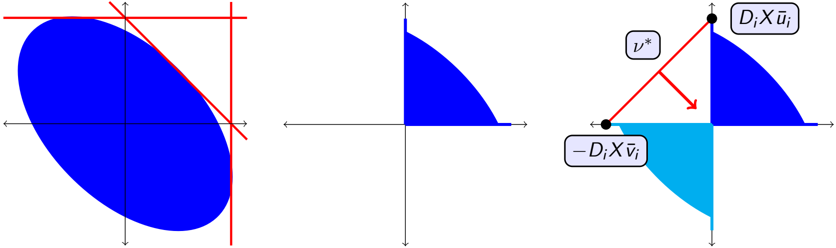
Definition (Pilanci and Ergen (2020))

The rectified ellipsoid is defined as the set

$$\mathcal{Q}_X = \left\{ (Xu)_+ \mid \|u\|_2 \leq 1 \right\}.$$

- It is the image of the ReLU mapping of the ellipsoid $\mathcal{E}_X = \{Xu \mid \|u\|_2 \leq 1\}$.

Geometric Intuition: The Rectified Ellipsoid



(a) $\mathcal{E}_X$

(b) $\mathcal{Q}_X$

(c) $\mathcal{Q}_X \cup -\mathcal{Q}_X$

- The dual variable ν^* **decides the "face"** where the points $\{D_i X \bar{u}_i\}_{i=1}^{P} \cup \{-D_i X \bar{v}_i\}_{i=1}^{P}$ lies on.
- Blue set corresponds to $\bar{u}_i$, cyan set corresponds to $\bar{v}_i$.

The Staircase of Connectivity: Motivation

- Denote $\mathrm{card}((u_i, v_i)_{i=1}^P)$ as the number of nonzero vectors in $\{u_i\}_{i=1}^P \cup \{v_i\}_{i=1}^P$.
- A solution map exists between Θ_m^* and the cardinality - constrained set

$$\mathcal{P}_m^* = \left\{(u_i, v_i)_{i=1}^P \in \mathcal{P}^* \mid \mathrm{card}((u_i, v_i)_{i=1}^P) \leq m\right\}$$

Result from Wang et al. (2021).

The Staircase of Connectivity: Motivation

- Though $\mathcal{P}^*$ is a connected set, $\mathcal{P}^*_m$ **might not be connected** due to cardinality constraints - phase transitional behavior in connectivity!
- As Θ^*_m and $\mathcal{P}^*_m$ is related by the solution map, connectivity properties of Θ^*_m can be deduced from that of $\mathcal{P}^*_m$.

The Staircase of Connectivity

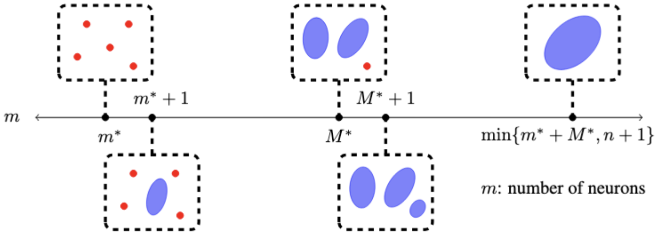


Figure: A schematic for the staircase of connectivity

- Red dots correspond to isolated points, where blue sets are connected components with more than one point.
- There exists critical widths m^*, M^* associated with the convex program that governs the phase tranistion.

The Staircase of Connectivity

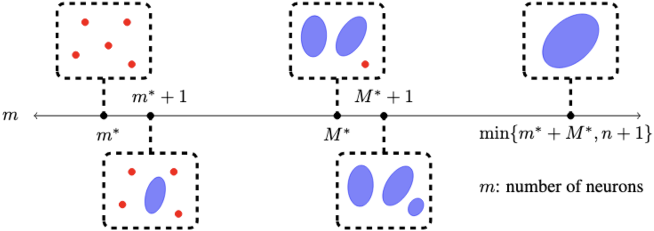


Figure: A schematic for staircase of connectivity

- When $m = m^*$, Θ^*_m is a set of finite isolated points.

The Staircase of Connectivity

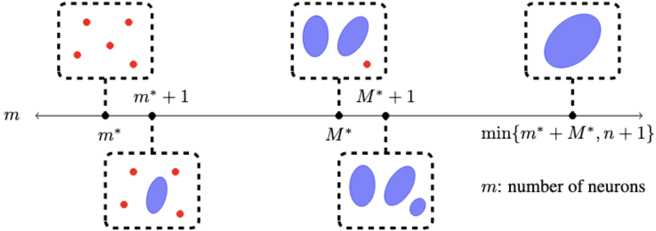


Figure: A schematic for staircase of connectivity

- When $m \geq m^* + 1$, there exists a path between two different optimal solutions in Θ^*_m

The Staircase of Connectivity

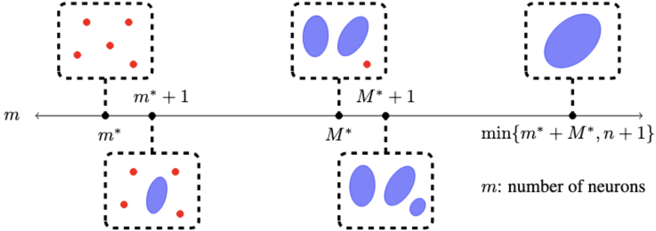


Figure: A schematic for staircase of connectivity

- When $m = M^*$, there exists an isolated point in Θ^*_m.

The Staircase of Connectivity

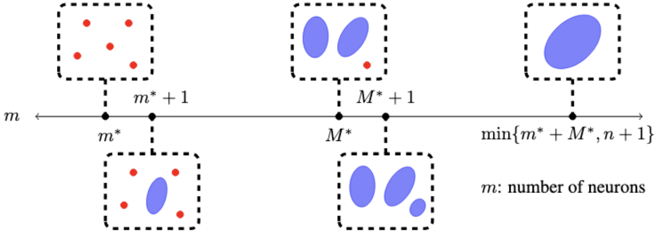


Figure: A schematic for staircase of connectivity

- When $m \geq M^* + 1$, there is no isolated point in Θ^*_m.
- Moreover, for an optimal solution $(w_i, \alpha_i)_{i=1}^m$, any permutation $(w_{\sigma(i)}, \alpha_{\sigma(i)})_{i=1}^m$ has a path inside Θ^*_m that connects the two solutions.

The Staircase of Connectivity

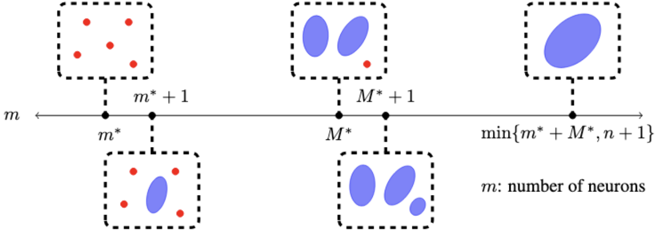


Figure: A schematic for staircase of connectivity

- When $m \geq \min\{M^* + m^*, n+1\}$, Θ^*_m is connected.

Relations with Existing Landscape Results

- Theoretical justification of Yang et al. (2021) that empirically observes larger models tend to have more connected optimal sets.
- Haeffele and Vidal (2017) shows that when $m \geq n + 1$, there is no spurious local minima. We further characterize that all sublevel sets are connected.
- Nguyen (2021) shows that when there is no regularization, Θ_m^* is connected when $m \geq n + 1$. We extend the result to the regularized case.
- Our analysis is also tightly connected to Simsek et al. (2021), who adds a neuron to connect permutations of optimal solutions.

Conclusion

- We derived novel characterizations of the loss landscape and global minima of neural networks by leveraging tools from convex optimization.
- An extension of these results to different nonconvex problems that has convex reformulations could be an interesting future direction.

Poster session: Hall 3 + Hall 2B #350, today 3pm – 5:30pm

Bibliography

Boursier, E. and Flammarion, N. (2023). Penalising the biases in norm regularisation enforces sparsity. *Advances in Neural Information Processing Systems*, 36:57795–57824.

Brea, J., Simsek, B., Illing, B., and Gerstner, W. (2019). Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape. *arXiv preprint arXiv:1907.02911*.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31.

Haeffele, B. D. and Vidal, R. (2017). Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339.

Mishkin, A. and Pilanci, M. (2023). Optimal sets and solution paths of relu networks. In *International Conference on Machine Learning*, pages 24888–24924. PMLR.

Nguyen, Q. (2021). A note on connectivity of sublevel sets in deep learning. *arXiv preprint arXiv:2101.08576*.

Pilanci, M. and Ergen, T. (2020). Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pages 7695–7705. PMLR.

Simsek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W., and Brea, J. (2021). Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning*, pages 9722–9732. PMLR.

Vidal, R., Zhu, Z., and Haeffele, B. D. (2022). Optimization landscape of neural networks. *Mathematical Aspects of Deep Learning*, page 200.

Wang, Y., Lacotte, J., and Pilanci, M. (2021). The hidden convex optimization landscape of regularized two-layer relu networks: an exact characterization of optimal solutions. In *International Conference on Learning Representations*.

Yang, Y., Hodgkinson, L., Theisen, R., Zou, J., Gonzalez, J. E., Ramchandran, K., and Mahoney, M. W. (2021). Taxonomizing local versus global structure in neural network loss landscapes. *Advances in Neural Information Processing Systems*, 34:18722–18733.