Effects of width-dependent model hyperparameters and ℓ_2-regularization on the loss landscape of two-layer ReLU networks

Haruka Eshima HARUKA.ESHIMA@OIST.JP
Okinawa Institute of Science and Technology, Okinawa, Japan

Makoto Yamada MAKOTO.YAMADA@OIST.JP
Okinawa Institute of Science and Technology, Okinawa, Japan

Abstract

Understanding deep neural networks remains a central challenge in machine learning. In particular, the theoretical properties of even two-layer ReLU networks, especially in the presence of weight decay, remain poorly understood. To this end, we derive a sufficient condition on the hyperparameter settings under which the global minima collapse to the zero solution. Interestingly, our experiments reveal that using AdamW as an optimizer prevents the collapse of the learned parameters, whereas using SGD does not, which may help explain the success of AdamW in deep learning training. In addition, when restricting the input dimension to one, we derive an analytical solution for the globally optimal parameter sets of two-layer ReLU networks and show that ℓ_2-regularization has a width-invariant effect on connectivity, but its dimensionality-reducing effect becomes stronger as the network width increases. These results provide insight into how width-dependent hyperparameters influence the geometry of regularized loss landscapes.

Keywords: Loss Landscape, Two-layer ReLU Networks, L2-reguralization

1. Introduction

Loss landscape analysis characterizes quantitative variations of the loss function in finite-width machine learning models, and studying the loss landscape of two-layer linear unit activation networks (ReLU) [29] has been a key topic of interest in the machine learning community [4, 17, 18, 34, 35]. Recent work has extended the theoretical analysis to ℓ_2-regularized two-layer ReLU networks, including characterizations of global optima and their geometric structure [4, 18]. However, the effect of explicit width-dependent hyperparameter scalings [15, 39] on these regularized loss landscapes does not appear to have been studied in a systematic way.

In this paper, we analyze the global minima of two-layer ReLU neural networks for ℓ_2-regularized loss in a width-dependent hyperparameter setting. We show that, when the ℓ_2-regularization coefficient scales faster than the scaling of the network, zero is the unique global minimum for a sufficiently overparametrized model, and experimentally show that whether the learned parameters collapse to zero depends on the choice of optimizer. We further analyze the effects of ℓ_2-regularization on the dimension and connectivity of global minima for one-dimensional input. Our results imply that the effect of ℓ_2-regularization on the connectivity of global minima is negligible regardless of how the model width is scaled up, whereas its dimensional reduction effect becomes stronger.

© H. Eshima & M. Yamada.

Problem Setting. The network model $f_\theta : \mathbb{R}^d \to \mathbb{R}$ is

$$f_\theta(x) = \frac{1}{\alpha} \sum_{j=1}^{m} (x u_j)_+ \omega_j, \tag{1}$$

where the trainable parameter of the model is $\theta = (u_i, \omega_i)_{i=1}^m \in (\mathbb{R}^d \times \mathbb{R})^m =: \Theta$. $U := (u_1, \ldots, u_m) \in \mathbb{R}^{d \times m}$ and $W := (\omega_1, \ldots, \omega_m)^T \in \mathbb{R}^m$ are the first and second layer weights, m is the number of hidden neurons, $(\cdot)_+ = \max\{\cdot, 0\}$ is the ReLU activation, and the hyperparameter $1/\alpha > 0$ is the scaling factor. Each component in the parameter is initialized independently by $u_{i,j}^0 \sim N(0, \tau_1^2)$ and $\omega_i^0 \sim N(0, \tau_2^2)$ where $\tau_1, \tau_2 \in \mathbb{R}$ are hyperparameters. By abuse of notation, for input data $X \in \mathbb{R}^{n \times d}$, we write $f_\theta(X)$ to denote the output of the model $(f_\theta(x_1), \ldots, f_\theta(x_n))^T \in \mathbb{R}^n$. We aim to analyze the effects of adding ℓ_2-regularization $\frac{\beta}{2} \sum_{j=1}^m (u_j^2 + \omega_j^2)$ to the loss function $\ell(\theta)$. The value of the hyperparameter $\beta > 0$ matches the weight decay coefficient for the gradient descent method. We introduce width-dependent hyperparameter setting $\alpha = m^a$, $\tau_1 = m^{-b_1}$, $\tau_2 = m^{-b_2}$, $\beta = m^{-\delta}$.

2. Effects of width-dependent ℓ_2-reguralization on loss landscape

2.1. Collapse of global optima due to ℓ_2-regularization

Theorem 1 shows that $\delta < a$ is a sufficient condition for a hyperparameter setting with which only the zero weight is the globally optimal parameter for the model (1) with sufficiently large m. This collapse of global optima to zero happens for most of the convex loss functions $\ell(\theta)$ used in practice. We will discuss this result with numerical experiments in Section 3.

Theorem 1 *We consider minimizing the following loss function*

$$\min_{\theta \in \Theta} L_\ell(\theta) = \ell(\theta) + \frac{\beta}{2} \sum_{j=1}^{m} (\|u_j\|_2^2 + \omega_j^2) \tag{2}$$

where ℓ is finite on Θ, and can be written as a convex function of $f_\theta(X)$. For example, $\ell(\theta) = \|f_\theta(X) - Y\|^2$.
If $\delta < a$, then for sufficiently large m, $\operatorname{argmin}_{\theta \in \Theta} L_\ell(\theta) = \{(0,0)_{i=1}^m\}$.

Theorem 1 and experimental results (Section 3 and Appendix F) tell us undesirable values of the weight decay coefficient β for the gradient descent algorithm.

2.2. Change in dimension and connectivity of global optima due to ℓ_2-regularization

When $\delta > a$, the regularization is not necessary strong enough to force the global minima to collapse, and our focus becomes how ℓ_2-regularization changes the geometry of the nontrivial global minima. This question is motivated by recent work showing that ℓ_2-regularized two-layer ReLU networks admit exact convex formulations, which is used to study optimal solution sets [28, 30]. We specifically focus on the one-dimensional ReLU setting (3). This setting exposes geometric structure that is difficult to access in the higher-dimensional setting considered by [28, 30], which then

allows us to derive closed-form descriptions of the ℓ_2-regularized or unregularized global minima (Theorem 2, 3) and to compare their dimension, boundedness, and connectivity directly.

$$\min_{\theta\in\mathbb{R}^{2m}} L(\theta)=\frac{1}{2}\left\|\frac{1}{\alpha}\sum_{j=1}^{m}(Xu_j)_+\omega_j-Y\right\|_2^2+\frac{\beta}{2}\sum_{j=1}^{m}(u_j^2+\omega_j^2). \tag{3}$$

For notations, we introduce $D(S) := \mathrm{Diag}(\mathbf{1}[X \geq 0])$ and $D(S^c) := \mathrm{Diag}(\mathbf{1}[X \leq 0])$ where $\mathbf{1}[X \leq 0]$ is an indicator vector with $(\mathbf{1}[X \leq 0])_i = 1[x_i \leq 0]$. Also, $X_S := \mathrm{Diag}(\mathbf{1}[X \geq 0])X$ and $X_{S^c} := \mathrm{Diag}(\mathbf{1}[X \leq 0])X$. To prevent trivial solutions, we assume that X has at least one positive element and one negative element (so $\|X_S\|_2 \neq 0$ and $\|X_{S^c}\|_2 \neq 0$), that the input training data X and the output training data Y are independent of m, and that $0 < \min\{|X_S^T Y|,\ |X_{S^C}^T Y|\}$.

Theorem 2 *The set of globally optimal parameters $\varphi^*(m) := \{\theta \in \mathbb{R}^{2m} \ : \ \theta = \mathrm{argmin}_{\theta\in\mathbb{R}^{2m}} L(\theta)$ with $\beta = 0\}$ for unregularized squared loss is*

$$\varphi^*(m)=\left\{(u_j,\omega_j)_{j=1}^m \ \middle|\ \sum_{j:\,u_j\geq 0} u_j\omega_j = \frac{\alpha X_S^\top Y}{\|X_S\|_2^2},\ \sum_{j:\,u_j\leq 0} u_j\omega_j = \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}\right\}. \tag{4}$$

Adding ℓ_2-regularization ($\beta > 0$) restricts the solution set because we need $|u_i| = |\omega_i|\ \forall i \in [m]$ for all globally optimal parameters $(u_i, \omega_i)_{i=1}^m$. (See the proof of Theorem 3 in Appendix B.2.) The visualization of global minima in Figure 1 (details are in Appendix E) illustrates that adding ℓ_2-regularization significantly restricts the set of globally optimal parameters.

Theorem 3 *The set of globally optimal parameters $\Theta^*(m) := \{\theta \in \mathbb{R}^{2m} \ : \ \theta = \mathrm{argmin}_{\theta\in\mathbb{R}^{2m}} L(\theta)$ with $\beta > 0\}$ for ℓ_2-regularized squared loss is*

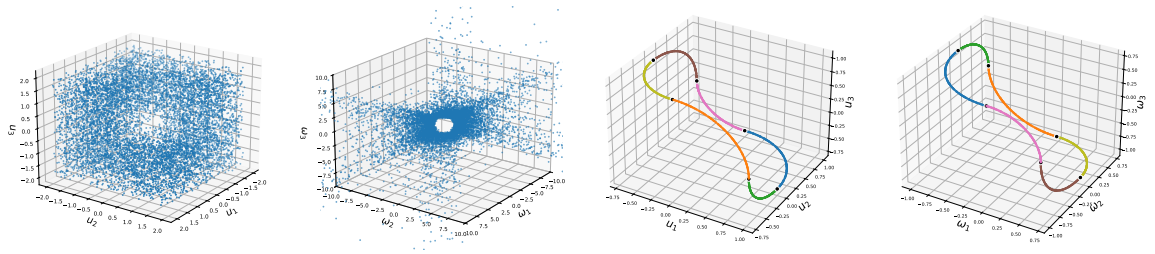
$$\Theta^*(m) = \left\{(u_i,\omega_i)_{i=1}^m \ \middle|\ \sum_{i:\,u_i\geq 0} u_i^2 = |\gamma_P^*|,\ \sum_{i:\,u_i\leq 0} u_i^2 = |\gamma_N^*|,\ \omega_i = \begin{cases}\mathrm{sign}(\gamma_P^*)\,u_i & (u_i > 0),\\ \mathrm{sign}(\gamma_N^*)\,u_i & (u_i < 0),\\ 0 & (u_i = 0)\end{cases}\right\} \tag{5}$$

where $\gamma_P^* = \alpha\frac{\mathcal{S}_{\alpha\beta}(X_S^T Y)}{\|X_S\|_2^2}$ and $\gamma_N^* = \alpha\frac{\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y)}{\|X_{S^c}\|_2^2}$. ($\mathcal{S}_{\alpha\beta}(b) := \mathrm{sign}(b)\max(|b| - \alpha\beta, 0) = \mathrm{sign}(b)\max(|b| - m^{a-\delta}, 0)$ is the soft-thresholding operator.)

The ReLU output is decomposed into the sum of different contributions based on activation patterns, and the original minimization problem (2) can be written as a convex problem [12, 30]. The proofs of Theorem 2, 3 (Appendix B.2) utilize the fact that we can remove the constraints for the convex problem introduced in [12, 30] if the data is one-dimensional (Proposition 25). This simplification is not feasible for general d-dimensional input because the activation patterns are more complicated.

2.2.1. CONNECTIVITY

In this section, we discuss how the connectivity of globally optimal solutions $\varphi^*(m)$ and $\Theta^*(m)$ changes with respect to m. For a set S, we say $x, y \in S$ are **connected** in S if $\exists$ a continuous function $f : [0, 1] \to S$ that satisfies $f(0) = x$ and $f(1) = y$. We say S is **connected** if, for any two points $x, y \in S$, x and y are connected in S. Theorem 4 shows phase transitional behavior of the connectivity of the global minima for the unregularized squared loss.



(a) Projection of $\varphi^*(m)$ in u-coordinates (b) Projection of $\varphi^*(3)$ in ω-coordinates (c) Projection of $\Theta^*(3)$ in u-coordinates (d) Projection of $\Theta^*(3)$ in ω-coordinates

Figure 1: Illustration of global optima of width-3 two-layer ReLU neural netwwork (1) in the parameter space. More details of this figure are in Appendix E.

Theorem 4 *We have the following connectivity results for the solution set $\varphi^*(\theta)$ to the unregularized squared loss.*
(1) For $m = 1$, $\varphi^*(m) = \emptyset$.
(2) For $m = 2$, $\varphi^*(m)$ has exactly 2 connected components.
(3) For $m \geq 3$, $\varphi^*(m)$ is connected.

If we add the ℓ_2-regularization term, we find critical widths $\underline{M}$ and $\overline{M}$ for phase transition behaviors, which depend on the hyperparameter setting.

Theorem 5 *Assume $\delta > a$. Define $M^*(m) = 1[|X_S^T Y| > m^{a-\delta}] + 1[|X_{S^c}^T Y| > m^{a-\delta}]$, $\underline{M} = \min\{m \in \mathbb{N}_{\geq 1} \mid M^*(m) \geq 1\}$ and $\overline{M} = \min\{m \in \mathbb{N}_{\geq 1} \mid M^*(m) = 2\}$. $\underline{M}$ and $\overline{M}$ are well-defined because $M^*(m) \in \{0, 1, 2\}$ is increasing with m and $M^*(m) = 2$ for sufficiently large m. We have the following connectivity results for the solution set $\Theta^*(\theta)$ to the ℓ_2-reguralized loss.*
(1) For $m < \underline{M}$, $\Theta^*(m)$ is a singleton ($\{(0,0)_{i=1}^m\}$).
(2) If $\underline{M} = \overline{M} = m = 1$, $\Theta^*(m) = \emptyset$.
(3) If $\underline{M} \leq m = 1 < \overline{M}$ or $\underline{M} \leq \overline{M} \leq m = 2$, $\Theta^*(m)$ is a finite set.
(4) Otherwise, $\Theta^*(m)$ is connected.

Theorem 5 implies that there are only three connectivity behaviors for $\Theta^*(m) \neq \emptyset$. As Kim et al. [18, Theorem 2] showed, this is not always the case for general d-dimensional input data. We provide explanations for this limited connectivity result for the one-dimensional input data from a perspective on the roles of unnecessary neurons (Appendix C.1) and a convex formulation of neural networks introduced by Pilanci and Ergen [30] (Appendix C.2).

2.2.2. DIMENSION

In this section, we compare the dimensionality and bounds for the set of optimal parameters. For a subset $A \subset \mathbb{R}^{2m}$, we define the **dimension** of A denoted by $\dim(A)$ to be the maximum k such that A contains a k-dimensional embedded C^1 submanifold (equivalently, the maximal stratum dimension). For unregularized squared loss, the globally optimal parameters are dense and spread out without a bound outside the sphere of radius $\|\theta^*\|_2$, where $\theta^* \in \Theta^*(\theta)$, in the parameter space $\mathbb{R}^{2m}$. The dimensional difference between $\varphi^*(m)$ and the parameter space $\mathbb{R}^{2m}$ is two, which is independent of m as long as m is sufficiently large.

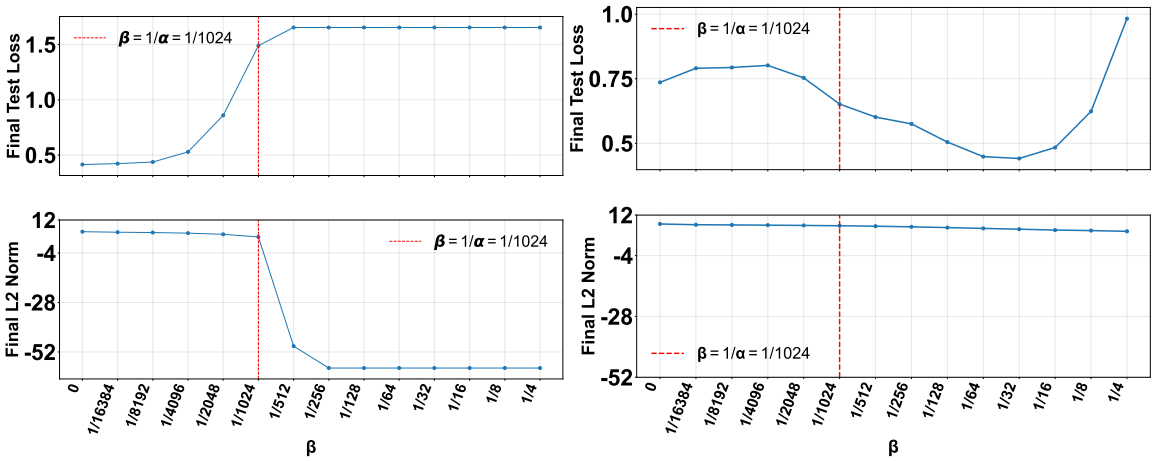


Figure 2: Final Test loss (above) and Final ℓ_2-norm of parameters (bottom, y-axis values are taken in logarithmic base 2) over weight decay coefficient β (plotted in logarithmic scale) after training width-1024 two-layer ReLU networks initialized by τ_1=τ_2=0.01 (b_1=b_2=0) and scaled by $1/\alpha = 1/1024$ with SGD (left) or AdamW (right) for 40000 epochs on Yacht Hydrodynamics [14].

Proposition 6 *For sufficiently large m, $\dim(\varphi^*(m)) = 2m - 2$.*

Proposition 7 *For sufficiently large m, $\varphi^*(m)$ is unbounded. Especially, $\forall n \geq \frac{a}{2}$, $\exists \varphi_n^*(m) \subset \varphi^*(m)$ s.t. $\forall \phi_n^* \in \varphi_n^*(m)$, $\|\phi_n^*\|_2 = \Theta(m^n)$ and $\dim(\varphi_n^*(m)) = 2m - 2$.*

We find that adding the ℓ_2-regularization term reduces the dimension of the set of optimal parameters by m and imposes a bound on the set. There is always $(m+2)$-dimensional difference between $\Theta^*(m)$ and the parameter space $\mathbb{R}^{2m}$. The dimensional difference grows with order m by increasing the width m.

Proposition 8 *Under Assumption $\delta > a$, for sufficiently large m, $\dim(\Theta^*(m)) = m - 2$.*

Proposition 9 *Under Assumption $\delta > a$, for sufficiently large m, $\Theta^*(m)$ is bounded and $\forall\, \theta^* \in \Theta^*(m)$, $\|\theta^*\|_2 = \sqrt{2\alpha \left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2} \right)} = \Theta(m^{\frac{a}{2}})$.*

3. Numerical Experiments

Theorem 1 itself does not inform us whether the learned parameter collapses to zero. For behaviors of learned parameters, several works discuss convergence to global minima using Adam [19] or stochastic gradient descent (SGD) (details are in Appendix A). We trained two-layer ReLU neural networks (1) with different widths (m= 64, 128, 256, 512, 1024, 2048) using the Yacht Hydrodynamics data [14] (squared loss), and MNIST [21] (cross entropy loss). We trained them with SGD or AdamW [24] using β as weight decay coefficient, with a fixed learning rate. (More details are in Appendix F.) Our numerical experiments show that when $\delta > a$ (i.e. $\beta > \frac{1}{\alpha}$), the learned parameters by SGD collapse as expected by Theorem 1 (e.g. Figure 2 left), while using AdamW [24]

prevents the learned parameters from collapsing to zero (e.g. Figure 2 right). This phenomenon is seen across different hyperparameter settings or widths. The results for other hyperparameter settings, different widths, and for MNIST are shown in Appendix F.

4. Conclusion

We analyzed the set of globally optimal parameters for two-layer ReLU networks under width-dependent hyperparameters and ℓ_2-regularization, and derived a sufficient condition for the collapse to the zero-weight solution, showing that width, output scaling, and weight decay can qualitatively affect the regularized loss landscape. For one-dimensional inputs, we explicitly characterized the globally optimal parameter sets, showing that ℓ_2-regularization has a width-invariant effect on connectivity while its dimensionality-reducing effect strengthens with width. Experiments show that SGD follows the predicted collapse, whereas AdamW prevents it; this may be related to decoupled weight decay, which is not equivalent to optimizing the explicit ℓ_2-regularized objective for adaptive methods [24], and to AdamW's implicit bias toward ℓ_∞-constrained optimization [38]. Explaining this optimizer-dependent behavior theoretically remains as future work.

References

[1] El Mehdi Achour, François Malgouyres, and Sébastien Gerchinovitz. The loss landscape of deep linear neural networks: a second-order analysis. *Journal of Machine Learning Research*, pages 1–76, 2024.

[2] Shunta Akiyama and Taiji Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In *International Conference on Machine Learning*, 2021.

[3] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019.

[4] Etienne Boursier, Matthew Bowditch, Matthias Englert, and Ranko Lazic. Benignity of loss landscape with weight decay requires both large overparametrization and initialization. In *High-dimensional Learning Dynamics 2025*.

[5] Davide Buffelli, Jamie McGowan, Wangkun Xu, Alexandru Cioba, Da-shan Shiu, Guillaume Hennequin, and Alberto Bernacchia. Exact, tractable gauss-newton optimization in deep reversible architectures reveal poor generalization. *Advances in Neural Information Processing Systems*, 2024.

[6] Lénaïc Chizat and Francis R. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems, 2018*.

[7] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

[8] Yaim Cooper. Global minima of overparameterized neural networks. *SIAM Journal on Mathematics of Data Science*, 3(2):676–691, 2021.

[9] Marvin F. da Silva, Felix Dangel, and Sageev Oore. Hide & seek: Transformer symmetries obscure sharpness & riemannian geometry finds it. In *International Conference on Machine Learning*, 2025.

[10] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, pages 1–65, 2024.

[11] Luke Eilers, Raoul-Martin Memmesheimer, and Sven Goedeke. A generalized neural tangent kernel for surrogate gradient learning. *Advances in Neural Information Processing Systems*, 2024.

[12] Tolga Ergen and Mert Pilanci. Global optimality beyond two layers: Training deep relu networks via convex programs. In *International Conference on Machine Learning*, 2021.

[13] Spencer Frei, Niladri S Chatterji, and Peter L Bartlett. Random feature amplification: Feature learning and generalization in neural networks. *Journal of Machine Learning Research*, 24 (303):1–49, 2023.

[14] Onnink R. Gerritsma, J. and A. Versluis. Yacht Hydrodynamics. UCI Machine Learning Repository, 1981.

[15] Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer. In *International Conference on Learning Representations*, 2026.

[16] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[17] Kedar Karhadkar, Michael Murray, Hanna Tseran, and Guido Montufar. Mildly overparameterized reLU networks have a favorable loss landscape. *Transactions on Machine Learning Research*, 2024.

[18] Sungyoon Kim, Aaron Mishkin, and Mert Pilanci. Exploring the loss landscape of regularized neural networks via convex duality. In *International Conference on Learning Representations*, 2025.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge, and Sanjeev Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. In *Advances in Neural Information Processing Systems*, 2019.

[21] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005.

[22] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, 2017.

[23] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, 2020.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[25] Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 2021.

[26] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 2018.

[27] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, 2019.

[28] Aaron Mishkin and Mert Pilanci. Optimal sets and solution paths of relu networks. In *International Conference on Machine Learning*, 2023.

[29] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010.

[30] Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, 2020.

[31] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[32] Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning*, 2021.

[33] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[34] Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Ridge regression with over-parametrized two-layer networks converge to ridgelet spectrum. In *International Conference on Artificial Intelligence and Statistics*, 2021.

[35] Yifei Wang, Jonathan Lacotte, and Mert Pilanci. The hidden convex optimization landscape of regularized two-layer reLU networks: an exact characterization of optimal solutions. In *International Conference on Learning Representations*, 2022.

[36] Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. *Advances in neural information processing systems*, 2019.

[37] Frank Zhengqing Wu, Berfin Simsek, and François Gaston Ged. Loss landscape of shallow reLU-like neural networks: Stationary points, saddle escape, and network embedding. In *International Conference on Learning Representations*, 2025.

[38] Shuo Xie and Zhiyuan Li. Implicit bias of adamw: ℓ_∞ norm constrained optimization. In *International Conference on Machine Learning*, 2024.

[39] Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, 2021.

[40] Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Understanding mode connectivity via parameter space symmetry. In *International Conference on Machine Learning*, 2025.

Appendix

Appendix Contents

A Related Work **10**

B Proofs for main **11**
B.1 Proofs for Subsection 2.1 11
B.2 Proofs for Subsection 2.2 13
B.3 Proofs for Subsection 2.2.1 16
B.4 Proofs for Subsection 2.2.2 22

C More analysis on connectivity result **26**
C.1 Roles of Permutation Symmetry and Inactive Neurons for Connectivity 26
C.2 Connectivity and Convex Formulation 27

D Proofs for Appendix **28**
D.1 Proofs for Appendix C.2 28

E Visualizations of global minima **37**

F More results from Numerical Experiments **37**
F.1 Yacht Hydrodynamics data 40
F.2 MNIST 40
F.3 Computing Environment 40

Appendix A. Related Work

Geometrical Analysis of Global minima.
There are many works analyzing the geometrical properties of the loss landscape [1, 9, 37]. Analyzing global minima has also been an active research topic [8, 20, 32, 40]. For example, Simsek et al. [32] explicitly describes the manifold of global minima.

Cooper [8] analyzes the dimension of the manifold in global optima for overparameterized neural networks. Zhao et al. [40] and Kuditipudi et al. [20] study global optima connectivity. The global optima for two-layer ReLU neural networks trained with ℓ_2-regularized loss are analyzed by Kim et al. [18]. While they use the convex formulation of neural networks introduced by Pilanci and Ergen [30] nicely to avoid the difficulty of analyzing non-convex loss, they did not provide an analytic solution for the globally optimal parameter set nor take into account the effects of scaling hyperparameters with respect to width. Inspired by their work, we explicitly consider the effects of width-dependent hyperparameter setting.

Convergence to Global minima.
Although the learned parameters do not necessarily converge to a global minimum, there are several works confirming the convergence to global minima [2, 6, 22, 23, 33, 34]. Akiyama and Suzuki [2] shows that, under a specific teacher-student setting, an overparameterized two-layer ReLU student trained with sparse/path-norm regularization, which is closely related to ℓ_2-regularization, and norm-dependent gradient descent can recover the teacher parameters with high probability. On the

other hand, Reddi et al. [31] shows that the Adam [19] algorithm fails to converge to a global optimum. We experimentally demonstrated differences in convergence results depending on the choice of optimizers in a setting distinct from previous work.

Hyperparameter setting and Training Dynamics.
Different training dynamics behaviors are observed to be dependent on a choice of hyperparmeters. In one regime (linear regime, lazy training, kernel regime), the training can be approximated by kernel gradient descent [7, 11]. In the other regime, nerual networks learn features beyond their initialization [10, 13]. A well-known example in this regime is mean-field regime [26, 27], and the training admits feature learning [39] or learns adaptively from samples [36]. These training dynamics regimes relate to the performance of neural networks. For example, the poor generalization performance of lazy training has been reported. [5, 7]. The hyperparameter settings for our numerical experiments includes both for the kernel regime and for the mean-field regime.

Appendix B. Proofs for main

B.1. Proofs for Subsection 2.1

Theorem 10 *(Theorem 1 in main) We consider minimizing the following loss function*

$$\min_{\theta\in\Theta} L_\ell(\theta) = \ell(\theta) + \frac{\beta}{2}\sum_{j=1}^{m}(u_j^2+\omega_j^2).$$

where ℓ is finite on Θ and can be written as a convex function of $f_\theta(X)$. For example, $\ell(\theta) = \|f_\theta(X)-Y\|^2$.
If $\delta < a$, then there exists $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$, $\operatorname{argmin}_{\theta\in\Theta} L_\ell(\theta) = \{(0,0)_{i=1}^m\}$.

Proof Theorem 10 is proven by the following lemmas. Lemma 11 proves that Theorem 10 holds when $\partial\tilde{\ell}(f_0(X)) \neq \emptyset$. Lemma 12 proves that $\partial\tilde{\ell}(f_0(X)) \neq \emptyset$ holds when ℓ is finite on Θ and can be written as a convex function of $f_\theta(X)$. ■

We now prove Lemma 11 and Lemma 12.

Lemma 11 *Assume that $\partial\tilde{\ell}(f_0(X)) \neq \emptyset$. If $\delta < a$, then there exists $m_0 \in \mathbb{N}$ such that for all $m \geq m_0$,*

$$\operatorname*{argmin}_{\theta\in\Theta} L_\ell(\theta) = \{(0,0)_{i=1}^m\}.$$

Proof $\ell(\theta) = \tilde{\ell}(f_\theta(X))$ for some convex function $\tilde{\ell}$. Fix an element $g_0 \in \partial\tilde{\ell}(f_0(X))$.
By the subgradient inequality, $\forall z \in \mathbb{R}^n,\ \tilde{\ell}(z) \geq \tilde{\ell}(f_0(X)) + \langle g_0, z\rangle$. By applying this inequality to $z = f_\theta(X)$, we obtain

$$L_\ell(\theta) \geq \tilde{\ell}(f_0(X)) + \langle g_0, f_\theta(X)\rangle + \frac{\beta}{2}\sum_{j=1}^{m}\big(\|u_j\|_2^2+\omega_j^2\big).$$

Now we are going to lower bound $\langle g_0, f_\theta(X)\rangle$.
For each j, by the operator norm inequality for a matrix X, $\|(Xu_j)_+\|_2 \leq \|Xu_j\|_2 \leq \|X\|_{\mathrm{op}}\|u_j\|_2$.

Hence,

$$\|f_\theta(X)\|_2 \le \frac{1}{\alpha}\sum_{j=1}^{m} |\omega_j|\,\|(Xu_j)_+\|_2 \le \frac{\|X\|_{\mathrm{op}}}{\alpha}\sum_{j=1}^{m} |\omega_j|\,\|u_j\|_2.$$

Using the basic inequality $2ab \le a^2 + b^2$, we further obtain$\sum_{j=1}^{m} |\omega_j|\,\|u_j\|_2 \le \frac{1}{2}\sum_{j=1}^{m}\big(\|u_j\|_2^2 + \omega_j^2\big)$.
Therefore,

$$\|f_\theta(X)\|_2 \le \frac{\|X\|_{\mathrm{op}}}{2\alpha}\sum_{j=1}^{m}\big(\|u_j\|_2^2 + \omega_j^2\big).$$

It follows that

$$\langle g_0, f_\theta(X)\rangle \ge -\|g_0\|_2\,\|f_\theta(X)\|_2 \ge -\frac{\|g_0\|_2\|X\|_{\mathrm{op}}}{2\alpha}\sum_{j=1}^{m}\big(\|u_j\|_2^2 + \omega_j^2\big).$$

By substituting this into the lower bound for $L_\ell(\theta)$, we obtain

$$L_\ell(\theta) \ge \tilde{\ell}(f_0(X)) + \frac{1}{2}\left(\beta - \frac{\|g_0\|_2\|X\|_{\mathrm{op}}}{\alpha}\right)\sum_{j=1}^{m}\big(\|u_j\|_2^2 + \omega_j^2\big).$$

In our hyperparameter setting, $\alpha\beta = m^{a-\delta}$, so the assumption $\delta < a$ implies $\alpha\beta \to \infty$ as $m \to \infty$. Therefore, there exists $m_0 \in \mathbb{N}$ such that for all $m \ge m_0$, $\beta > \frac{\|g_0\|_2\|X\|_{\mathrm{op}}}{\alpha}$. For such m, we have

$$L_\ell(\theta) \ge \tilde{\ell}(f_0(X)) + c_m\sum_{j=1}^{m}\big(\|u_j\|_2^2 + \omega_j^2\big) \text{ for some positive constant } c_m > 0.$$

On the other hand,

$$L_\ell((0,0)_{i=1}^m) = \tilde{\ell}(f_0(X)).$$

Thus, for every nonzero θ,

$$L_\ell(\theta) > L_\ell((0,0)_{i=1}^m).$$

Hence, $(0,0)_{i=1}^m$ is the unique global minimizer of L_ℓ for all sufficiently large m. ■

Lemma 12 *If $\tilde{\ell}$ is finite and convex on all of $\mathbb{R}^n$, the assumption $\partial\tilde{\ell}(f_0(X)) \neq \emptyset$ holds. If $\tilde{\ell}$ is differentiable at 0, one may take $g_0 = \nabla\tilde{\ell}(f_0(X))$ in the proof of Lemma 11.*

Proof Recall that the subdifferential of $\tilde{\ell}$ at 0 is $\partial\tilde{\ell}(f_0(X)) = \Big\{g \in \mathbb{R}^n : \ \tilde{\ell}(z) \ge \tilde{\ell}(f_0(X)) + \langle g, z\rangle \quad \forall z \in \mathbb{R}^n\Big\}$.
If $\tilde{\ell}$ is differentiable at 0, $\partial\tilde{\ell}(f_0(X)) = \{\nabla\tilde{\ell}(f_0(X))\}$.

Now we assume more generally that $\tilde{\ell} : \mathbb{R}^n \to \mathbb{R}$ is finite and convex. Consider the epigraph

$$\mathrm{epi}(\tilde{\ell}) = \{(z, r) \in \mathbb{R}^n \times \mathbb{R} : r \ge \tilde{\ell}(z)\}.$$

As $\tilde{\ell}$ is convex and finite everywhere, $\tilde{\ell}$ is locally Lipschitz at every point, and so is continuous on all of $\mathbb{R}^n$. Take any sequence $(z_k, r_k) \in \mathrm{epi}(\tilde{\ell})$ such that $(z_k, r_k) \to (z, r) \in \mathbb{R}^n \times \mathbb{R}$. Then,

$r_k \geq \tilde{\ell}(z_k)$ for all k, and by continuity of $\tilde{\ell}$, $r \geq \tilde{\ell}(z)$. Hence, $(z, r) \in \text{epi}(\tilde{\ell})$, and so $\text{epi}(\tilde{\ell})$ is closed.

As $\tilde{\ell}$ is convex, $\text{epi}(\tilde{\ell})$ is a convex set. The point $(0, \tilde{\ell}(f_0(X)))$ lies on the boundary of $\text{epi}(\tilde{\ell})$. By the supporting hyperplane theorem, there exist non-zero $(a, b) \neq (0, 0) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$\forall (z, r) \in \text{epi}(\tilde{\ell}),\ \langle a, z\rangle + br \geq \langle a, 0\rangle + b\tilde{\ell}(f_0(X)).$$

We prove $b > 0$ by contradiction. If $b < 0$, then for any fixed z, letting $r \to \infty$ contradicts the inequality above. If $b = 0$, $\langle a, z\rangle \geq 0$ for all $z \in \mathbb{R}^n$. Applying this also to $-z$ yields $\langle a, z\rangle = 0$ for all z. This leads to $a = 0$, a contradiction. Therefore, $b > 0$.

For every $z \in \mathbb{R}^n$, the point $(z, \tilde{\ell}(z))$ belongs to $\text{epi}(\tilde{\ell})$, so $\langle a, z\rangle + b\tilde{\ell}(z) \geq b\tilde{\ell}(f_0(X))$. By rearranging the equations,

$$\forall z \in \mathbb{R}^n,\ \tilde{\ell}(z) \geq \tilde{\ell}(f_0(X)) + \left\langle -\frac{a}{b},\, z\right\rangle.$$

Hence, $-\frac{a}{b} \in \partial\tilde{\ell}(f_0(X))$ and therefore $\partial\tilde{\ell}(f_0(X)) \neq \emptyset$. ■

B.2. Proofs for Subsection 2.2

Theorem 13 *(Theorem 2 in main) The set of global optimal parameters*

$$\varphi^*(m) = \{\theta \in \mathbb{R}^{2m}\ :\ \arg\min_{\theta\in\mathbb{R}^{2m}} \frac{1}{2}\,\|f_\theta(x) - Y\|_2^2\}$$

for squared loss is

$$\varphi^*(m) := \Big\{(u_j, \omega_j)_{j=1}^m \ \Big|\ \sum_{j:\,u_j\geq 0} u_j\omega_j = \frac{\alpha X_S^\top Y}{\|X_S\|_2^2},\ \sum_{j:\,u_j\leq 0} u_j\omega_j = \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}\Big\}.$$

Proof By writing $\gamma_P = \sum_{i:u_i\geq 0} u_i\omega_i$ and $\gamma_N = \sum_{i:u_i\leq 0} u_i\omega_i$,

$$\|f_\theta(x) - Y\|_2^2 = \Big\|\frac{1}{\alpha}\sum_{j=1}^m (Xu_j)_+\omega_j - Y\Big\|_2^2 = \|\,\gamma_P\frac{X_S}{\alpha} + \gamma_N\frac{X_{S^c}}{\alpha} - Y\,\|_2^2.$$

By solving the KKT condition, (LHS) is minimized at $(\gamma_P,\ \gamma_N) = (\gamma_P^*, \gamma_N^*)$ where

$$0 = \gamma_P^*\|\frac{X_S}{\alpha}\|_2^2 - \frac{X_S}{\alpha}^T Y,\ 0 = \gamma_N^*\|\frac{X_{S^c}}{\alpha}\|_2^2 - \frac{X_{S^c}}{\alpha}^T Y$$
$$\iff \gamma_P^* = \frac{\alpha X_S^T Y}{\|X_S\|_2^2},\ \gamma_N^* = \frac{\alpha X_{S^c} Y}{\|X_{S^c}\|_2^2}.$$

Hence,

$$\Big\|\frac{1}{\alpha}\sum_{j=1}^m (Xu_j)_+\omega_j - Y\Big\|_2^2 \geq \|\,\gamma_P^*\frac{X_S}{\alpha} + \gamma_N^*\frac{X_{S^c}}{\alpha} - Y\,\|_2^2.$$

with equality iff $\sum_{j:\,u_j\geq 0} u_j\omega_j = \frac{\alpha X_S^\top Y}{\|X_S\|_2^2}$ and $\sum_{j:\,u_j\leq 0} u_j\omega_j = \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}$. ■

For the one-dimensional input case, we prove that the minimum value for the original non-convex problem (3) with $\beta > 0$ equals the minimum value for the convex problem (6) in Lemma 14.

Lemma 14 *Consider the optimization problem*

$$\min_{\gamma_P,\gamma_N\in\mathbb{R}} \frac{1}{2}\| \gamma_P \frac{X_S}{\alpha} + \gamma_N \frac{X_{S^c}}{\alpha} - Y \|_2^2 + \beta(|\gamma_P| + |\gamma_N|). \tag{6}$$

The optimal argument (γ_P^*, γ_N^*) is uniquely determined as

$$(\gamma_P^*, \gamma_N^*) = \left(\alpha \frac{\mathcal{S}_{\alpha\beta}(X_S^T Y)}{\|X_S\|_2^2}, \alpha \frac{\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y)}{\|X_{S^c}\|_2^2} \right). \tag{7}$$

where $\mathcal{S}_\beta(b) := sign(b)\max(|b| - \beta, 0)$ is the softmax function.

Proof Note that the minimization problem (6) is a convex minimization problem. By KKT condition,

$$0 \in \{\gamma_P \| \frac{X_S}{\alpha}\|_2^2 - \frac{X_S}{\alpha}^T Y + \beta S \mid S \in \partial|\gamma_P|\}$$

at $\gamma_P = \gamma_P^*$. $\partial|\gamma_P|$ is the subdifferential of the absolute value function $|\gamma_P|$ at γ_P found as

$$\partial\,|\gamma_P| = \begin{cases} \{1\}, & \gamma_P > 0, \\ [-1,\,1], & \gamma_P = 0, \\ \{-1\}, & \gamma_P < 0. \end{cases}$$

Substituting $\partial|\gamma_P|$ gives the solution form as follows

$$\gamma_P^* = \frac{sign(\frac{X_S}{\alpha}^T Y)\ \max(|\frac{X_S}{\alpha}^T Y| - \beta, 0)}{\|\frac{X_S}{\alpha}\|_2^2} = \frac{\mathcal{S}_\beta(\frac{X_S}{\alpha}^T Y)}{\|\frac{X_S}{\alpha}\|_2^2} = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_S^T Y)}{\|X_S\|_2^2}.$$

$\gamma_N^* = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y)}{\|X_{S^c}\|_2^2}$ is found by replacing X_S by X_{S^c}. ■

We derive the exact solution to the global minima for regularized squared loss by using Lemma 14

Theorem 15 *(Theorem 3 in the main). The global optimum $\Theta^*(m) = \{\theta \in \mathbb{R}^{2m} \ :\ \arg\min_{\theta\in\mathbb{R}^{2m}} L(\theta)\}$ with $\beta > 0$ is*

$$\Theta^*(m) = \left\{ (u_i, \omega_i)_{i=1}^m \;\middle|\; \sum_{i:\,u_i\geq 0} u_i^2 = |\gamma_P^*|,\ \sum_{i:\,u_i\leq 0} u_i^2 = |\gamma_N^*|,\ \omega_i = \begin{cases} \mathrm{sign}(\gamma_P^*)\, u_i & (u_i > 0), \\ \mathrm{sign}(\gamma_N^*)\, u_i & (u_i < 0), \\ 0 & (u_i = 0) \end{cases} \right\} \tag{8}$$

where $\gamma_P^* = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_S^T Y)}{\|X_S\|_2^2}$ and $\gamma_N^* = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y)}{\|X_{S^c}\|_2^2}$.

Proof By writing

$$\gamma_P = \sum_{i:u_i\geq 0} u_i\omega_i, \quad \gamma_N = \sum_{i:u_i\leq 0} u_i\omega_i, \quad X_S = \mathrm{Diag}(\mathbf{1}[X \geq 0])X, \quad \text{and} \quad X_{S^c} = \mathrm{Diag}(\mathbf{1}[X \leq 0])X,$$

we can write $\left\|\frac{1}{\alpha}\sum_{j=1}^m (Xu_j)_+\omega_j - Y\right\|_2^2$ as $\| \gamma_P \frac{X_S}{\alpha} + \gamma_N \frac{X_{S^c}}{\alpha} - Y \|_2^2$. Also,

$$\begin{aligned}
&\frac{\beta}{2}\sum_{j=1}^m (u_j^2+\omega_j^2) \\
&\geq \frac{\beta}{2}\sum_{j:u_j>0}(u_j^2+\omega_j^2) + \frac{\beta}{2}\sum_{j:u_j<0}(u_j^2+\omega_j^2) \quad \text{(equality holds iff } u_j=0 \Rightarrow \omega_j=0) && (9)\\
&\geq \beta\sum_{j:u_j\geq 0}|u_j||\omega_j| + \beta\sum_{j:u_j\leq 0}|u_j||\omega_j| = \beta\sum_{j:u_j\geq 0}|u_j\omega_j| + \beta\sum_{j:u_j\leq 0}|u_j\omega_j| \quad \text{(equality holds iff } \forall\, j,\ |u_j|=|\omega_j|) && (10)\\
&\geq \beta\left|\sum_{j:u_j\geq 0} u_j\omega_j\right| + \beta\left|\sum_{j:u_j\leq 0} u_j\omega_j\right| = \beta|\gamma_P| + \beta|\gamma_N| && (11)
\end{aligned}$$

(equality holds iff $\forall j, i$ s.t. the product $u_j u_i > 0,\ \mathrm{sign}(u_j\omega_j) = \mathrm{sign}(u_i\omega_i)$).

Hence,

$$\begin{aligned}
&\min_{\{u_i,\omega_i\}_{i=1}^m \in (\mathbb{R}\times\mathbb{R})^m} \frac{1}{2}\left\|\frac{1}{\alpha}\sum_{j=1}^m (Xu_j)_+\omega_j - Y\right\|_2^2 + \frac{\beta}{2}\sum_{j=1}^m (u_j^2+\omega_j^2) \\
&\geq \min_{\gamma_P,\gamma_N\in\mathbb{R}} \frac{1}{2}\| \gamma_P \frac{X_S}{\alpha} + \gamma_N \frac{X_{S^c}}{\alpha} - Y \|_2^2 + \beta(|\gamma_P| + |\gamma_N|) && (12)\\
&= \frac{1}{2}\| \gamma_P^* \frac{X_S}{\alpha} + \gamma_N^* \frac{X_{S^c}}{\alpha} - Y \|_2^2 + \beta(|\gamma_P^*| + |\gamma_N^*|).
\end{aligned}$$

The equality in the last holds by Lemma 14 where (γ_P^*, γ_N^*) is defined as (7).
By (9), (10) and (11),

$$\frac{1}{2}\left\|\frac{1}{\alpha}\sum_{j=1}^m (Xu_j)_+\omega_j - Y\right\|_2^2 + \frac{\beta}{2}\sum_{j=1}^m (u_j^2+\omega_j^2) = \frac{1}{2}\left\|\gamma_P^* \frac{X_S}{\alpha} + \gamma_N^* \frac{X_{S^c}}{\alpha} - Y\right\|_2^2 + \beta|\gamma_P^*| + \beta|\gamma_N^*|$$

holds iff $(u_i,\omega_i)_{i=1}^m = (u_i^*,\omega_i^*)_{i=1}^m$ such that

$$|\gamma_P^*| = \left|\sum_{i:u_i^*\geq 0} u_i^*\omega_i^*\right| = \left|\sum_{i:u_i^*\geq 0} u_i^{*2}\right| = \sum_{i:u_i^*\geq 0} u_i^{*2}, \tag{13}$$

$$|\gamma_N^*| = \left|\sum_{i:u_i^*\leq 0} u_i^*\omega_i^*\right| = \sum_{i:u_i^*\leq 0} u_i^{*2}, \tag{14}$$

$$\omega_i^* = \begin{cases} \mathrm{sign}(\gamma_P^*)u_i^* & (u_i^*>0), \\ \mathrm{sign}(\gamma_N^*)u_i^* & (u_i^*<0), \\ 0 & (u_i^*=0). \end{cases} \tag{15}$$

(10) and (11) imply that $\sum_{i:u_i^*\geq 0} u_i^*\omega_i^* = \sum_{i:u_i^*\geq 0} u_i^{*2}$ or $-\sum_{i:u_i^*\geq 0} u_i^{*2}$. This implies (13). Similarly, (10) and (11) imply (14).

To satisfy $\gamma_P^* = \sum_{i:u_i^*\geq 0} u_i^*\omega_i^*$ and $\gamma_N^* = \sum_{i:u_i^*\leq 0} u_i^*\omega_i^*$ under (10) and (11), we need (15). Hence, $(u_i^*, \omega_i^*)_{i=1}^m \in \Theta^*(m)$ if and only if it satisfies (13) (14) (15). ■

B.3. Proofs for Subsection 2.2.1

Theorem 16 *(Theorem 4 in main) We have the phase transitional behavior of the solution set for the unregularized squared loss.*
(1) For $m = 1$, $\varphi^*(m) = \emptyset$.
(1) For $m = 2$, $\varphi^*(m)$ has exactly 2 connected components.
(2) For $m \geq 3$, $\varphi^*(m)$ is connected.

Proof (1) $m = 1$:
Since $0 < \min\{|X_S^T Y|,\ |X_{S^C}^T Y|\}$, we need at least two non-zero neurons to satisfy $\sum_{j:\,u_j\geq 0} u_j\omega_j = \frac{\alpha X_S^\top Y}{\|X_S\|_2^2}$ and $\sum_{j:\,u_j\leq 0} u_j\omega_j = \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}$.

(2) $m = 2$:
The sign pattern for u_1 and u_2 is $(u_1, u_2) = (+, -)$ or $(-, +)$. For $(u_1, u_2) = (+, -)$, $(\omega_1,\ \omega_2) = \left(\frac{\alpha X_S^\top Y}{u_1\|X_S\|_2^2},\ \frac{\alpha X_{S^c}^\top Y}{u_2\|X_{S^c}\|_2^2}\right)$. For $(u_1, u_2) = (-, +)$, $(\omega_1,\ \omega_2) = \left(\frac{\alpha X_{S^c}^\top Y}{u_1\|X_{S^c}\|_2^2},\ \frac{\alpha X_S^\top Y}{u_2\|X_S\|_2^2}\right)$.
Hence, $\varphi^*(m)$ consists of two connected components $\varphi_1^*(m)$ and $\varphi_2^*(m)$ such that

$$\varphi_1^*(m) = \left\{(u_i, \omega_i)_{i=1}^2 \middle| u_1 > 0,\ u_2 < 0,\ \omega_1 = \frac{\alpha X_S^\top Y}{u_1\|X_S\|_2^2},\ \omega_2 = \frac{\alpha X_{S^c}^\top Y}{u_2\|X_{S^c}\|_2^2}\right\},$$
$$\varphi_2^*(m) = \left\{(u_i, \omega_i)_{i=1}^2 \middle| u_2 > 0,\ u_1 < 0,\ \omega_2 = \frac{\alpha X_S^\top Y}{u_2\|X_S\|_2^2},\ \omega_1 = \frac{\alpha X_{S^c}^\top Y}{u_1\|X_{S^c}\|_2^2}\right\}.$$

$\varphi_1^*(m)$ and $\varphi_2^*(m)$ are disjoint because $u_1 \neq 0$ and $u_2 \neq 0$.

(3) $m \geq 3$:
For simplicity, write $A := \frac{\alpha X_S^\top Y}{\|X_S\|_2^2}$ and $B := \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}$.
Fix any $\theta = (u, \omega) \in \varphi^*(m)$, where $u = (u_1, \ldots, u_m)$ and $\omega = (\omega_1, \ldots, \omega_m)$. Define the strict sign index sets

$$P(u) := \{j \in [m] : u_j > 0\}, \qquad N(u) := \{j \in [m] : u_j < 0\}.$$

We must have $P(u) \neq \emptyset$ and $N(u) \neq \emptyset$. Choose any indices $p \in P(u)$ and $n \in N(u)$.

Keep u fixed and define $\omega(t)$ for $t \in [0, 1]$ by

$$\omega_j(t) := (1 - t)\omega_j \quad \text{for all } j \notin \{p, n\},$$

and

$$\omega_p(t) := \omega_p + \frac{t}{u_p} \sum_{\substack{j\in P(u)\\ j\neq p}} u_j\omega_j, \qquad \omega_n(t) := \omega_n + \frac{t}{u_n} \sum_{\substack{j\in N(u)\\ j\neq n}} u_j\omega_j.$$

Then, for all $t \in [0,1]$,

$$\sum_{j:\, u_j \geq 0} u_j \omega_j(t) = u_p \omega_p(t) + \sum_{\substack{j \in P(u) \\ j \neq p}} u_j (1-t) \omega_j = \sum_{j \in P(u)} u_j \omega_j = A,$$

and likewise

$$\sum_{j:\, u_j \leq 0} u_j \omega_j(t) = u_n \omega_n(t) + \sum_{\substack{j \in N(u) \\ j \neq n}} u_j (1-t) \omega_j = \sum_{j \in N(u)} u_j \omega_j = B.$$

Hence $\theta(t) \coloneqq (u, \omega(t)) \in \varphi^*(m)$ for all t. At $t = 1$, we have $\omega_j(1) = 0$ for all $j \notin \{p, n\}$, so only the two indices p and n carry the constraints $u_p \omega_p(1) = A,\ u_n \omega_n(1) = B$ and all other products are $u_j \omega_j(1) = 0$.

After this deformation, for every $j \notin \{p, n\}$, keep ω_j fixed and define $u_j(t)$ for $t \in [0,1]$ by

$$u_j(t) \coloneqq u_j(1-t).$$

For other indices, define as follows

$$u_p(t) \coloneqq \begin{cases} u_p & \text{if} \quad u_p = 1, \\ \frac{u_p}{1+(u_p-1)t} & \text{if} \quad u_p \neq 1, \end{cases} \qquad \omega_p \coloneqq \frac{A}{u_p(t)},$$

$$u_n(t) \coloneqq \begin{cases} u_n & \text{if} \quad u_n = -1, \\ \frac{u_n}{1+(-u_n-1)t} & \text{if} \quad u_n \neq -1, \end{cases} \qquad \omega_n \coloneqq \frac{B}{u_n(t)}.$$

Then, $u_p(t) > 0$ and $u_n(t) < 0$ for all $t \in [0,1]$. $u_j(t)\omega_j = 0$ for all $t \in [0,1]$. $u_p(t)\omega_p(t)$ and $u_n(t)\omega_n(t)$ are unchanged over $t \in [0,1]$. Hence, both constraints remain unchanged.
Thus, $\theta \in \varphi^*(m)$ is connected to a point in $\varphi^*_{min}(m)$ where

$$\varphi^*_{min}(m) = \left\{ (u_i, \omega_i)_{i=1}^m \middle| \exists\, p, n \in [m]\ s.t.\ (u_j, \omega_j) = (0,0) \text{ for all } j \notin \{p,n\},\ u_p = 1,\ \omega_p = A,\ u_n = -1,\ \omega_n = -B \right\}$$

To prove the connectivity of $\varphi^*(m)$, it is enough to show that all elements in $\varphi^*_{min}(m)$ are connected in $\varphi^*(m)$.
All elements in $\varphi^*_{min}(m)$ are permutations of each other. Denote by S_m the symmetric group on $\{1, 2, \ldots, m\}$. S_m is the set of all permutations on $\{1, 2, \ldots, m\}$. It is enough to show that

$$\forall (u_i, \omega_i)_{i=1}^m \in \varphi^*_{min}(m),\ \forall \sigma \in S_m,\ \exists \text{a continuous path in } \varphi^*(m) \text{ connecting } (u_i, \omega_i)_{i=1}^m \text{ and } (u_{\sigma(i)}, \omega_{\sigma(i)})_{i=1}^m.$$

Pick any $(u_i,\ \omega_i)_{i=1}^m \in \varphi^*_{min}(m)$. As $m \geq 3$ and $(u_i,\ \omega_i)_{i=1}^m$ has exactly two non-zero elements, $\exists\, p, n, j \in [m]$ such that $u_p = 1,\ u_n = -1,\ u_j = 0$. Also, for all $j \in [m]$, the set of transpositions $\mathcal{T}_j \coloneqq \{(i,j) \mid i \in [m]\ s.t.\ i \neq j\}$ (We denote by (i,j) a transposition) generates S_m, so it is enough to show that

$$\forall\, T \in \mathcal{T}_j \text{ where } u_j = 0,\ \exists \text{ a continuous path in } \varphi^*(m) \text{ connecting } (u_i, \omega_i)_{i=1}^m \text{ and } (u_{T(i)}, \omega_{T(i)})_{i=1}^m.$$

Take any $T \in \mathcal{T}_j$.

(case 1) If $u_{T(j)} = 0$ ($T = (i, j)$ for $i \neq n, p$), then $(u_i, \omega_i)_{i=1}^m = (u_{T(i)}, \omega_{T(i)})_{i=1}^m$.

(case 2) Consider the case $u_{T(j)} = u_p$ ($T = (p, j)$).
Construct a path C as

$$C(s) = (u_i(s), \omega_i(s))_{i=1}^m, \ s \in [0, 1]$$

s.t.

$$\begin{aligned}
&(u_p(s), \omega_p(s)) = (\sqrt{1-s}, A\sqrt{1-s}),\\
&(u_j(s), \omega_j(s)) = (\sqrt{s}, A\sqrt{s}),\\
&(u_i(s), \omega_i(s)) = (u_i, \omega_i) \ \forall\, i \neq j, p.
\end{aligned}$$

$C(s)$ is well-defined and connected. As $u_j(s),\ u_p(s) \geq 0$ and $u_j(s)\omega_j(s) + u_p(s)\omega_p(s) = A\ \forall s \in [0, 1]$, $C(s) \in \varphi^*(m)$. $C(0) = (u_i, \omega_i)_{i=1}^m$ and $C(1) = (u_{T(i)}, \omega_{T(i)})_{i=1}^m$. Thus, $C(s)$ is a continuous path in $\varphi^*(m)$ connecting $(u_i, \omega_i)_{i=1}^m$ and $(u_{T(i)}, \omega_{T(i)})_{i=1}^m$.

(case 3) Consider the case $u_{T(j)} = u_n$ ($T = (n, j)$).
Construct a path C as

$$C(s) = (u_i(s), \omega_i(s))_{i=1}^m, \ s \in [0, 1]$$

s.t.

$$\begin{aligned}
&(u_n(s), \omega_n(s)) = (-\sqrt{1-s}, -B\sqrt{1-s}),\\
&(u_j(s), \omega_j(s)) = (-\sqrt{s}, -B\sqrt{s}),\\
&(u_i(s), \omega_i(s)) = (u_i, \omega_i) \ \forall\, i \neq j, n.
\end{aligned}$$

$C(s)$ is well-defined and connected. As $u_j(s),\ u_n(s) \leq 0$ and $u_j(s)\omega_j(s) + u_n(s)\omega_n(s) = B\ \forall s \in [0, 1]$, $C(s) \in \varphi^*(m)$. $C(0) = (u_i, \omega_i)_{i=1}^m$ and $C(1) = (u_{T(i)}, \omega_{T(i)})_{i=1}^m$. Thus, $C(s)$ is a continuous path in $\varphi^*(m)$ connecting $(u_i, \omega_i)_{i=1}^m$ and $(u_{T(i)}, \omega_{T(i)})_{i=1}^m$.

Thus, the claim is proved. ■

We firstly prove the connectivity result of global optimal parameters $\Theta^*(m)$ using notations $\alpha = m^a$ and $\beta = m^{-\delta}$ in two different ways.

Theorem 17 *We have a critical width $M^* = 1[\mathcal{S}_{\alpha\beta}(X_S^T Y) \neq 0] + 1[\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y) \neq 0]$ that determines the phase transitional behavior of the solution set.*
(1) For $M^* = 0$, $\Theta^*(m)$ is a singleton ($\{(0, 0)_{i=1}^m\}$)
(2) For $m < M^*$, $\Theta^*(m) = \emptyset$
(3) For $m = M^* > 0$, $\Theta^*(m)$ is a finite set.
(4) For $m > M^* > 0$, $\Theta^*(m)$ is connected.

The first proof is simply to use the explicit form of $\Theta^*(m)$. The second proof follows the principal ideas introduced by Kim et al. [18]. We provide the first proof in this section, and the second proof is given in the next section (Appendix D.1).
Proof (The first proof)

$$M^* = 1[\mathcal{S}_{\alpha\beta}(X_S^T Y) \neq 0) + 1[\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y) \neq 0] = 1[|\gamma_P^*| \neq 0] + 1[|\gamma_N^*| \neq 0].$$

(1) $M^* = 0$:
$M^* = 0$, implies $|\gamma_P^*| = 0$ and $|\gamma_N^*| = 0$. For $(u_i, \omega_1)_{i=1}^m \in \Theta^*(m)$, $\sum_{i:\, u_i \geq 0} u_i^2 = 0$ and $\sum_{i:\, u_i \leq 0} u_i^2 = 0$, so $\forall i \in [m],\ u_i = 0$. Also, $\forall i \in [m],\ |u_i| = |\omega_i|$, so $\forall i \in [m],\ \omega_i = 0$.

(2) $m < M^*$:
$\Theta^*(m) = \emptyset$ because we need at least M^* non-zero neurons for $\sum_{i:\, u_i \geq 0} u_i^2 = |\gamma_P^*|$ and $\sum_{i:\, u_i \leq 0} u_i^2 = |\gamma_N^*|$ to hold.

(3) $m = M^*$:
To satisfy (13) (14) (15),

$$(u_i, \omega_i)_{i=1}^{M^*} \in \Theta^*(M^*) \Rightarrow \forall i \in [M^*],\ (u_i,\ \omega_i) \in \Big\{\text{non-zero elements in } \{(\sqrt{|\gamma_P^*|}, sign(\gamma_P^*)\sqrt{|\gamma_P^*|}), (-\sqrt{|\gamma_N^*|}, -sign(\gamma_N^*)\sqrt{|\gamma_N^*|})\}\Big\}$$

Hence, $|\Theta^*(m)| \leq 2^{M^*}$ and $\Theta^*(m)$ is a finite set.

(4) $m > M^*$:
Define the u-projection

$$U^*(m) := \left\{ u \in \mathbb{R}^m \,\middle|\, \sum_{i:u_i>0} u_i^2 = |\gamma_P^*|,\ \sum_{i:u_i<0} u_i^2 = |\gamma_N^*| \right\}.$$

Define $\Omega : \mathbb{R}^m \to \mathbb{R}^m$ coordinatewise by

$$\Omega_i(u) = \begin{cases} \operatorname{sign}(\gamma_P^*) u_i & (u_i > 0), \\ \operatorname{sign}(\gamma_N^*) u_i & (u_i < 0), \\ 0 & (u_i = 0). \end{cases}$$

$|\operatorname{sign}(\gamma_P^*)|,\ |\operatorname{sign}(\gamma_N^*)| \leq 1$ implies that $|\Omega_i(a) - \Omega_i(b)| \leq |a - b|$, so Ω_i is Lipschitz, hence is continuous. As each Ω_i is continuous, Ω is continuous. We can write $\Theta^*(m)$ as

$$\Theta^*(m) = \{(u, \Omega(u)) |\ u \in U^*(m)\}.$$

Therefore, the map $F : U^*(m) \to \Theta^*(m)$ defined as $F(u) = (u, \Omega(u))$ is a homeomorphism with inverse given by the projection $(u, \omega) \mapsto u$. Thus, it suffices to prove that $U^*(m)$ is path-connected. Define the minimal optimal subset as

$$U_{min}^*(m) := \left\{ \sqrt{|\gamma_P^*|}\, e_i - \sqrt{|\gamma_N^*|}\, e_j \,\middle|\, i, j \in \{1, \ldots, m\},\ i \neq j \right\} \subset U^*(m),$$

where $(e_i)_{i=1}^m$ are the standard basis vectors.
We prove the following two claims.

1. Elements in $U_{min}^*(m)$ are connected to each other in $U^*(m)$ when $m > M^*$.

   Proof We prove by a direct construction of a path. Denote by S_m the symmetric group on $\{1, 2, \ldots, m\}$. S_m is the set of all permutations on $\{1, 2, \ldots, m\}$. Denote $u = (u_1, \ldots, u_m) \in$

$\mathbb{R}^m$ by $(u_i)_{i=1}^m$.
From the construction of $U^*_{min}(m)$, it is enough to show that

$\forall (u_i)_{i=1}^m \in U^*_{min}(m)$, $\forall \sigma \in S_m$, $\exists$a continuous path in $U^*(m)$ connecting $(u_i)_{i=1}^m$ and $(u_{\sigma(i)})_{i=1}^m$.

Pick any $(u_i)_{i=1}^m \in U^*_{min}(m)$. As $m > M^* = 1(|\gamma^*_P| \neq 0) + 1(|\gamma^*_N| \neq 0)$, $\exists j \in [m]$ such that $u_j = 0$. Also, for all $j \in [m]$, the set of transpositions $T_j := \{(i,j) \mid i \in [m] \ s.t. \ i \neq j\}$ (We denote by (i,j) a transposition) generates S_m, so it is enough to show that

$\forall t \in T_j$ where $u_j \neq 0$, $\exists$ a continuous path in $U^*(m)$ connecting $(u_i)_{i=1}^m$ and $(u_{T(i)})_{i=1}^m$.

Take any $t \in T_j$. If $u_{t(j)} = 0$, $(u_i)_{i=1}^m = (u_{t(i)})_{i=1}^m$. If $u_{t(j)} \neq 0$, construct a path C as

$$C(s) = (u_i(s))_{i=1}^m, \ s \in [0,1]$$

s.t.

$$u_i(s) = \begin{cases} u_i & \text{if} \quad i \neq j \text{ and } i \neq t(j), \\ u_{t(j)}\sqrt{s} & \text{if} \quad i = j, \\ u_{t(j)}\sqrt{1-s} & \text{if} \quad i = t(j). \end{cases}$$

By direct calculation, $C(s)$ is a continuous path in $U^*(m)$ that connects $(u_i)_{i=1}^m$ and $(u_{t(i)})_{i=1}^m$. Thus, the claim holds. ■

2. Elements in $U^*(m)$ are connected to an element in $U^*_{min}(m)$.

Proof We prove by direct construction of a path along which we decrease the number of non-zero elements. Take any $u \in U^*(m)$ and apply the following steps.

Step 1
Define sets of indices.

$$P := \{i \in [m] : u_i > 0\}, \ \ N := \{i \in [m] : u_i < 0\}$$

If $|P| \leq 1$ and $|N| \leq 1$, u is a point in $U^*_{min}(m)$. If not, move to Step 2.

Step 2
Since $|P| > 1$ or $|N| > 1$, $\exists k < l$ such that $k, l \in P$ or $k, l \in N$. For such k, l, construct a merging path $M(s)$ as

$$M(s) = (u_i(s))_{i=1}^m, \ s \in [0,1]$$

s.t.

$$u_i(s) = \begin{cases} u_i & \text{if} \quad i \neq k \text{ and } i \neq l, \\ \sqrt{u_k^2 u_l^2}\cos(\theta(1-s)) & \text{if} \quad i = k, \\ \sqrt{u_k^2 u_l^2}\sin(\theta(1-s)) & \text{if} \quad i = l. \end{cases}$$

$\theta \in [0, 2\pi)$ satisfies $u_k = \sqrt{u_k^2 u_l^2}\cos\theta$ and $u_l = \sqrt{u_k^2 u_l^2}\sin\theta$. By direct calculation, $M(s)$ is a continuous path in $U^*(m)$ and $M(0) = u$. Apply Step 1 to $M(1)$ and repeat this merging process until we reach a point in $U^*_{min}(m)$. The number of non-zero elements in $M(1)$ is less than that of $M(0)$, so the process terminates in a finite number of steps. ■

These two claims prove that $\Theta^*(m)$ is connected for $m > M^*$. ■

The second proof is provided in Appendix D.1.

Theorem 18 *Assume $a < \delta$. Define $M^*(m) = 1[|X_S^T Y| > m^{a-\delta}] + 1[|X_{S^c}^T Y| > m^{a-\delta}]$. Define*

$$\underline{M} = \min\{m \in \mathbb{N}_{\geq 1} \mid M^*(m) \geq 1\},$$
$$\overline{M} = \min\{m \in \mathbb{N}_{\geq 1} \mid M^*(m) = 2\}.$$

$M^*(m) \in \{0, 1, 2\}$ is increasing with m and $M^*(m) = 2$ for sufficiently large m. Hence, $\underline{M}$ and $\overline{M}$ are well-defined.
We have the following connectivity results.
(1) For $m < \underline{M}$, $\Theta^*(m)$ is a singleton ($\{(0,0)_{i=1}^m\}$).
(2) If $\underline{M} = \overline{M} = m = 1$, $\Theta^*(m) = \emptyset$.
(3) If $\underline{M} \leq m = 1 < \overline{M}$ or $\underline{M} \leq \overline{M} \leq m = 2$, $\Theta^*(m)$ is a finite set.
(4) Otherwise, $\Theta^*(m)$ is connected.

Proof $m^{a-\delta}$ is strictly decreasing and $m^{a-\delta} \to 0$ as $m \to \infty$. Hence, $M^*(m) \in \{0, 1, 2\}$ is increasing with m and $M^*(m) = 2$ for sufficiently large m.
As the product $\alpha\beta$ depends on m, $\gamma_P^* = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_S^T Y)}{\|X_S\|_2^2}$ and $\gamma_N^* = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y)}{\|X_{S^c}\|_2^2}$ depends on m.

$$M^*(m) = 1[|X_S^T Y| > m^{a-\delta}] + 1[|X_{S^c}^T Y| > m^{a-\delta}] = 1[|\gamma_P^*| > 0] + 1[|\gamma_N^*| > 0].$$

Also,

$$M^*(m) = 0 \iff m < \underline{M}, \tag{16}$$
$$M^*(m) = 1 \iff \underline{M} \leq m < \overline{M}, \tag{17}$$
$$M^*(m) = 2 \iff \overline{M} \leq m. \tag{18}$$

(1) $m < \underline{M}$:
From (16), $M^*(m) = 0$, so the result from Theorem 17 (1) implies that $\Theta^*(m)$ is a singleton.
(2) $\underline{M} = \overline{M} = m = 1$:
We need at least $M^*(m)$ non-zero neurons for $\sum_{i:\, u_i \geq 0} u_i^2 = |\gamma_P^*|$ and $\sum_{i:\, u_i \leq 0} u_i^2 = |\gamma_N^*|$ to hold. Hence, $\Theta^*(m) = \emptyset$ iff $m < M^*(m)$. As $M^*(m) \in \{0, 1, 2\}$, $m < M^*(m)$ happens iff $m = 1$ and $M^(1) = 2$. As $1 \leq \underline{M} \leq \overline{M}$ and by (18), $M^(1) = 2$ holds iff $\underline{M} = \overline{M} = 1 = m$.
(3) $\underline{M} \leq m = 1 < \overline{M}$ or $\underline{M} \leq \overline{M} \leq m = 2$:
As $M^*(m) \in \{0, 1, 2\}$, $m = M^*(m)$ iff $1 = m = M^*(1)$ or $2 = m = M^*(2)$. By (17), $1 = m = M^*(1)$ iff $\underline{M} \leq 1 = m < \overline{M}$. By (18), $2 = m = M^*(2)$ iff $\overline{M} \leq m = 2$.

(4) Otherwise:
From the above arguments, $m < M^*(m)$ iff conditions for (2) hold, $m = M^*(m)$ iff conditions for (3) hold, and $M^*(m) = 0$ iff conditions for (1) hold. Thus, $m > M^*(m) > 0$ for this case. By the result from Theorem 17 (4), $\Theta^*(m)$ is connected. ■

B.4. Proofs for Subsection 2.2.2

Notation: For a subset $A \subset \mathbb{R}^d$, we define $\dim(A)$ to be the maximum k such that A contains a k-dimensional embedded C^1 submanifold (equivalently, the maximal stratum dimension).

Lemma 19 *Under Assumption $\delta > a$, for sufficiently large m,*

$$\mathcal{S}_{\alpha\beta}(X_S^\top Y) \neq 0 \wedge \mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y) \neq 0$$

Proof Denote $\{\min\{|X_S^T Y|,\ |X_{S^C}^T Y|\}$ by C. By our assumption, C is positive and independent of m. For $m > 1$,

$$\begin{aligned} & \mathcal{S}_{\alpha\beta}(X_S^\top Y) \neq 0 \wedge \mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y) \neq 0 \\ \iff\ & \beta < \frac{1}{\alpha}\min\{|X_S^T Y|,\ |X_{S^C}^T Y|\} \iff m^{-\delta} < m^{-(a)}C \\ \iff\ & -d < -a + \log_m C \iff a - d < \log_m C. \end{aligned}$$

$\log_m C \to 0$ as $m \to \infty$. Under Assumption $\delta > a$, $a - d < \log_m C$ holds for sufficiently leargе m. ■

Proposition 20 *(Proposition 8 in main) Under Assumption $\delta > a$, for sufficiently large m,*

$$\dim(\Theta^*(m)) = m - 2.$$

Proof By the explicit form of $\Theta^*(m)$ as in Theorem 15, ω_i is uniquely determined by u_i for every i; hence the degrees of freedom of $\Theta^*(m)$ are entirely captured by the vector $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$. We have two conditions $\sum_{i:\, u_i \geq 0} u_i^2 = |\gamma_P^*|$ and $\sum_{i:\, u_i \leq 0} u_i^2 = |\gamma_N^*|$. Since γ_P^* (resp. γ_N^*) is a nonzero scalar multiple of $\mathcal{S}_{\alpha\beta}(X_S^\top Y)$ (resp. $\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)$), we have

$$\gamma_P^* \neq 0 \wedge \gamma_N^* \neq 0 \iff \mathcal{S}_{\alpha\beta}(X_S^\top Y) \neq 0 \wedge \mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y) \neq 0.$$

By Lemma 19, this holds for sufficiently large m. Therefore, both of the quadratic equalities impose nontrivial conditions on u.

To satisfy these equalities, we need both $\{i : u_i > 0\}$ and $\{i : u_i < 0\}$ to be non-empty. The two constraints are independent because they involve disjoint sets of indices ($\{i : u_i \geq 0\} \neq \{i : u_i \leq 0\}$). Hence, in $\mathbb{R}^m$, the dimension is reduced by exactly one for each condition. ■

Proposition 21 *(Proposition 9 in main) Under Assumption $\delta > a$, for sufficiently large m, $\Theta^*(m)$ is bounded and*

$$\begin{aligned} \forall\, \theta^* \in \Theta^*(m), \quad \|\theta^*\|_2 &= \sqrt{2\alpha\left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2}\right)} \\ &= \Theta\big(m^{\frac{a}{2}}\big). \end{aligned}$$

Proof By the explicit form of $\Theta^*(m)$ in Theorem 3, $\forall\, \theta^* = (u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$,

$$\begin{aligned}\|\theta^*\|_2^2 &= 2\left(\sum_{i:\, u_i \geq 0} u_i^2 + \sum_{i:\, u_i \leq 0} u_i^2\right)\\ &= 2(|\gamma_P^*| + |\gamma_N^*|)\\ &= 2\alpha\left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2}\right)\end{aligned}$$

$\|\theta^*\|_2^2 = \Theta(m^a) \Rightarrow \|\theta^*\|_2 = \Theta(m^{\frac{a}{2}})$. ■

For the rest of the proofs, keep in mind that we made an assumption that $0 < \min\{|X_S^T Y|,\ |X_{SC}^T Y|\}$.

Lemma 22 *Under Assumption $\delta > a$, for sufficiently large m,*

$$\forall \phi^* \in \varphi^*(m),\ \forall \theta^* \in \Theta^*(m), \quad \|\phi^*\|_2 \geq \|\theta^*\|_2.$$

Proof $\forall\, \phi^* \in \varphi^*(m), \forall\, \theta^* \in \Theta^*(m)$,

$$\begin{aligned}\|\phi^*\|_2^2 &= \sum_{i=1}^m u_i^2 + \sum_{i=1}^m \omega_i^2\\ &\geq 2\sum_{i=1}^m |u_i\omega_i| \quad \text{(by AM-GM inequality)}\\ &= 2\sum_{i:u_i>0} |u_i\omega_i| + 2\sum_{i:u_i<0} |u_i\omega_i|\\ &\geq 2\left|\sum_{i:u_i>0} u_i\omega_i\right| + 2\left|\sum_{i:u_i<0} u_i\omega_i\right|\\ &= 2\left|\frac{\alpha X_S^\top Y}{\|X_S\|_2^2}\right| + 2\left|\frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}\right| \quad \text{(by Theorem 13)}\\ &= 2\alpha\left(\frac{|X_S^\top Y|}{\|X_S\|_2^2} + \frac{|X_{S^c}^\top Y|}{\|X_S^c\|_2^2}\right)\\ &\geq 2\alpha\left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2}\right)\\ &= \|\theta^*\|_2^2 \quad \text{(by Theorem 21)}\end{aligned}$$

■

Proposition 23 *(Proposition 6 in main) For sufficiently large m,*

$$\dim(\varphi^*(m)) = 2m - 2.$$

Proof For notations, write

$$c_1 := \frac{\alpha X_S^\top Y}{\|X_S\|_2^2} \quad \text{and} \quad c_2 := \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}.$$

We assume $m \geq 2$. As $0 < \min\{|X_S^\top Y|,\ |X_{S^c}^\top Y|\}$, we have $c_1 \neq 0$ and $c_2 \neq 0$. Hence any $(u_j, \omega_j)_{j=1}^m \in \varphi^*(m)$ must have at least one strictly positive u_j and at least one strictly negative u_j.

Fix a sign pattern by choosing a partition $P, N \subset \{1, \ldots, m\}$ with

$$P \neq \varnothing, \quad N \neq \varnothing, \quad P \cup N = \{1, \ldots, m\}, \quad P \cap N = \varnothing,$$

and consider the open orthant

$$\mathcal{U}_{P,N} := \Big\{(u, \omega) \in \mathbb{R}^{2m} :\ u_j > 0\ (j \in P),\ u_j < 0\ (j \in N)\Big\}.$$

On $\mathcal{U}_{P,N}$ we have $\{j : u_j \geq 0\} = P$ and $\{j : u_j \leq 0\} = N$, so the defining conditions of $\varphi^*(m)$ become the two smooth equations

$$f_1(u, \omega) := \sum_{j \in P} u_j \omega_j - c_1 = 0,$$
$$f_2(u, \omega) := \sum_{j \in N} u_j \omega_j - c_2 = 0.$$

Let $F := (f_1, f_2) : \mathbb{R}^{2m} \to \mathbb{R}^2$. Then

$$\varphi^*(m) \cap \mathcal{U}_{P,N} = \big\{(u, \omega) \in \mathcal{U}_{P,N} :\ F(u, \omega) = 0\big\} = F^{-1}(0) \cap \mathcal{U}_{P,N}.$$

Take any $(u, \omega) \in \mathcal{U}_{P,N}$. Choose $j \in P$ and $k \in N$ (possible since both are nonempty). Then

$$\frac{\partial f_1}{\partial \omega_j}(u, \omega) = u_j \neq 0, \qquad \frac{\partial f_2}{\partial \omega_j}(u, \omega) = 0,$$

and

$$\frac{\partial f_2}{\partial \omega_k}(u, \omega) = u_k \neq 0, \qquad \frac{\partial f_1}{\partial \omega_k}(u, \omega) = 0.$$

Hence the 2×2 submatrix of $DF(u, \omega)$ formed by the columns corresponding to ω_j and ω_k is

$$\begin{pmatrix} u_j & 0 \\ 0 & u_k \end{pmatrix},$$

which is invertible. Therefore $\operatorname{rank} DF(u, \omega) = 2$ at every $(u, \omega) \in \mathcal{U}_{P,N}$. In particular, $0 \in \mathbb{R}^2$ is a regular value of F on $\mathcal{U}_{P,N}$. By the preimage theorem (a version of the implicit function theorem), $F^{-1}(0) \cap \mathcal{U}_{P,N}$ is a submanifold of $\mathcal{U}_{P,N}$ with codimension 2. Hence,

$$\dim\big(F^{-1}(0) \cap \mathcal{U}_{P,N}\big) = 2m - 2.$$

The set $\varphi^*(m)$ is the union over all such admissible sign patterns (P, N) of the pieces $\varphi^*(m) \cap \mathcal{U}_{P,N}$, together with boundary parts where some $u_j = 0$. Each interior piece has dimension $2m - 2$, while boundary parts (where at least one additional equality $u_j = 0$ holds) have dimension at most $2m - 3$. Therefore, the (maximal) manifold dimension of $\varphi^*(m)$ is

$$\dim(\varphi^*(m)) = 2m - 2 \quad \text{for } m \geq 2.$$

■

Proposition 24 *(Proposition 7 in main) For sufficiently large m, $\varphi^*(m)$ is unbounded. Especially, $\forall a \geq \frac{a}{2}$, $\exists \varphi_a^*(m) \subset \varphi^*(m)$ s.t. $\forall \phi_a^* \in \varphi_a^*(m)$, $\|\phi_a^*\|_2 = \Theta(m^a)$ and $\dim(\varphi_a^*(m)) = 2m-2$.*

Proof For simplicity, define the nonzero constants $b_1 := \frac{X_S^\top Y}{\|X_S\|_2^2} \neq 0$, $b_2 := \frac{X_{S^c}^\top Y}{\|X_{S^c}\|_2^2} \neq 0$ and write $c_1(m) := \alpha b_1$, $c_2(m) := \alpha b_2$ where $\alpha = m^a$. Assume $m > 2$. Then,$\forall t > 0$, $\phi(t) = (u_i(t), \omega_i(t))_{i=1}^m \in \varphi^*(m)$ where $\phi(t)$ is defined as

$$u_1 = t,\ \omega_1 = \frac{c_1(m)}{t},\ u_2 = -t,\ \omega_2 = -\frac{c_2(m)}{t},\ u_i = 0,\ \omega_i = 0 \quad (i \in \{3,\ \dots, m\}).$$

Its squared norm is $\|\phi(t)\|_2^2 = 2t^2 + \frac{c_1(m)^2 + c_2(m)^2}{t^2} \to \infty$ as $t \to \infty$. Hence, $\varphi^*(m)$ is unbounded for $m > 2$.

As in the proof for Proposition 23, define a sign pattern partition $P, N \subset \{1, \dots, m\}$ with

$$P \neq \varnothing, \quad N \neq \varnothing, \quad P \cup N = \{1, \dots, m\}, \quad P \cap N = \varnothing,$$

and consider the open orthant

$$\mathcal{U}_{P,N} := \Big\{(u, \omega) \in \mathbb{R}^{2m} :\ u_j > 0\ (j \in P),\ u_j < 0\ (j \in N)\Big\}.$$

We set $P = \{1\}$ and $N = \{2,\ \dots, m\}$ and denote $\mathcal{U}_{P,N}$ by $\mathcal{U}$. From the proof for Proposition 23, $\dim\big(\varphi^*(m) \cap \mathcal{U}\big) = 2m-2$.

Pick any $a \geq \frac{a}{2}$. Introduce notations $\beta := \min\{|b_1|, |b_2|\} > 0$ and $\kappa := \sqrt{\frac{\beta}{2}}$.
Define $\varphi_a^*(m)$ to be the subset of $\varphi^*(m) \cap \mathcal{U}$ consisting of points satisfying

$$u_1 \in (m^a, 2m^a), \qquad u_2 \in (-2m^a, -m^a),$$

and for $j = 3, \dots, m$,

$$u_j \in \Big(-\kappa\, m^{\frac{a-1}{2}},\ -\frac{\kappa}{2}\, m^{\frac{a-1}{2}}\Big), \qquad \omega_j \in \Big(-\kappa\, m^{\frac{a-1}{2}},\ \kappa\, m^{\frac{a-1}{2}}\Big),$$

with (ω_1, ω_2) determined by the constraints:

$$\omega_1 = \frac{c_1(m)}{u_1}, \qquad \omega_2 = \frac{c_2(m) - \sum_{j=3}^m u_j \omega_j}{u_2}.$$

From the proof for Proposition 23, $\varphi^*(m) \cap \mathcal{U}$ is an embedded submanifold of $\mathbb{R}^{2m}$ (by the implicit function theorem), so the manifold topology on $\varphi^*(m) \cap \mathcal{U}$ is the subspace topology inherited from $\mathbb{R}^{2m}$. The set of points satisfying the constraints is open in $\mathbb{R}^{2m}$. Hence, $\varphi_a^*(m)$ is an open subset of the $(2m-2)$-dimensional manifold $\varphi^*(m) \cap \mathcal{U}$. Thus, $\dim(\varphi_a^*(m)) = 2m-2$.

Take any $\phi \in \varphi_a^*(m)$. By construction, $|u_1| + |u_2| = \Theta(m^a) \ \Rightarrow\ \|\phi\|_2 \geq \sqrt{u_1^2 + u_2^2} = \Omega(m^a)$. Also, $|\omega_1| = \left|\frac{c_1(m)}{u_1}\right| = \Theta\big(\frac{m^a}{m^a}\big) = \Theta\big(m^{a-a}\big)$. For $j \geq 3$ we have $|u_j \omega_j| \leq \kappa^2 m^{a-1}$, hence

$$\left|\sum_{j=3}^m u_j \omega_j\right| \leq (m-2)\kappa^2 m^{a-1} \leq \kappa^2 m^a.$$

Therefore, for all sufficiently large m,

$$\left|c_2(m) - \sum_{j=3}^{m} u_j\omega_j\right| \geq |c_2(m)| - \left|\sum_{j=3}^{m} u_j\omega_j\right| \geq |c_2(m)| - \kappa^2 m^a \geq \left(|b_2| - \kappa^2\right) m^a \geq \frac{\beta}{2} m^a.$$

Also,

$$\left|c_2(m) - \sum_{j=3}^{m} u_j\omega_j\right| \leq |c_2(m)| + \left|\sum_{j=3}^{m} u_j\omega_j\right| \leq |c_2(m)| + \kappa^2 m^a \leq \left(|b_2| + \kappa^2\right) m^a \leq \frac{3|b_2|}{2} m^a.$$

Hence, $\left|c_2(m) - \sum_{j=3}^{m} u_j\omega_j\right| = \Theta(m^a)$. By using $|u_2| = \Theta(m^a)$, we get $|\omega_2| = \Theta\left(\frac{m^a}{m^a}\right) = \Theta\left(m^{a-a}\right)$.

Moreover, by construction, $|u_j| = |\omega_j| = \Theta\left(m^{\frac{a-1}{2}}\right) \quad (j \geq 3)$. Hence,

$$\sum_{j=3}^{m}(u_j^2 + \omega_j^2) = O\left(2(m-2)m^{a-1}\right) = O(m^a)\,.$$

Their total contribution satisfies $\left(\sum_{j=3}^{m}(u_j^2 + \omega_j^2)\right)^{1/2} = O\left(m^{\frac{a}{2}}\right)$. Since $a \geq \frac{a}{2}$, we have $m^{\frac{a}{2}} \leq m^a$, so $\left(\sum_{j=3}^{m}(u_j^2 + \omega_j^2)\right)^{1/2} = O(m^a)$. Also, $a \geq \frac{a}{2}$ implies $a - a \leq a$, so $|\omega_1|, |\omega_2| = O(m^a)$. By construction, $|u_1|, |u_2| = O(m^a)$. Combining these bounds yields

$$\|\phi\|_2 = \left(\sum_{j=1}^{m}(u_j^2 + \omega_j^2)\right)^{1/2} \leq \sqrt{u_1^2 + u_2^2} + \sqrt{\omega_1^2 + \omega_2^2} + \left(\sum_{j=3}^{m}(u_j^2 + \omega_j^2)\right)^{1/2} = O(m^a).$$

Together with the lower bound $\|\phi\|_2 = \Omega(m^a)$, we conclude $\forall \phi \in \varphi_a^*(m),\ \|\phi\|_2 = \Theta(m^a)$. This completes the proof. ■

Appendix C. More analysis on connectivity result

C.1. Roles of Permutation Symmetry and Inactive Neurons for Connectivity

Both of the two proofs in Appendix B.3 for Theorem 5 exploit two basic facts about the neural network model: hidden neurons have a permutation symmetry, and overparameterized models have unnecessary hidden neurons to express an optimal function. In this section, we explain their roles for the connectivity result.

The neurons in the neural network model (1) have a permutation symmetry i.e. changing $(u_i, v_i)_{i=1}^m$ to $(u_{\pi(i)}, v_{\pi(i)})_{i=1}^m$ for a permutation π does not change its output or the training loss (3). The ℓ_2 regularization forces any neuron (u_i, ω_i) in optimal parameter $(u_i, v_i)_{i=1}^m \in \Theta^*(m)$ to be an inactive neuron i.e. $(u_i, v_i) = (0, 0)$ or an active neuron $u_i \neq 0$ and $\omega_i \neq 0$. (See Proposition 30 in Appendix.) We define, by applying the concept of minimal optimal networks defined in Kim et al. [18] to one-dimensional data, the set of Minimal Optimal Solutions as

$$\Theta^*_{\min}(m) := \left\{(u_j, \omega_j)_{j=1}^m \;\middle|\; \forall\, p \neq q \in [m], \omega_p\omega_q > 0 \Rightarrow u_p u_q < 0\right\}. \tag{19}$$

This is the set of parameters that has the least possible number of active neurons. We find that all elements in $\Theta^*_{min}(m)$ are permutations of each other. (See Appendix Lemma 35). In Figure 1 (c) (d), Minimal Optimal Solutions are denoted by black points. A key observation from the connectivity result is that once we have an inactive neuron in $(u_i, \omega_i)_{i=1}^m \in \Theta^*_{min}(m)$, $\Theta^*(m)$ becomes a connected set. This is because an inactive neuron plays a role in proving that permutations $\{(u_{\pi(i)}, v_{\pi(i)})_{i=1}^m \,\big|\, \pi \text{ is a permutation on } \{1, \ldots .\text{m}\}\}$ are connected in $\Theta^*(m)$. (See Collorary 39 in Appendix.) Additionally, a merging process, a process that is analogous to the merging process defined in Kim et al. [18], proves that all the elements in $\Theta^*(m)$ are connected to a point in $\Theta^*_{min}(m)$. (See Lemma 36). Paths created in the merging process are illustrated by colored paths in Figure 1 (c) (d).

C.2. Connectivity and Convex Formulation

In this section, we provide an explanation of why the connectivity phase transition is restricted for the one-dimensional input case (Theorem 5) from a perspective on a convex formulation introduced by Pilanci and Ergen [30]. Kim et al. [18] explored the connectivity of global optimal parameters for general d-dimensional input data by applying the convex formulation. We follow their principle strategies, but our proof is simpler because the analysis on one-dimensional input is enough for our purpose. (See Appendix D.1.)

Firstly, we find the convex formulation of the non-convex problem (3). The convex formulation of the training problem is introduced by [30]. By restricting our attention to one-dimensional input, the convex formulation can be written as follows. By abuse of notation, we denote (v_1, v_2, t_1, t_2) by $(v_i, t_i)_{i=1}^2$.

Proposition 25 *Consider the convex problem given as a cone-constrained group LASSO*

$$\min_{\substack{v_1,v_2,t_1,t_2\in\mathbb{R}\\ v_1,t_1\geq 0\\ v_2,t_2\leq 0}} L_{conv}(v_1, v_2, t_1, t_2), \tag{20}$$

$$\begin{aligned} \text{where } L_{conv}(v_1, v_2, t_1, t_2) &= \frac{1}{2}\left\|\Big((v_1 - t_1)D(S) + (v_2 - t_2)D(S^c)\Big)\frac{X}{\alpha} - Y\right\|_2^2 \\ &+ \beta\big(|v_1| + |v_2| + |t_1| + |t_2|\big). \end{aligned} \tag{21}$$

The convex problem (20) and the non-convex problem (3) have identical optimal value when $m \geq M^* = \sum_{i\in\{1,2\}:v_i^*\neq 0} 1 + \sum_{i\in\{1,2\}:t_i^*\neq 0} 1$ where $(v_i^*, t_i^*)_{i=1}^2$ is an optimal solution to (20).

Remark 26 *M^* defined in Theorem 5 and M^* defined in Proposition 25 are the same. This is because γ_P^* and γ_N^* introduced in Theorem 3 can be written as $\gamma_P^* = v_1^* - t_1^*$, $\gamma_N^* = v_2^* - t_2^*$ by Proposition 27. We call M^* the critical value.*

Since the problem is convex, we can solve (20) directly to find optimal solutions that satisfy the constraints. It turns out that the solution set is a singleton.

Proposition 27 *The set of global optimal solutions*

$$\mathcal{P}^* = \left\{ (v_1, v_2, t_1, t_2) \;\middle|\; \underset{\substack{v_1,v_2,t_1,t_2\in\mathbb{R}\\ v_1,t_1\geq 0\\ v_2,t_2\leq 0}}{\arg\min} L_{\text{conv}}(v_1, v_2, t_1, t_2) \right\}$$

for the convex problem (20) is $\mathcal{P}^* = \{(v_1^*, v_2^*, t_1^*, t_2^*)\}$ *where*

$$(v_1^*, t_1^*) = \begin{cases} \left(\alpha \frac{X_S^\top Y - \alpha\beta}{\|X_S\|_2^2}, 0\right) & \text{if } X_S^\top Y > \alpha\beta, \\ (0,0) & \text{if } -\alpha\beta \le X_S^\top Y \le \alpha\beta, \\ \left(0, -\alpha \frac{X_S^\top Y + \alpha\beta}{\|X_S\|_2^2}\right) & \text{if } X_S^\top Y < -\alpha\beta, \end{cases}$$

$$(v_2^*, t_2^*) = \begin{cases} \left(0, \alpha \frac{-X_{S^c}^\top Y + \alpha\beta}{\|X_{S^c}\|_2^2}\right) & \text{if } X_{S^c}^\top Y > \alpha\beta, \\ (0,0) & \text{if } -\alpha\beta \le X_{S^c}^\top Y \le \alpha\beta, \\ \left(\alpha \frac{X_{S^c}^\top Y + \alpha\beta}{\|X_{S^c}\|_2^2}, 0\right) & \text{if } X_{S^c}^\top Y < -\alpha\beta. \end{cases}$$

The staircase connectivity shown in [18] arises because we can relate the connectivity results of the cardinarity constraint set $\mathcal{P}^*(m) \subseteq \mathcal{P}^*$ defined as

$$\mathcal{P}^*(m) := \Big\{ (u_i, v_i)_{i=1}^P \;\Big|\; (u_i, v_i)_{i=1}^P \in \mathcal{P}^*, \sum_{i \in [P]: v_i^* \neq 0} 1 + \sum_{i \in [P]: t_i^* \neq 0} 1 \le m \Big\}. \tag{22}$$

with the connectivity results of $\Theta^*(m)$, and $\mathcal{P}^*(m)$ changes with respect to m. For 1-dimensional case, $\mathcal{P}^*$ is a singleton, so $\mathcal{P}^*(m) = \mathcal{P}^*$ for any $m \ge M^*$ and we observe limited connectivity phase transitions.

Appendix D. Proofs for Appendix

D.1. Proofs for Appendix C.2

In general, convex problems are easier to deal with than non-convex problems. Kim et al. [18] showed that the staircase connectivity of global optimal solutions to the original non-convex training loss minimization problem can be found by analyzing the connectivity of global optimal solutions to its convex formulation. We apply this strategy to our problem with 1-dimensional input data. Thanks to the simplicity coming from 1-dimensional input, we provide a simpler proof without using all the notations introduced by Kim et al. [18].

A benefit from deriving connectivity via convex formulation is that we can derive the connectivity without knowing the explicit form of $\Theta^*(m)$. Therefore, we pretend as if we did not know the explicit form (8) throughout Section D.1.

Firstly, we find the convex formulation of our non-convex problem (3).

Proposition 28 *(Proposition 25 in Appendix C.2) Consider the convex problem given as a cone-constrained group LASSO*

$$\min_{v_1, v_2, t_1, t_2} \frac{1}{2} \left\| \Big((v_1 - t_1) D(S) + (v_2 - t_2) D(S^c) \Big) \frac{X}{\alpha} - Y \right\|_2^2 + \beta \big(|v_1| + |v_2| + |t_1| + |t_2| \big)$$
$$\text{s.t.} \quad v_1 \ge 0,\ t_1 \ge 0,\ v_2 \le 0,\ t_2 \le 0 \tag{23}$$

where $D(S) = \mathrm{Diag}(\mathbf{1}[X \geq 0])$ *and* $D(S^c) = \mathrm{Diag}(\mathbf{1}[X \leq 0])$.
The convex problem (23) and the non-convex problem (3) have identical optimal value when $m \geq M^* = \sum_{i\in\{1,2\}:v_i^*\neq 0} 1 + \sum_{i\in\{1,2\}:t_i^*\neq 0} 1$ *where* $\{v_i^*, t_i^*\}_{i=1}^2$ *is an optimal solution to (23).*

Proof By our assumption in Appendix C.2, X has at least one positive element and one negative element. Hence,

$$\{\mathrm{Diag}(\mathbf{1}[Xu \geq 0] \mid u \in \mathbb{R}\} = \{\mathrm{Diag}(\mathbf{1}[X \geq 0]),\ \mathrm{Diag}(\mathbf{1}[X \leq 0]),\ I_n\}.$$

(3) can be equivalently written as

$$L(\theta) = \frac{1}{2}\|\sum_{j=1}^{m}(\frac{X}{\alpha}u_j)_+\omega_j - Y\|_2^2 + \frac{\beta}{2}\sum_{j=1}^{m}(u_j^2 + \omega_j^2) \tag{24}$$

Consider the following convex problem

$$\min_{\{v_i,t_i\}_{i=1}^3} \frac{1}{2}\Big\|\Big((v_1 - t_1)D(S) + (v_2 - t_2)D(S^c) + (v_3 - t_3)I_n\Big)\frac{X}{\alpha} - Y\Big\|_2^2 \quad + \beta\sum_{i=1}^{3}\big(|v_i| + |t_i|\big)$$
$$\text{s.t.} \quad (2D_i - I_n)Xv_i \geq 0, \quad (2D_i - I_n)Xt_i \geq 0,\ \forall i \in [3]. \quad (D_1 = D(S),\ D_2 = D(S^c),\ D_3 = I_n) \tag{25}$$

By applying Theorem 1 in Pilanci and Ergen [30] to our problem with $d = 1$ and input matrix $\frac{X}{\alpha}$, the convex problem (25) and the non-convex problem (3) have identical optimal values if $m \geq M^* = \sum_{i\in[3]:v_i^*\neq 0} 1 + \sum_{i\in[3]:t_i^*\neq 0} 1$ where $\{v_i^*, t_i^*\}_{i=1}^2$ is an optimal solution to (23). As X has both positive and negative elements,

$$\begin{aligned} &(2D_i - I_n)Xv_i \geq 0, \quad (2D_i - I_n)Xt_i \geq 0,\ \forall i \in [3] \\ \iff &v_1 \geq 0,\ t_1 \geq 0,\ v_2 \leq 0,\ t_2 \leq 0,\ v_3 = t_3 = 0 \end{aligned}$$

Rewriting (25) with this condition gives (23). ■

To find the connectivity of $\Theta^*(m)$, we use this convex formulation. We find the solution set $\mathcal{P}^*$ to the convex problem (23) is a singleton.

Proposition 29 *(Proposition 27 in Appendix C.2) The set of global optimal solutions*

$$\mathcal{P}^* = \left\{ (v_1, v_2, t_1, t_2) \;\middle|\; \begin{array}{c} \arg\min \\ {\scriptstyle v_1,v_2,t_1,t_2\in\mathbb{R}} \\ {\scriptstyle v_1,t_1\geq 0} \\ {\scriptstyle v_2,t_2\leq 0} \end{array} L_{\mathrm{conv}}(v_1, v_2, t_1, t_2) \right\}$$

for the convex problem (23) is $\mathcal{P}^* = \{(v_1^*, v_2^*, t_1^*, t_2^*)\}$ *where*

$$(v_1^*, t_1^*) = \begin{cases} \left(\alpha\frac{X_S^TY - \alpha\beta}{\|X_S\|_2^2}, 0\right) & \text{if} \quad X_S^TY > \alpha\beta \\ (0,0) & \text{if} \quad -\alpha\beta \leq X_S^TY \leq \alpha\beta \\ \left(0, -\alpha\frac{X_S^TY + \alpha\beta}{\|X_S\|_2^2}\right) & \text{if} \quad X_S^TY < -\alpha\beta \end{cases}$$

$$(v_2^*, t_2^*) = \begin{cases} \left(0, \alpha\frac{-X_{S^c}^TY + \alpha\beta}{\|X_{S^c}\|_2^2}\right) & \text{if} \quad X_{S^c}^TY > \alpha\beta, \\ (0,0) & \text{if} \quad -\alpha\beta \leq X_{S^c}^TY \leq \alpha\beta, \\ \left(\alpha\frac{X_{S^c}^TY + \alpha\beta}{\|X_{S^c}\|_2^2}, 0\right) & \text{if} \quad X_{S^c}^TY < -\alpha\beta. \end{cases}$$

Proof

$$L_{conv}(v_1, v_2, t_1, t_2) = \frac{1}{2}\Big\|\Big((v_1 - t_1)D(S) + (v_2 - t_2)D(S^c)\Big)\frac{X}{\alpha} - Y\Big\|_2^2 + \beta\big(|v_1| + |v_2| + |t_1| + |t_2|\big).$$

By the KKT condition, at an optimal,

$$0 \in \partial_{v_1} L_{conv} = (v_1 - t_1)\Big\|\frac{X_S}{\alpha}\Big\|^2 - \frac{X_S^T Y}{\alpha} + \beta\partial|v_1|,$$
$$0 \in \partial_{t_1} L_{conv} = (t_1 - v_1)\Big\|\frac{X_S}{\alpha}\Big\|^2 + \frac{X_S^T Y}{\alpha} + \beta\partial|t_1|.$$

To satisfy the above conditions and the constraint $v_1, t_1 \geq 0$, optimal v_1^*, t_1^* are uniquely defined as

$$(v_1^*, t_1^*) = \begin{cases} \Big(\alpha\frac{X_S^T Y - \alpha\beta}{\|X_S\|_2^2}, 0\Big) & \text{if} \quad X_S^T Y > \alpha\beta \\ (0, 0) & \text{if} \quad -\alpha\beta \leq X_S^T Y \leq \alpha\beta \\ \Big(0, -\alpha\frac{X_S^T Y + \alpha\beta}{\|X_S\|_2^2}\Big) & \text{if} \quad X_S^T Y < -\alpha\beta \end{cases}$$

We can find optimal v_2^*, t_2^* by replacing X_S by X_{S^c} and by replacing the constraint by $v_2, t_2 \leq 0$. ■

We state a constraint about the form of solutions in $\Theta^*(m)$. This constraint comes from the ℓ_2-regularization.

Proposition 30 $\forall (u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$, $|u_i| = |\omega_i|$ *holds for all* $i \in [m]$.

Proof Take any $(u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$. We have two cases to consider.

(Case 1) $u_i = 0$ (or $\omega_i = 0$):
The choice of ω_i (or $u_i = 0$) does not change the model function $f_\theta(X)$ defined in (**??**). For an optimal parameter, we need $\omega_i = 0$ (or $u_i = 0$) to minimize the ℓ_2-regularization term. This implies that $u_i = 0 \iff \omega_i = 0$. Hence, $|u_i| = |\omega_i|$.

(Case 2) $u_i \neq 0$:
From Case 1, $u_i \neq 0 \Rightarrow \omega_i \neq 0$. $\forall r > 0$, changing (u_i, ω_i) to $(ru_i, \omega_i/r)$ does not change the model function $f_\theta(X)$. By AM-GM inequality, $r^2 u_i^2 + \frac{\omega_i^2}{r^2} \geq 2|u_i||\omega_i|$ with equality iff $r^2 u_i^2 = \frac{\omega_i^2}{r^2}$. For $r > 0$, $r^2 u_i^2 = \frac{\omega_i^2}{r^2}$ implies $r = \frac{|\omega_i|}{|u_i|}$. Hence, by minimality, $(u_i, \omega_i) = (ru_i, \omega_i/r) = (\frac{|\omega_i|}{|u_i|}u_i, \frac{|u_i|}{|\omega_i|}\omega_i)$. This implies $|u_i| = |\omega_i|$. ■

We take the following steps to prove Theorem 17.

1. Construct functions that map elements in $\mathcal{P}^*$ and elements in $\Theta^*(m)$. (Definition 31 and Definition 32)

2. Introduce the notion of Minimal Optimal Solution. (Definition 34)

3. Prove that Minimal Optimal Solutions are permutations of each other. (Lemma 35)

4. Prove that any optimal solution is connected to a Minimal Optimal Solution. (Lemma 36)
5. For $m = M^*$, prove that all the optimal solutions are Minimal Optimal Solutions. (Lemma 37)
6. For $m > M^*$, prove that all the permutation solutions are connected in $\Theta^*(m)$. (Lemma 38)

Step 3 and Step 5 together imply that $\Theta^*(M^*)$ is finite. Step 3, Step 4, and Step 6 together imply that $\Theta^*(m)$ is connected when $m > M^*$.

Definition 31 *Suppose $m \geq M^*$. Define $\Psi : \mathcal{P}^* \to \Theta^*(m)$ as*

$$\Psi((v_i^*, t_i^*)_{i=1}^2) = \left(\frac{v_i^*}{\sqrt{|v_i^*|}},\ \sqrt{|v_i^*|}\right)_{v_i^* \neq 0} \oplus \left(\frac{t_i^*}{\sqrt{|t_i^*|}},\ -\sqrt{|t_i^*|}\right)_{t_i^* \neq 0} \oplus (0,0)^{m-M^*}.$$

Definition 32 *Suppose $m \geq M^*$. Define $\Phi : \Theta^*(m) \to \mathcal{P}^*$ as*

$$\begin{aligned}
\Phi((u_i, \omega_i)_{i=1}^m) &= (v_i, t_i)_{i=1}^2 \\
&:= \begin{cases} v_1 = \sum_{i \in \mathcal{I}} u_i |\omega_i| & \textit{where} \quad \mathcal{I} = \{i \mid \omega_i > 0, u_i > 0\}, \\ v_2 = \sum_{i \in \mathcal{I}} u_i |\omega_i| & \textit{where} \quad \mathcal{I} = \{i \mid \omega_i > 0, u_i < 0\}, \\ t_1 = \sum_{i \in \mathcal{I}} u_i |\omega_i| & \textit{where} \quad \mathcal{I} = \{i \mid \omega_i < 0, u_i > 0\}, \\ t_2 = \sum_{i \in \mathcal{I}} u_i |\omega_i| & \textit{where} \quad \mathcal{I} = \{i \mid \omega_i < 0, u_i < 0\}. \end{cases} \\
&\left(\sum_{i \in \mathcal{I}} u_i |\omega_i| = 0 \quad \textit{if} \quad \mathcal{I} = \emptyset.\right)
\end{aligned}$$

For simplicity, we take $\alpha = 1$ in the following arguments. We can prove the same connectivity result for our case by replacing X by $\frac{X}{\alpha}$ ($\alpha > 0$).

Proposition 33 *Suppose $m \geq M^*$. The maps Ψ and Φ are well-defined.*

Proof The function values for Φ and Ψ are uniquely determined for each input, so it is enough to show that $\forall (u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$, $\Phi((u_i, \omega_i)_{i=1}^m) = (v_i^*, t_i^*)_{i=1}^2 \in \mathcal{P}^*$ and for $(v_i^*, t_i^*)_{i=1}^2 \in \mathcal{P}^*$, $\Psi((v_i^*, t_i^*)_{i=1}^2) \in \Theta^*(m)$.
To prove the first part, by Proposition 28, it is enough to show that $L_{conv}(\Phi((u_i, \omega_i)_{i=1}^m)) =$

$L((u_i,\omega_i)_{i=1}^m)$ holds for $(u_i,\omega_i)_{i=1}^m \in \Theta^*(m)$. We denote $\Phi((u_i,\omega_i)_{i=1}^m)$ by $(v_i^*,t_i^*)_{i=1}^2$.

$$\begin{aligned}
(v_1^*-t_1^*)D(S)X &= \left(\sum_{\substack{i:u_i>0\\ \omega_i>0}} u_i|\omega_i| - \sum_{\substack{i:u_i>0\\ \omega_i<0}} u_i|\omega_i|\right) D(S)X \\
&= \sum_{i:u_i>0} u_i\omega_i D(S)X \\
&= \sum_{i:u_i>0} (Xu_i)_+\omega_i. \\
(v_2^*-t_2^*)D(S^c)X &= \left(\sum_{\substack{i:u_i<0\\ \omega_i>0}} u_i|\omega_i| - \sum_{\substack{i:u_i<0\\ \omega_i<0}} u_i|\omega_i|\right) D(S^c)X \\
&= \sum_{i:u_i<0} u_i\omega_i D(S^c)X \\
&= \sum_{i:u_i<0} (Xu_i)_+\omega_i.
\end{aligned}$$

$\forall\mathcal{I}, \forall i,j\in\mathcal{I}$, $\text{sign}\,(u_i|\omega_i|) = \text{sign}(u_j|\omega_j|)$, so $|\sum_{i\in\mathcal{I}} u_i|\omega_i|| = \sum_{i\in\mathcal{I}}|u_i||\omega_i| = \frac{1}{2}\sum_{i\in\mathcal{I}} u_i^2+\omega_i^2$. The last equality holds by the result from Proposition 30.
Hence, $L_{conv}(\Phi((u_i,\omega_i)_{i=1}^m)) = \frac{1}{2}\|\sum_{j=1}^m (Xu_j)_+\omega_j - Y\|_2^2 + \beta(|v_1^*|+|v_2^*|+|t_1^*|+|t_2^*|) = L((u_i,\omega_i)_{i=1}^m)$.

To prove the second part, by Proposition 28, it is enough to show that $L_{conv}((v_i^*,t_i^*)_{i=1}^2) = L(\Psi((v_i^*,t_i^*)_{i=1}^2))$ holds for $(v_i^*,t_i^*)_{i=1}^2\in\mathcal{P}^*$. We denote $\Psi((v_i^*,t_i^*)_{i=1}^2)$ by $(u_i,\omega_i)_{i=1}^m$.

$$\begin{aligned}
L(\Psi((v_i^*,t_i^*)_{i=1}^2)) =& \frac{1}{2}\|\sum_{i\in[2],v_i^*\neq 0}\left(X\frac{v_i^*}{\sqrt{|v_i^*|}}\right)_+\sqrt{|v_i^*|} - \sum_{i\in[2],t_i^*\neq 0}\left(X\frac{t_i^*}{\sqrt{|t_i^*|}}\right)_+\sqrt{|t_i^*|} - Y\|_2^2 \\
&+\frac{\beta}{2}\sum_{i\in[2],v_i^*\neq 0}\left(\left(\frac{v_i^*}{\sqrt{|v_i^*|}}\right)^2+\left(\sqrt{|v_i^*|}\right)^2\right)+\frac{\beta}{2}\sum_{i\in[2],t_i^*\neq 0}\left(\left(\frac{t_i^*}{\sqrt{|t_i^*|}}\right)^2+\left(\sqrt{|t_i^*|}\right)^2\right) \\
=& \frac{1}{2}\|\sum_{i\in[2],v_i^*\neq 0}(Xv_i^*)_+ - \sum_{i\in[2],t_i^*\neq 0}(Xt_i^*)_+ - Y\|_2^2 + \beta\sum_{i\in[2],v_i^*\neq 0}|v_i^*| + \beta\sum_{i\in[2],t_i^*\neq 0}|t_i^*| \\
=& \frac{1}{2}\|\sum_{i\in[2],v_i^*\neq 0}(Xv_i^*)_+ - \sum_{i\in[2],t_i^*\neq 0}(Xt_i^*)_+ - Y\|_2^2 + \beta\sum_{i\in[2]}(|v_i^*|+|t_i^*|)
\end{aligned}$$

As $v_1^*,t_1^*\geq 0$ and $v_2^*,t_2^*\leq 0$, $\sum_{i\in[2],v_i^*\neq 0}(Xv_i^*)_+ = D(S)Xv_1^* + D(S^c)Xv_2^*$ and $\sum_{i\in[2],t_i^*\neq 0}(Xt_i^*)_+ = D(S)Xt_1^* + D(S^c)Xt_2^*$. Hence, $L(\Psi((v_i^*,t_i^*)_{i=1}^2)) = L_{conv}((v_i^*,t_i^*)_{i=1}^2)$. ■

Definition 34 *The set of Minimal Optimal Solutions is defined for $m\geq M^*$ as*

$$\Theta_{min}^*(m) := \{(u_j,\omega_j)_{j=1}^m | \forall p\neq q\in[m],\ \omega_p\omega_q>0 \Rightarrow u_pu_q<0\}$$

A Minimal Optimal Solution has the least possible number of active neurons. The Lemma 35 proves that all the Minimal Optimal Solutions are permutations of each other.

Lemma 35 *Denote the set of all the permutations on $\{1, \ldots, m\}$ by S_m. $\forall \pi \in S_m$, we call $(u_{\pi(i)}, \omega_{\pi(i)})_{i=1}^m$ a permutation of $(u_i, \omega_i)_{i=1}^m$. Then, every element in $\Theta^*_{min}(m)$ is a permutation of $\Psi((v_i^*, t_i^*)_{i=1}^2)$.*

Proof $\Psi((v_i^*, t_i^*)_{i=1}^2) \in \Theta^*_{min}(m)$ by its construction as Definition 31, $v_1^*, t_1^* \geq 0$ and $v_2^*, t_2^* \leq 0$. Hence, it is enough to show that $\forall (u_i, \omega_i)_{i=1}^m \in \Theta^*_{min}(m)$, $(u_i, \omega_i)_{i=1}^m$ is a permutation of $\Psi((v_i^*, t_i^*)_{i=1}^2)$.
For any element in $\Theta^*_{min}(m)$, the minimality shown in Definition 34 implies that $\mathcal{I}$ introduced in Definition 32 is a singleton or an empty set. Take any $(u_i, \omega_i)_{i=1}^m \in \Theta^*_{min}(m)$. As $\mathcal{I}$ is a singleton,

$$\begin{aligned}\Phi((u_i, \omega_i)_{i=1}^m) &= (v_i^*, t_i^*)_{i=1}^2 \\ &= \begin{cases} v_1^* = u_i|\omega_i| & \text{if } \exists!\, i \; s.t. \; \omega_i > 0 \wedge u_i > 0, \; 0 \; otherwise, \\ t_1^* = u_i|\omega_i| & \text{if } \exists!\, i \; s.t. \; \omega_i < 0 \wedge u_i > 0, \; 0 \; otherwise, \\ v_2^* = u_i|\omega_i| & \text{if } \exists!\, i \; s.t. \; \omega_i > 0 \wedge u_i < 0, \; 0 \; otherwise, \\ t_2^* = u_i|\omega_i| & \text{if } \exists!\, i \; s.t. \; \omega_i < 0 \wedge u_i < 0, \; 0 \; otherwise. \end{cases}\end{aligned}$$

Hence, $M^* = \sum_{i \in \{1,2\}: v_i^* \neq 0} 1 + \sum_{i \in \{1,2\}: t_i^* \neq 0} 1 = \sum_{i=1}^m 1(u_i \omega_i \neq 0)$.
We write $P = \sum_{i=1}^m 1(\omega_i > 0)$ and $N = \sum_{i=1}^m 1(\omega_i < 0)$. Using the result from Proposition 30, $P + Q = \sum_{i=1}^m 1(u_i \omega_i \neq 0)$, so $M^* = P + Q$. Denote the P positive elements $\{\omega_i | \omega_i > 0\}$ by $\{\omega_{j_1}, \ldots, \omega_{j_P}\}$. Denote the N negative elements $\{\omega_i | \omega_i < 0\}$ by $\{\omega_{k_1}, \ldots, \omega_{k_N}\}$.
Then,

$$\Psi((v_i^*, t_i^*)_{i=1}^2) = \Psi(\Phi((u_i, \omega_i)_{i=1}^m)) = \left(\frac{u_{j_i}|\omega_{j_i}|}{\sqrt{|u_{j_i}\omega_{j_i}|}}, \sqrt{|u_{j_i}\omega_{j_i}|}\right)_{i=1}^P \oplus \left(\frac{u_{k_i}|\omega_{k_i}|}{\sqrt{|u_{k_i}\omega_{k_i}|}}, -\sqrt{|u_{k_i}\omega_{k_i}|}\right)_{i=1}^Q \oplus (0,0)^{m-M^*}.$$

From Proposition 30, $\forall j, \; \sqrt{|u_j \omega_j|} = |\omega_j|$. Therefore,

$$\begin{aligned}\Psi((v_i^*, t_i^*)_{i=1}^2) = \Psi(\Phi((u_i, \omega_i)_{i=1}^m)) &= (u_{j_i}, \; |\omega_{j_i}|)_{i=1}^P \oplus (u_{k_i}, \; -|\omega_{k_i}|)_{i=1}^N \oplus (0,0)^{m-M^*} \\ &= (u_{j_i}, \; \omega_{j_i})_{i=1}^P \oplus (u_{k_i}, \; \omega_{k_i})_{i=1}^N \oplus (0,0)^{m-M^*}\end{aligned}$$

This is a permutation of $(u_i, \omega_i)_{i=1}^m$. ■

Lemma 36 *For any $(u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$, there exists a point in $\Theta^*_{min}(m)$ that is connected to $(u_i, \omega_i)_{i=1}^m$.*

Proof It is enough to show that $\forall (u_i, \omega_i)_{i=1}^m \in \Theta^*(m) \setminus \Theta^*_{min}(m)$, we can construct a continuous path that connects $(u_i, \omega_i)_{i=1}^m$ and a point in $\Theta^*_{min}(m)$.
As $(u_i, \omega_i)_{i=1}^m \notin \Theta^*_{min}(m)$, wlog, $\omega_1 \omega_2 > 0$ and $u_1 u_2 > 0$. Denote $\text{sign}(\omega_1) = \text{sign}(\omega_2)$ by s. Define a path

$$C(t) = \left(\frac{u_1|\omega_1| + t u_2|\omega_2|}{\sqrt{|u_1\omega_1 + t u_2 \omega_2|}}, \sqrt{|u_1\omega_1 + t u_2\omega_2|}s\right) \oplus \left(\sqrt{1-t}\frac{u_2|\omega_2|}{\sqrt{|u_2\omega_2|}}, \sqrt{(1-t)|u_2\omega_2|}s\right) \oplus (u_j, \omega_j)_{j=3}^m$$

where $t \in [0, 1]$.
$C(t)$ is a part of the path connecting $(u_i, \omega_i)_{i=1}^m$ and a point in $\Theta^*_{min}(m)$. To show this, we prove the following claims.

1. $C(t)$ is well-defined and is continuous.

   Proof $\omega_1\omega_2 > 0$ and $u_1u_2 > 0$ imply that $\text{sign}(u_1\omega_1) = \text{sign}(u_2\omega_2) \neq 0$. Hence, $|u_1\omega_1 + tu_2\omega_2| \neq 0$ for $t \in [0, 1]$. Also, $\omega_2 \neq 0$ implies that $|u_2\omega_2| \neq 0$ by Proposition 30. The well-definedness of $C(t)$ implies that $C(t)$ is continuous on $[0, 1]$. ■

2. The number of zero neurons in $C(1)$ is more than that of $C(0)$.

   Proof By direct calculation, $C(0) = (u_i, \omega_i)_{i=1}^m$ and $C(1) = \left(\frac{u_1|\omega_1|+u_2|\omega_2|}{\sqrt{|u_1\omega_1+u_2\omega_2|}}, \sqrt{|u_1\omega_1 + u_2\omega_2|}s\right) \oplus (0, 0) \oplus (u_j, \omega_j)_{j=3}^m$. As $(u_1, \omega_1) \neq (0, 0)$ and $(u_2, \omega_2) \neq (0, 0)$, the claim is true. ■

3. $C(t)$ is a path in $\Theta^*(m)$.

   Proof By substituting $\theta = C(t)$ to $f_\theta(X)$ in (1), the sum of the first two terms is

$$
\begin{aligned}
&\left(X\frac{u_1|\omega_1| + tu_2|\omega_2|}{\sqrt{|u_1\omega_1 + tu_2\omega_2|}}\right)_+ \sqrt{|u_1\omega_1 + tu_2\omega_2|}s + \left(X\sqrt{1-t}\frac{u_2|\omega_2|}{\sqrt{|u_2\omega_2|}}\right)_+ \sqrt{(1-t)|u_2\omega_2|}s \\
=&\left(X(u_1|\omega_1| + tu_2|\omega_2|)\right)_+ s + (1-t)\left(Xu_2|\omega_2|\right)_+ s \\
=&\left(Xu_1|\omega_1|\right)_+ s + \left(tXu_2|\omega_2|\right)_+ s + (1-t)\left(Xu_2|\omega_2|\right)_+ s \quad (\text{since } \text{sign}(u_1|\omega_1|) = \text{sign}(u_2|\omega_2|)) \\
=&(Xu_1)_+\omega_1 + (Xu_2)_+\omega_2
\end{aligned}
$$

   Hence, changing $(u_i, \omega_i)_{i=1}^m$ to $C(t)$ does not change the model function.

$$
\begin{aligned}
&\left(\frac{u_1|\omega_1| + tu_2|\omega_2|}{\sqrt{|u_1\omega_1 + tu_2\omega_2|}}\right)^2 + \left(\sqrt{|u_1\omega_1 + tu_2\omega_2|}s\right)^2 + \left(\sqrt{1-t}\frac{u_2|\omega_2|}{\sqrt{|u_2\omega_2|}}\right)^2 + \left(\sqrt{(1-t)|u_2\omega_2|}s\right)^2 \\
=&\frac{(u_1|\omega_1| + tu_2|\omega_2|)^2}{|u_1\omega_1 + tu_2\omega_2|} + |u_1\omega_1 + tu_2\omega_2| + (1-t)\frac{(u_2|\omega_2|)^2}{|u_2\omega_2|} + (1-t)|u_2\omega_2| \\
=&\frac{(u_1|\omega_1| + tu_2|\omega_2|)^2}{|u_1\omega_1 + tu_2\omega_2|} + |u_1\omega_1 + tu_2\omega_2| + 2(1-t)|u_2\omega_2| \\
=&2|u_1\omega_1 + tu_2\omega_2| + 2(1-t)|u_2\omega_2| \quad (\text{since } \text{sign}(\omega_1) = \text{sign}(\omega_2)) \\
=&2|u_1\omega_1| + 2|u_2\omega_2| \quad (\text{since } \text{sign}(u_1\omega_1) = \text{sign}(u_1\omega_1)) \\
=&u_1^2 + \omega_1^2 + u_2^2 + \omega_2^2
\end{aligned}
$$

   Hence, changing $(u_i, \omega_i)_{i=1}^m$ to $C(t)$ does not change the ℓ_2-regularization term.
   From the above arguments, changing $(u_i, \omega_i)_{i=1}^m$ to $C(t)$ does not change the loss value i.e. $L(u_i, \omega_i)_{i=1}^m) = L(C(t))$. Hence, $\forall t \in [0, 1]$, $C(t)$ is in $\Theta^*(m)$. ■

We can see that $C(t)$ merges the two active neurons $\{(u_1, \omega_1), (u_2, \omega_2)\}$ to generate one active neuron $\left(\frac{u_1|\omega_1|+u_2|\omega_2|}{\sqrt{|u_1\omega_1+u_2\omega_2|}}, \sqrt{|u_1\omega_1 + u_2\omega_2|}s\right)$ and one inactive neuron (0,0). We repeat this merging process until we cannot find such a pair, i.e., we reach a point in $\Theta^*_{min}(m)$. This process should

terminate since the merging process strictly decreases the number of active neurons. When the merging process ends, we concatenate all the paths. Then, we have a continuous path in $\Theta^*(m)$ starting from $(u_i, \omega_i)_{i=1}^m$. At the end of the path, we have a point in $\Theta^*_{min}(m)$. ■

Lemma 37 $\Theta^*(M^*) = \Theta^*_{min}(M^*)$

Proof Assume that there exists $A \in \Theta^*(M^*) \setminus \Theta^*_{min}(M^*)$. By applying the merging process defined in Lemma 36 to A, we get a point $B \in \Theta^*_{min}(M^*)$ that has at least one inactive neuron $(0, 0)$. So, the number of non-zero neurons in B is at most $M^* - 1$. From the proof in Lemma 35, for $B = (u_i, \omega_i)_{i=1}^m \in \Theta^*_{min}(M^*)$, $M^* = \sum_{i=1}^m 1(u_i\omega_i \neq 0)$ i.e. the number of non-zero neurons is M^*. ■

Lemma 38 *For $m \geq M^* + 1$, all permutation solutions are connected.*

The idea of the proof is inspired by Kim et al. [18]. We prove it using notions of group theory, similarly to our other proofs (e.g., the proof for Theorem 16 (3), the first proof for Theorem 17).
Proof We use S_m to denote the symmetric group on $\{1, 2, \ldots, m\}$. S_m is the set of all permutations on $\{1, 2, \ldots, m\}$.
Using Lemma 35 and Lemma 36, it is enough to show that

$$\forall (u_i, \omega_i)_{i=1}^m \in \Theta^*_{min}(m),\ \forall \sigma \in S_m,\ \exists \text{a continuous path in } \Theta^*(m) \text{ connecting } (u_i, \omega_i)_{i=1}^m \text{ and } (u_{\sigma(i)}, \omega_{\sigma(i)})_{i=1}^m.$$

We take any $(u_i, \omega_i)_{i=1}^m \in \Theta^*_{min}(m)$. From the same argument in Lemma 35, $m > M^* = \sum_{i=1}^m 1(u_i\omega_i \neq 0)$ implies that $(u_i, \omega_i)_{i=1}^m$ has at least one inactive neuron (0,0). Denote the index for the inactive neuron as $j \in [m]$, so $(u_j, \omega_j) = (0, 0)$. As for all $j \in [m]$, $\mathcal{T}_j = \{(i, j) \mid i \in [m]\ s.t.\ i \neq j\}$ (We use (i, j) to denote a transposition) generates S_m, it is enough to show

$\forall T \in \mathcal{T}_j$ where $u_j \neq 0$, $\exists$ a continuous path in $\Theta^*(m)$ connecting $(u_i, \omega_i)_{i=1}^m$ and $(u_{T(i)}, \omega_{T(i)})_{i=1}^m$.

For $(u_{T(j)}, \omega_{T(j)}) = (0, 0)$, applying T does not change the parameters. So, it is enough to think about the case where $(u_{T(j)}, \omega_{T(j)}) \neq (0, 0)$ We construct a path

$$C(t) = (u_j(t), \omega_j(t))_{j=1}^m \in (\mathbb{R} \times \mathbb{R})^m,\ t \in [0, 1]$$

s.t.

$$(u_i(t), \omega_i(t)) = \begin{cases} (u_i, \omega_i) \quad \text{if} \quad i \neq j \text{ and } i \neq T(j), \\ \left(u_{T(j)}|\omega_{T(j)}|\sqrt{\frac{t}{|u_{T(j)}\omega_{T(j)}|}}, \sqrt{t|u_{T(j)}\omega_{T(j)}|}s\right) \quad \text{if} \quad i = j, \\ \left(u_{T(j)}|\omega_{T(j)}|\sqrt{\frac{1-t}{|u_{T(j)}\omega_{T(j)}|}}, \sqrt{(1-t)|u_{T(j)}\omega_{T(j)}|}s\right) \quad \text{if} \quad i = T(j) \end{cases}$$

where $s = \text{sign}(\omega_{T(j)})$. We prove the following claims.

1. $C(t)$ is well-defined and is continuous.

   Proof Proposition 30 and $(u_{T(j)}, \omega_{T(j)}) \neq (0,0)$ imply that $|u_{T(j)}\omega_{T(j)}| \neq 0$. So, division by $|u_{T(j)}\omega_{T(j)}|$ is possible and well-defined. The well-definedness of $C(t)$ implies that $C(t)$ is continuous on $[0,1]$. ■

2. $C(0) = (u_i, \omega_i)_{i=1}^m$ and $C(1) = (u_{T(i)}, \omega_{T(i)})_{i=1}^m$

   Proof By direct calculation. ■

3. $C(t)$ is a continuous path in $\Theta^*(m)$.

   Proof

$$
\begin{aligned}
&\left(Xu_{T(j)}|\omega_{T(j)}|\sqrt{\frac{t}{|u_{T(j)}\omega_{T(j)|}}}\right)_+ \sqrt{t|u_{T(j)}\omega_{T(j)}|}s \\
&+\left(Xu_{T(j)}|\omega_{T(j)}|\sqrt{\frac{1-t}{|u_{T(j)}\omega_{T(j)|}}}\right)_+ \sqrt{(1-t)|u_{T(j)}\omega_{T(j)}|}s \\
=&t\left(Xu_{T(j)}\right)_+ \omega_{T(j)} + (1-t)\left(Xu_{T(j)}\right)_+ \omega_{T(j)} \\
=&\left(Xu_{T(j)}\right)_+ \omega_{T(j)} \\
=&\left(Xu_j\right)_+ \omega_j + \left(Xu_{T(j)}\right)_+ \omega_{T(j)}
\end{aligned}
$$

   Hence, changing $(u_i, \omega_i)_{i=1}^m$ to $C(t)$ does not change the function $f(X)$.

$$
\begin{aligned}
&\left(u_{T(j)}|\omega_{T(j)}|\sqrt{\frac{t}{|u_{T(j)}\omega_{T(j)|}}}\right)^2 + \left(\sqrt{t|u_{T(j)}\omega_{T(j)}|}s\right)^2 \\
&+\left(u_{T(j)}|\omega_{T(j)}|\sqrt{\frac{1-t}{|u_{T(j)}\omega_{T(j)|}}}\right)^2 + \left(\sqrt{(1-t)|u_{T(j)}\omega_{T(j)}|}s\right)^2 \\
&= t|u_{T(j)}\omega_{T(j)}| + (1-t)|u_{T(j)}\omega_{T(j)}| \\
&= |u_{T(j)}\omega_{T(j)}| \\
&\leq u_j^2 + \omega_j^2 + u_{T(j)}^2 + \omega_{T(j)}^2
\end{aligned}
$$

   Hence, changing $(u_i, \omega_i)_{i=1}^m$ to $C(t)$ does not increase the ℓ_2 term. By the optimality of $(u_i, \omega_i)_{i=1}^m$, equality holds in the last inequality.

   From the above arguments, changing $(u_i, \omega_i)_{i=1}^m$ to $C(t)$ does not change the loss value, so $\forall t \in [0,1]$, $C(t)$ is in $\Theta^*(m)$. ■

From above arguments, $C(t)$ is a continuous path in $\Theta^*(m)$ connecting $(u_i, \omega_i)_{i=1}^m$ and $(u_{T(i)}, \omega_{T(i)})_{i=1}^m$.

■

Corollary 39

$$\forall (u_i, \omega_i)_{i=1}^m \in \Theta^*(m) \ s.t. \ \exists \ k \in [m] \ s.t \ (u_k, v_k) = (0, 0), \ \forall \sigma \in S_m,$$
$$\exists \textit{a continuous path in } \Theta^*(m) \textit{ connecting } (u_i, \omega_i)_{i=1}^m \textit{ and } (u_{\sigma(i)}, \omega_{\sigma(i)})_{i=1}^m.$$

Proof This is a corollary from the proof of Lemma 38. ■

The second proof for Theorem 17

Proof (The second proof)

$$M^* = 1[\mathcal{S}_{\alpha\beta}(X_S^T Y) \neq 0) + 1[\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y) \neq 0] = 1[|\gamma_P^*| \neq 0] + 1[|\gamma_N^*| \neq 0].$$

(1) $M^* = 0$:
$M^* = 0$, implies $|\gamma_P^*| = 0$ and $|\gamma_N^*| = 0$. For $(u_i, \omega_1)_{i=1}^m \in \Theta^*(m)$, $\sum_{i:\, u_i \geq 0} u_i^2 = 0$ and $\sum_{i:\, u_i \leq 0} u_i^2 = 0$, so $\forall i \in [m], \ u_i = 0$. Also, $\forall i \in [m], \ |u_i| = |\omega_i|$, so $\forall i \in [m], \ \omega_i = 0$.

(2) $m < M^*$:
$\Theta^*(m) = \emptyset$ because we need at least M^* non-zero neurons for $\sum_{i:\, u_i \geq 0} u_i^2 = |\gamma_P^*|$ and $\sum_{i:\, u_i \leq 0} u_i^2 = |\gamma_N^*|$ to hold.

(3) $m = M^*$:
As a set of permutations of a finite number of neurons is a finite set, Lemma 35 implies that $\Theta_{min}^*(m)$ is a finite set for all m. By this and Lemma 37, $\Theta^*(M^*)$ is a finite set.

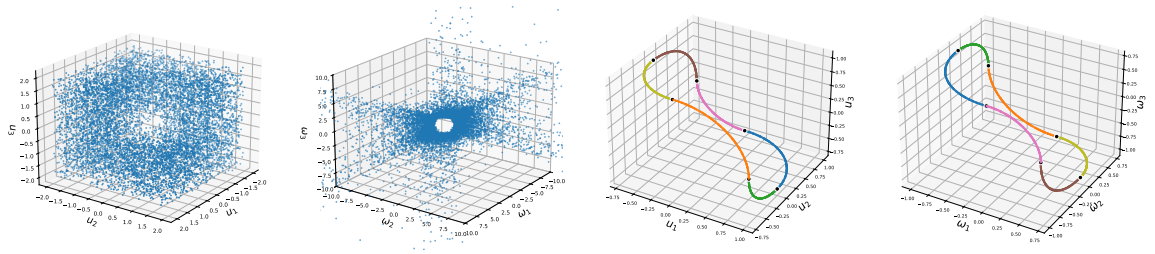
(4) $m > M^*$:
By Lemma 35 and Lemma 38, all the elements in $\Theta_{min}^*(m)$ are connected in $\Theta^*(m)$. By this and Lemmma 36, all elemets in $\Theta^*(m)$ are connected, so $\Theta^*(m)$ is connected. ■

Appendix E. Visualizations of global minima

The Figure 3 (a) (b) show samples from $\varphi^*(3)$ and (c) (d) show all the points in $\Theta^*(3)$. We can visually see that adding ℓ_2-regularization limits the set of globally optimal parameters. We also provide visualizations for global minima of other data sets and the regularization coefficient (Figure 4, 5, 6).

Appendix F. More results from Numerical Experiments

We train two-layer ReLU neural networks (1) with different widths (m=64, 128, 256, 512, 1024, 2048) using the Yacht Hydrodynamics data [14] (squared loss), and MNIST [21] (cross entropy loss). For Yacht Hydrodynamics data [14] (6-dimensional input), the network is trained with squared loss. For MNIST data [21] (784-dimensional input), the network is trained with the cross-entropy loss. We train them with stochastic gradient descent (SGD), and AdamW [24] using β as weight decay coefficient. For SGD, adding β as a weight decay coefficient is equivalent to adding the explicit ℓ_2-regularization. Whereas, for AdamW, we use squared loss or cross entropy as the minimization object and use β as a weight decay to push learned parameters towards zero. We



(a) Projection of $\varphi^*(m)$ in u-coordinates (b) Projection of $\varphi^*(3)$ in ω-coordinates (c) Projection of $\Theta^*(3)$ in u-coordinates (d) Projection of $\Theta^*(3)$ in ω-coordinates

Figure 3: (Figure 1 in main) (a) (b) **Projections of samples from $\varphi^*(3)$ onto u and ω coordinates** (Theorem 2). The blue points are samples from $\varphi^*(3)$ for $X = [1, -1]^T$, $Y = [-\frac{2}{3}, \frac{1}{2}]^T$, $\alpha = 3^1$. u values are sampled with constraint that its ℓ_∞ norm is in $[10^{-100}, 2.0]$. For each sample of u, ω that satisfies the constraints is sampled from the ω coordinate. (c) (d) **Projections of $\Theta^*(3)$ onto u and ω coordinates** (Theorem 3). The dataset and scaling is the same $X = [1, -1]^T$, $Y = [-\frac{2}{3}, \frac{1}{2}]^T$, $\alpha = 3^1$, and we additionally have weight decay $\beta = 3^{-2}$ ($\gamma_P^* = -1$, $\gamma_N^* = -0.5$). The black dots correspond to Minimal Optimal Solutions defined in Eq. 19. Different colors for u represent different sign patterns. For each u, corresponding ω has the same color in ω-coordinates.

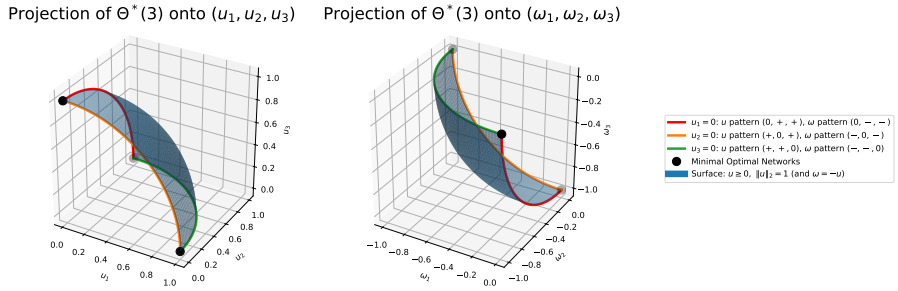


Figure 4: The projections of $\Theta^*(3)$ onto u and ω coordinates for $X = [1, -1]^T$, $Y = [-\frac{4}{9}, \frac{1}{18}]^T$, $\alpha = 3^1$, $\beta = 3^{-3}$ ($\gamma_P^* = -1$, $\gamma_N^* = 0$). The black dots correspond to Minimal Optimal Solutions defined in (19). Both the surface and the colored boundaries are globally optimal parameters. For each u, corresponding ω has the same color in ω coordinate.

use 0.9 as the first moment decay and 0.999 as the second moment decay for AdamW. For all the experiments, the learning rate is set to be 0.01.

All the plots show final test loss, which does not include the ℓ_2-regularization term, (above) and final ℓ_2-norm of parameters (bottom, y-values taken in logarithmic scale base 2). The plots show the mean of values obtained from three independent experiments.

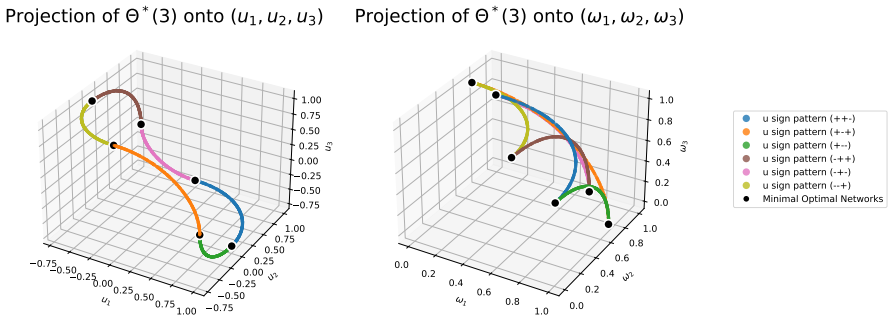


Figure 5: The projections of $\Theta^*(3)$ onto u and ω coordinates for $X = [1, -1]^T$, $Y = [\frac{2}{3}, \frac{1}{2}]^T$, $\alpha = 3^1$, $\beta = 3^{-2}$ ($\gamma_P^* = 1$, $\gamma_N^* = -0.5$). The black dots correspond to Minimal Optimal Solutions defined in (19). For each u, corresponding ω has the same color in ω coordinate.

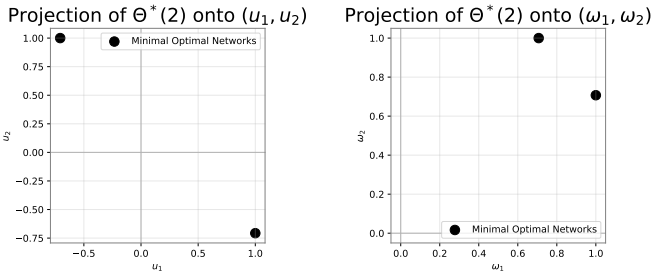


Figure 6: The projections of $\Theta^*(2)$ onto u and ω coordinates for $X = [1, -1]^T$, $Y = [1, \frac{4}{3}]^T$, $\alpha = 2^1$, $\beta = 2^{-2}$ ($\gamma_P^* = 1$, $\gamma_N^* = -0.5$). The black dots correspond to Minimal Optimal Solutions defined in (19). As $m = M^*$, $\Theta^*(m)$ is a finite set.

Table 1: Runtime hardware and software.

CPU	
Model name	AMD EPYC 7763 64-Core Processor
# CPU(s)	128
GPU	
Product Name	NVIDIA A100-SXM4-80GB
CUDA Version	12.2
PyTorch	
Version	2.7.1

We show the results when we trained the network with SGD (left) and AdamW (right). Across all the experimental settings, training with AdamW prevents the learned parameter from collapsing to zero.

F.1. Yacht Hydrodynamics data

- Figures 7 and Figure 8 show results for networks with scaling and initialization in the neural tangent kernel setting [3, 7, 16, 25] ($a = 0.5$, $b_1 = b_2 = 0$).
- Figures 9 and Figure 10 show results for networks with scaling and initialization in the mean field setting [6, 25, 26, 33] ($a = 0.5$, $b_1 = b_2 = 0$).
- Figures 11 and Figure 12 show results for small initialization ($b_1 = b_2 = 0.5$) for different scalings of the network ($a = 0.5$ for Figure11 and $a = 1$ for Figure 12).

F.2. MNIST

- Figures 13 and Figure 14 show results for networks with scaling and initialization in the neural tangent kernel setting [3, 7, 16, 25] ($a = 0.5$, $b_1 = b_2 = 0$).
- Figures 15 and Figure 16 show results for networks with scaling and initialization in the mean field setting [6, 25, 26, 33] ($a = 0.5$, $b_1 = b_2 = 0$).

F.3. Computing Environment

Experiments ran on NVIDIA A100-SXM4-80GB GPUs (CUDA 12.2) and AMD EPYC 7763 CPUs. Table 1 provides detailed hardware and software specifications.

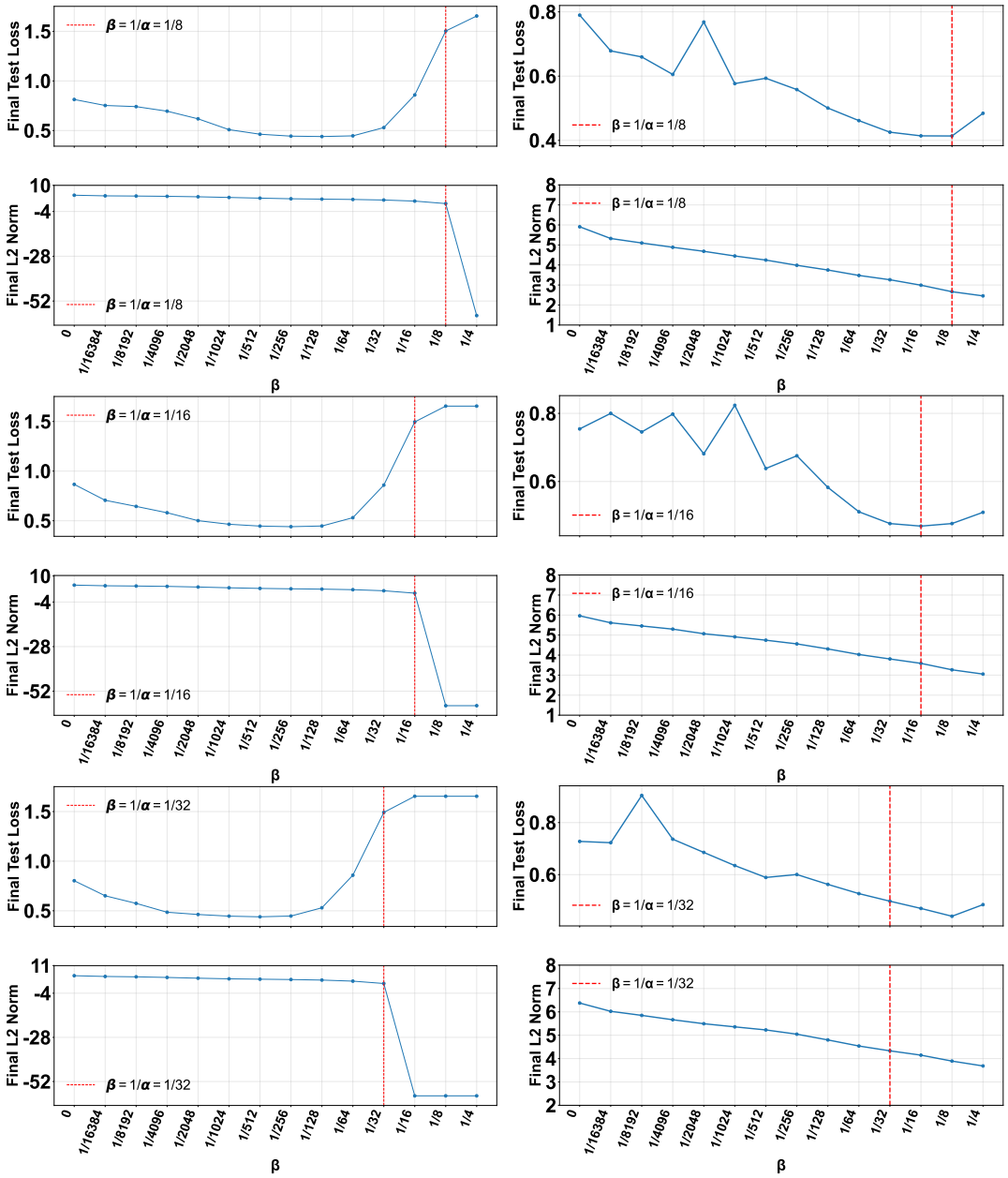


Figure 7: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/\sqrt{m}$ ($a = 0.5$) (NTK regime) and trained for 40000 epochs on Yacht Hydrodynamics. Left column: SGD. Right column: AdamW. Rows show widths 64, 256, and 1024.

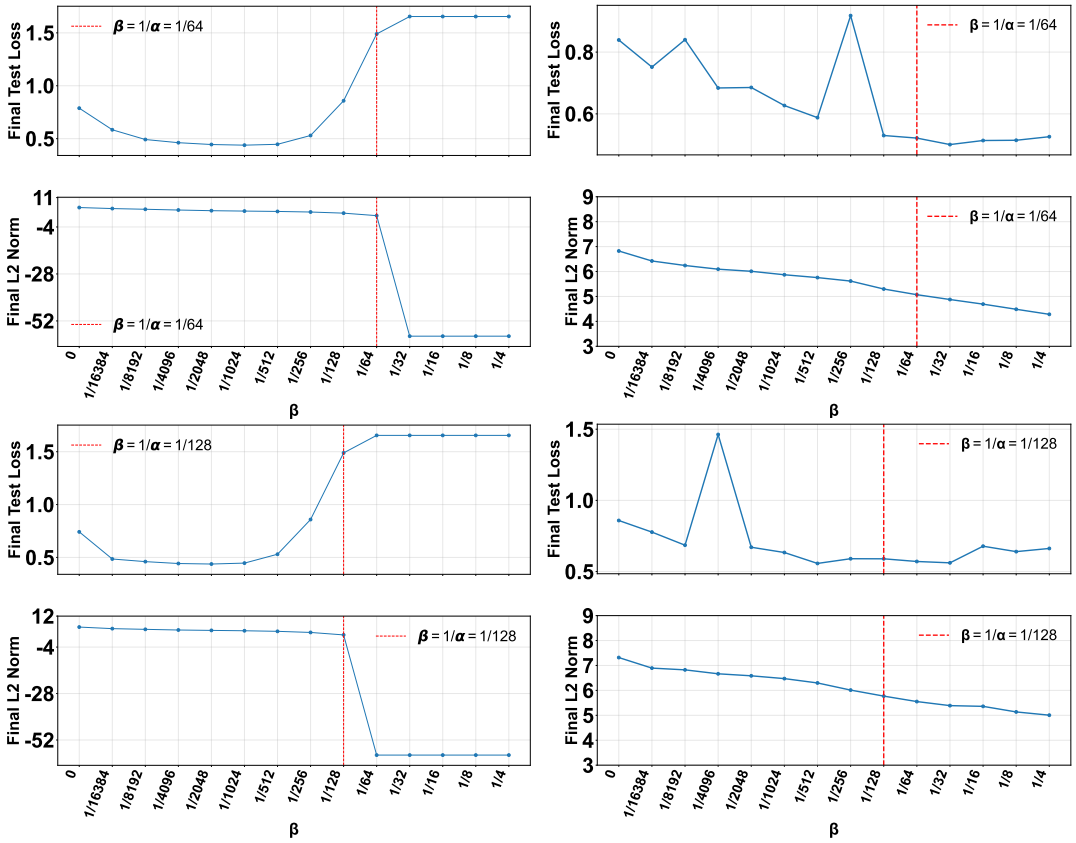


Figure 8: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/\sqrt{m}$ ($a = 0.5$) (NTK regime) and trained for 40000 epochs on Yacht Hydrodynamics. Left column: SGD. Right column: AdamW. Rows show widths 4096 and 16384.

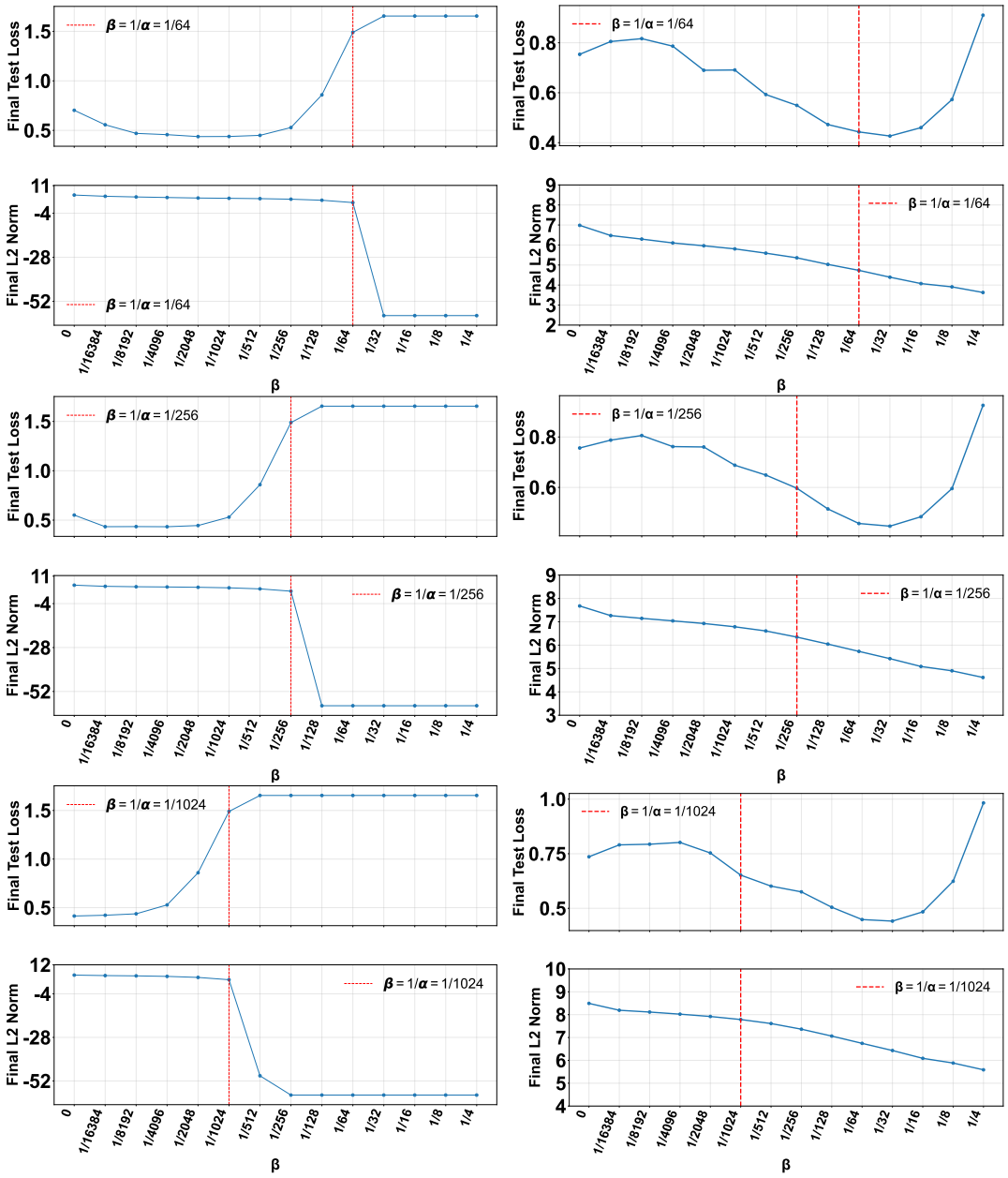


Figure 9: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/m$ ($a = 1$) (mean field regime) and trained for 40000 epochs on Yacht Hydrodynamics. Left column: SGD. Right column: AdamW. Rows show widths 64, 256, and 1024.

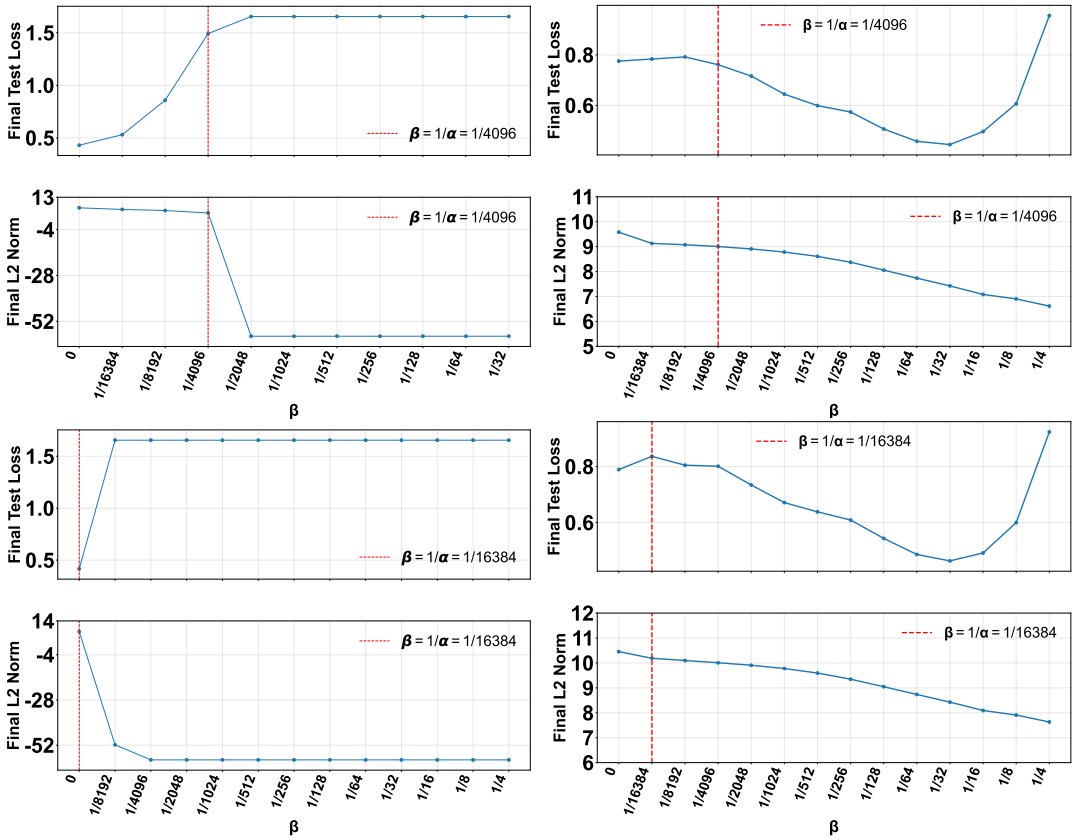


Figure 10: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/m$ ($a = 1$) (mean field regime) and trained on Yacht Hydrodynamics. Left column: SGD. Right column: AdamW. Rows show widths 4096 (240000 epochs) and 16384 (700000 epochs).

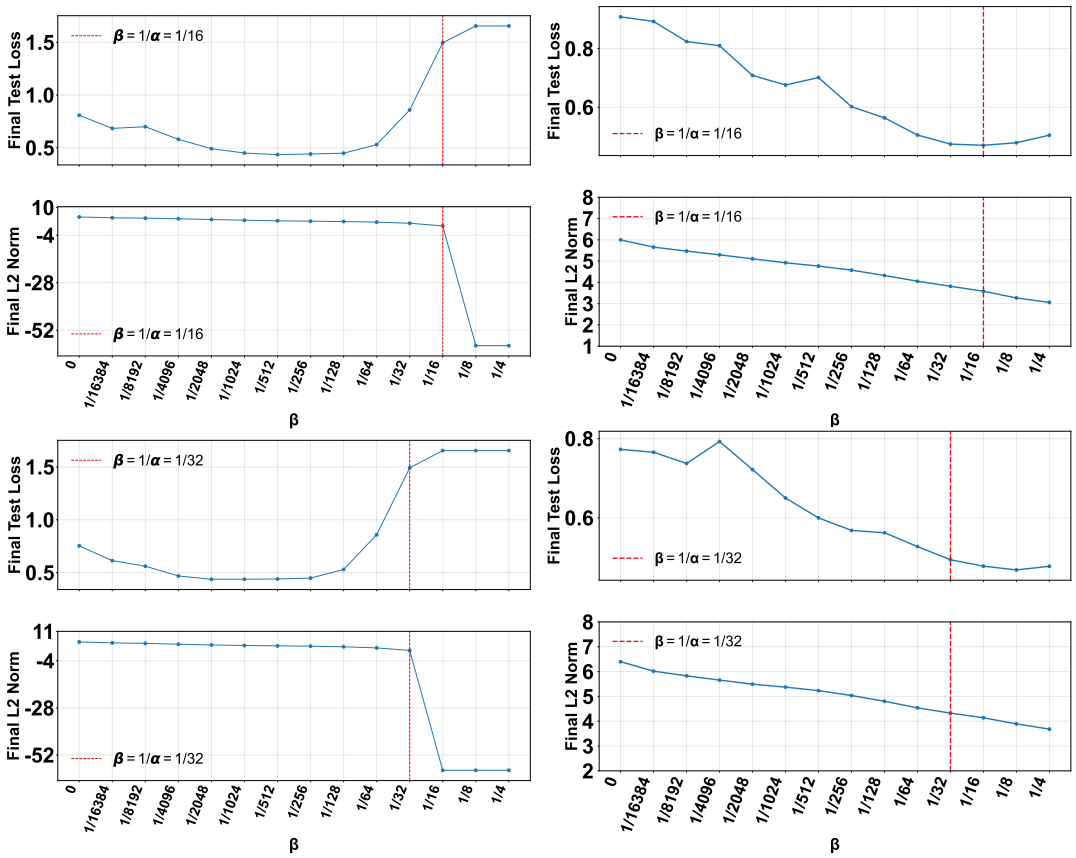


Figure 11: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01/\sqrt{m}$ ($b_1 = b_2 = 0.5$), scaled by $1/\alpha = 1/\sqrt{m}$ ($a = 0.5$) and trained for 40000 epochs on Yacht Hydrodynamics. Left column: SGD. Right column: AdamW. Rows show widths 256 and 1024.

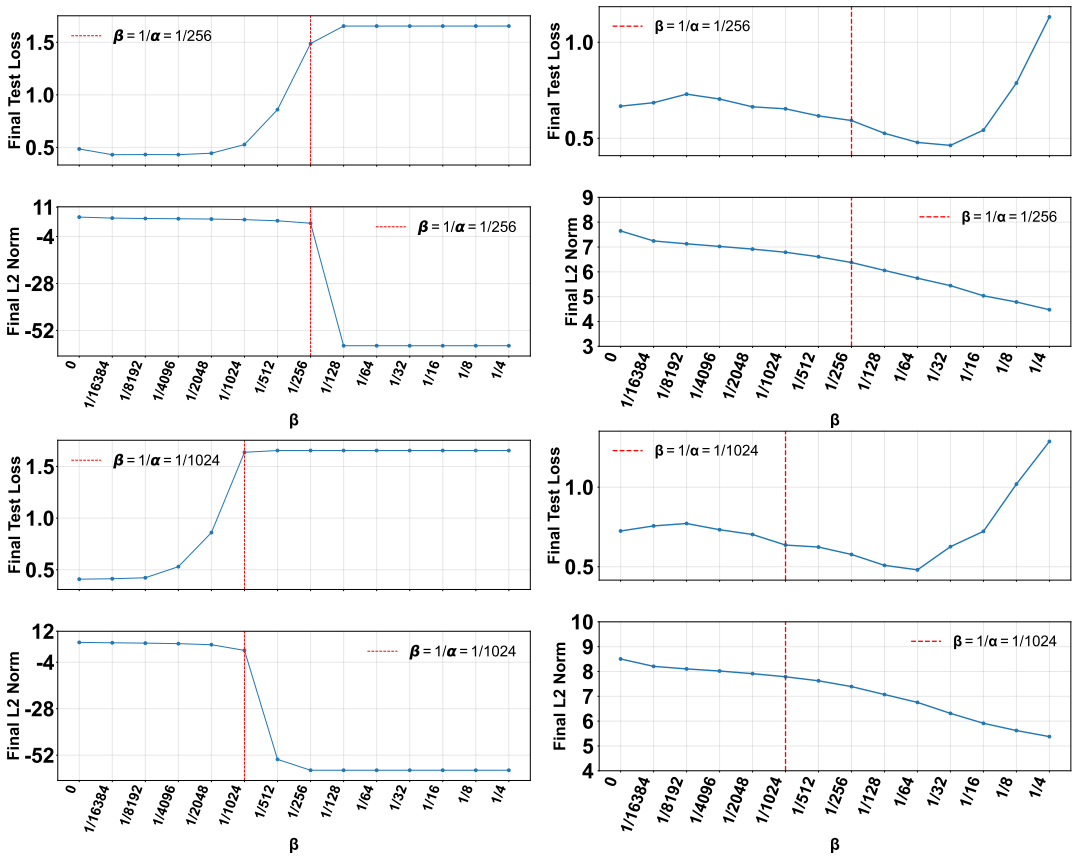


Figure 12: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01/\sqrt{m}$ ($b_1 = b_2 = 0.5$), scaled by $1/\alpha = 1/m$ ($a = 1$) and trained for 40000 epochs on Yacht Hydrodynamics. Left column: SGD. Right column: AdamW. Rows show widths 256 and 1024.

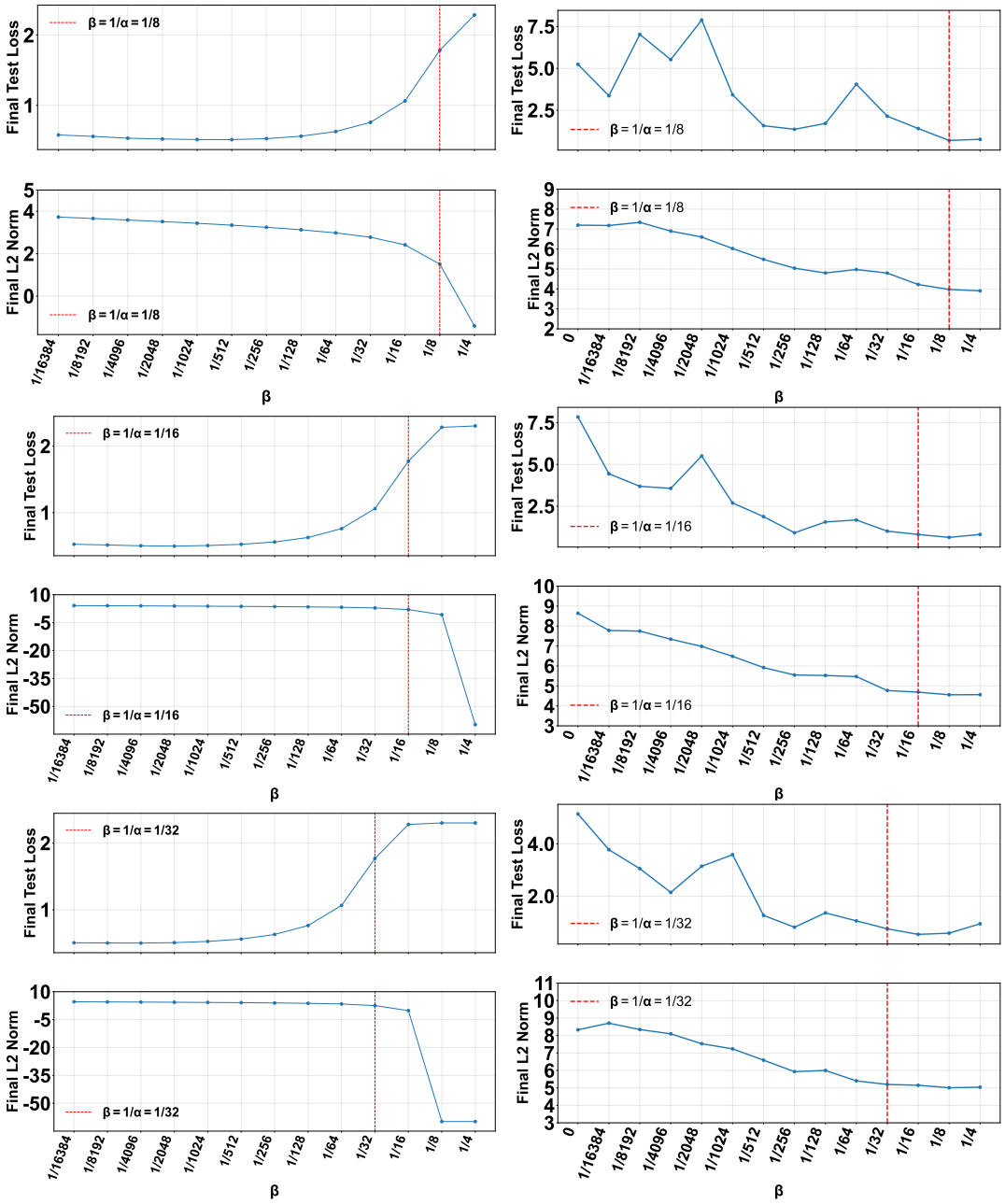


Figure 13: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/\sqrt{m}$ ($a = 0.5$) (NTK regime) and trained for 100000 epochs on MNIST. Left column: SGD. Right column: AdamW. Rows show widths 64, 256, and 1024.

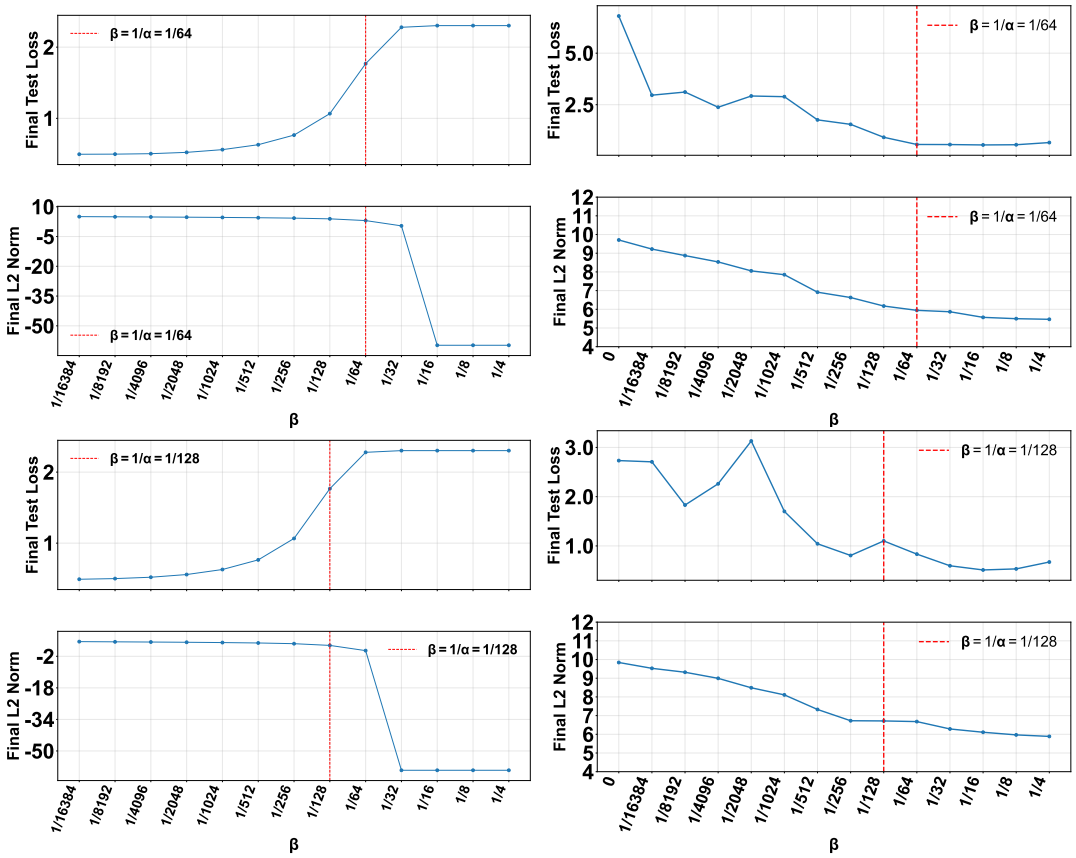


Figure 14: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/\sqrt{m}$ ($a = 0.5$) (NTK regime) and trained for 100000 epochs on MNIST. Left column: SGD. Right column: AdamW. Rows show widths 4096 and 16384.

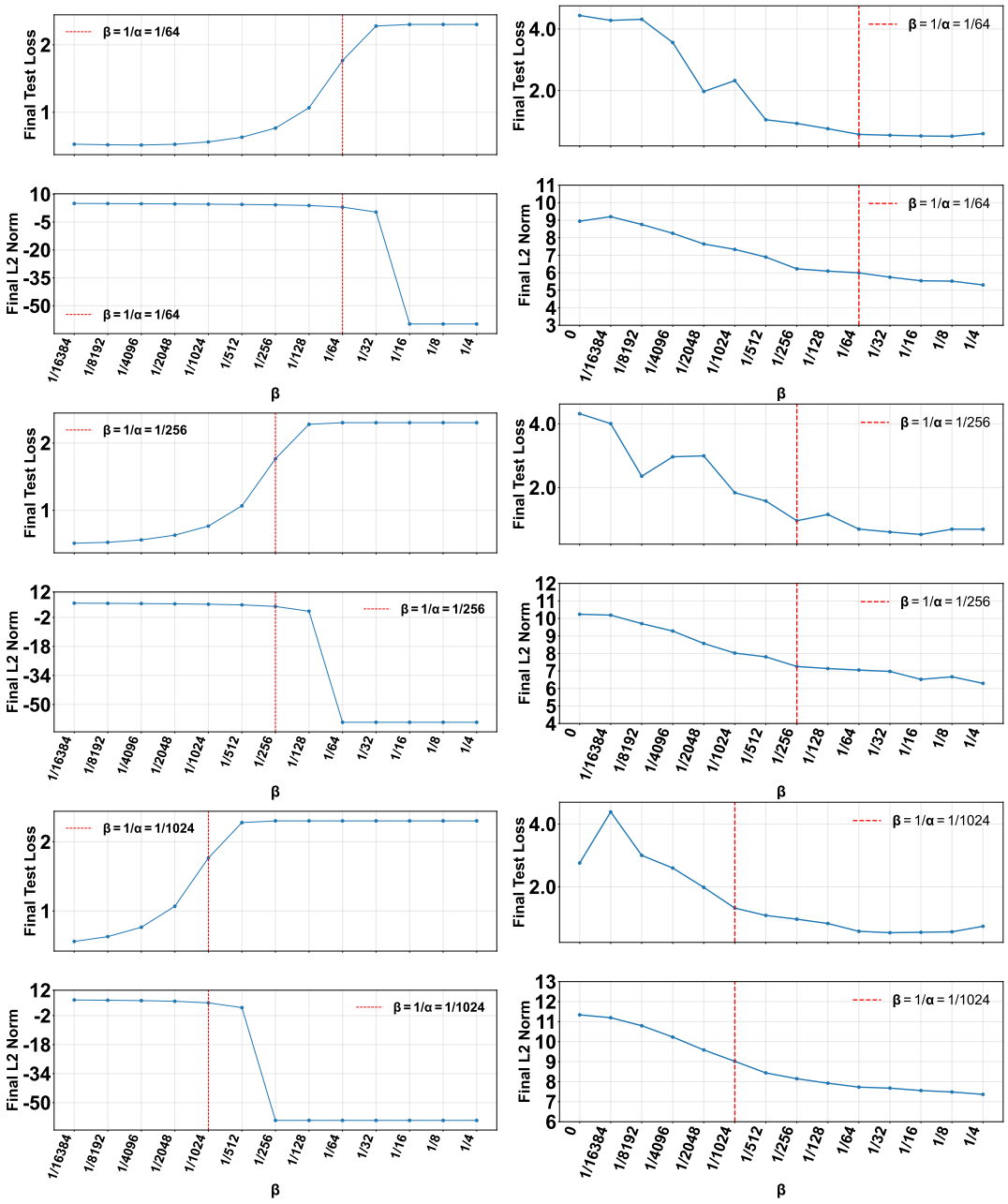


Figure 15: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/m$ ($a = 1$) (mean field regime) and trained for 100000 epochs on MNIST. Left column: SGD. Right column: AdamW. Rows show widths 64, 256, and 1024.

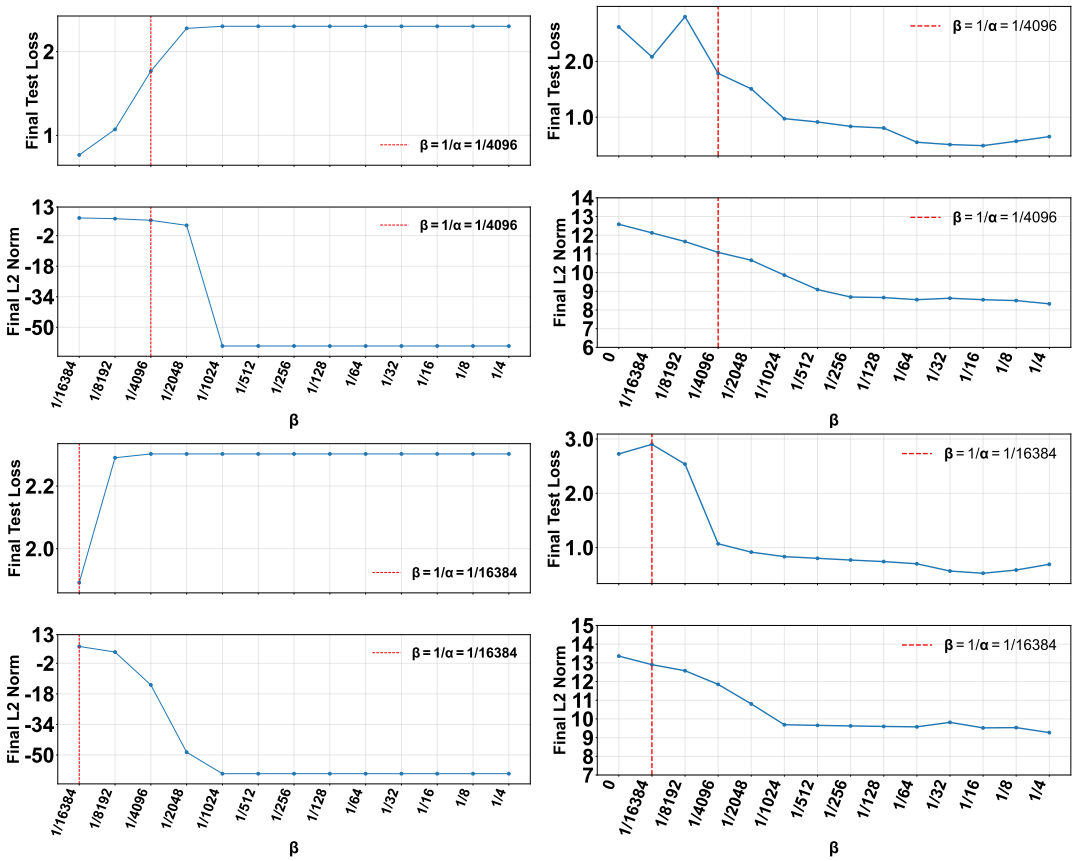


Figure 16: Two-layer ReLU networks initialized by $\tau_1 = \tau_2 = 0.01$ ($b_1 = b_2 = 0$), scaled by $1/\alpha = 1/m$ ($a = 1$) (mean field regime) and trained for 100000 epochs on MNIST. Left column: SGD. Right column: AdamW. Rows show widths 4096 and 16384.