Effects of ℓ_2-regularization and Hyperparameters on the Loss Landscape of Two-layer ReLU Networks for One-dimensional Data

Haruka Eshima

Abstract

Understanding deep neural networks (DNNs) is a central challenge in machine learning, and characterizing their loss landscapes is a key direction toward this goal. The loss landscape governs optimization dynamics and strongly influences generalization and stability of learned solutions. Even if overparameterization or choices of hyperparameters and loss functions do not change the class of functions a model can represent, they can substantially alter the geometry of the loss landscape and the locations of initialization, leading to qualitatively different learning behaviors. In this paper, we aim to explain the characterization of distinctive learning behaviors of two-layer ReLU neural networks depending on hyperparameters from a viewpoint of their loss landscape. For the first part of this work, we focus on a two-layer ReLU with one-dimensional data and then characterize the set of globally optimal parameters analytically. We analyze them with or without ℓ_2-regularization so that we can see the effects of the regularization. Based on the relationship between their geometrical properties and the locations of initial parameters, we make a separation of the hyperparameter space. Our separation provides a possible explanation for the distinctive dynamical behaviors of two-layer ReLU networks minimizing squared loss, which are often analyzed at the infinite-width limit.

1. Introduction

Loss landscape analysis characterizes quantitative variations of the loss function in finite-width machine learning models, and studying two-layer ReLU networks has been a key topic of interest in the machine learning community (Li & Yuan, 2017; Safran & Shamir, 2018; Li et al., 2020; Luo et al., 2021; Kim et al., 2025). Until recently, such analyses had been largely restricted to unregularized models. Recently, Kim et al. (2025) characterized the ℓ_2-regularized loss function by the lens of convex analysis. Specifically, they derived the convex reformulation of two-layer ReLU neural network (Pilanci & Ergen, 2020) and showed the mode connectivity of global optima for ℓ_2-regularized loss function with respect to the number of hidden neurons. Thanks to their work, global optima are found to be all connected if the number of hidden neurons is sufficiently large. While this work provides important insights into the structure of global optima, other geometric properties such as the dimension or boundedness of the optimal parameter set remain less explored. Moreover, settings in which hyperparameters depend on network width, such as the abc-parameterization (Yang & Hu, 2020), have not been explicitly addressed.

The choice of hyperparameters has been extensively studied in the context of training dynamics regimes, both theoretically (Williams et al., 2019; Kosson et al., 2024; Kunin et al., 2024; Jacot et al., 2025) and experimentally (Luo et al., 2021; Kosson et al., 2025). One well-studied regime is the linear regime (Lee et al., 2019; Geiger et al., 2020), in which the training dynamics can be approximated by a linear ordinary differential equation in function space (Yang & Hu, 2020; Chizat et al., 2019); the neural tangent kernel (NTK) (Jacot et al., 2018) is a well-known example. Another important regime is the feature learning regime (Yang & Hu, 2020; Xu & Zheng, 2024), in which neural networks learn data-dependent representations beyond their initialization. A key challenge in the feature learning regime is that training dynamics exhibit qualitatively distinct behaviors depending on the choice of hyperparameters (Luo et al., 2021; Yang & Hu, 2020). While weight decay (i.e., ℓ_2-regularization) is a fundamental component of modern neural network training (D'Angelo et al., 2024), these dynamical regimes have not yet been fully understood from a loss landscape perspective. Moreover, existing analyses are largely based on the infinite-width setting.

In this paper, we aim to characterize the loss landscape of ℓ_2-regularized two-layer neural networks with one-dimensional inputs and to elucidate the geometric properties induced by changes in hyperparameters. More specifically, we analytically

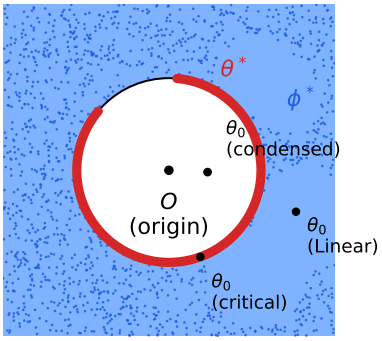


Figure 1. **A caricature of parameter space.** The figure illustrates the locations of global optima $\varphi^*(m)$ for the squared loss (densely scattered out in the blue region outside the circle of radius $\log_m \|\theta^*\|_2 = \frac{a_1+a_2}{2}$ (Proposition 4.12)), global optima $\theta^*(m)$ for the ℓ_2-regularized squared loss (the red region on the circumference (Theorem 4.11)), and the initial parameter θ_0 depending on the hyperparameter settings ($\log_m \|\theta_0\|_2 = \frac{1-2b_1}{2}$ (Proposition 5.2)). The scaling is after taking $\log_m(.)$ on norms. θ_0 is outside the circle for the linear regime, on the circumference for the critical regime, inside the circle for the condensed regime. These regimes are found by Luo et al. (2021) based on training dynamics behaviors at the infinite-width limit. (Theorem 5.3).

derive the globally optimal parameters, both with and without ℓ_2-regularization, under width-dependent hyperparameter scaling. Our geometric analysis of the global optima reveals a separation in the hyperparameter space, which corresponds to the phase diagram based on training dynamics regimes at the infinite-width limit identified by Luo et al. (2021). (See Figure 1.) Finally, we compare our theoretical predictions with numerical experiments by Luo et al. (2021), demonstrating strong agreement between theory and experiments.

Our Contributions. Our contributions are categorized into two perspectives. First, in Section 4, we provide geometric descriptions of the set of global optima $\Theta^*(m)$ for ℓ_2-regularized loss in the parameter space $\mathbb{R}^{2m}$.

- We show that adding ℓ_2 regularization does not affect the connectivity result. We provide two explanations: one based on symmetries among hidden neurons, and another via a convex formulation of neural networks proposed by Pilanci & Ergen (2020).
- We show that, for the squared loss, the set of globally optimal parameters is substantially larger than under the ℓ_2-regularized squared loss: it is unbounded and has a higher dimension.

Second, in Section 5, based on the geometric descriptions of the globally optimal parameters, we discuss the relationship between the locations of the initialized parameters and the locations of the globally optimal parameters for squared loss with or without ℓ_2-regularization.

- We classify the hyperparameter space based on this relationship. Our classification boundary matches the learning phase transition boundary for a squared loss function derived in an infinite-width setting by Luo et al. (2021).
- We further find a region in hyperparameter space where globally optimal parameters for minimizing squared loss are guaranteed to be in the neighborhood of an initialized parameter with high probability.

2. Related Work

Geometry of loss landscape for overparametrized network. There are many works analyzing the geometrical properties of the loss landscape (Achour et al., 2024; Wu et al., 2025; da Silva et al., 2025). Simsek et al. (2021) explicitly describes the manifold of global minima. Cooper (2021) analyzes the dimension of the manifold in global optima for overparameterized neural networks. Fukumizu et al. (2019) studies the landscape of the training error for overparameterized neural networks. **?** showed that every local minima are global minima for deep linear neural networks and deep ReLU neural networks but assuming independent activation.

Geometry of loss landscape for two-layer ReLU neural networks. Tian (2017) analyzes critical points of population squared loss for two-layer ReLU neural networks in a teacher-student setting, assuming that the input data are

d-dimensional spherical Gaussian input. They do not consider the effects of l_2-regularization. The l_2-regularized loss landscape is analyzed by Kim et al. (2025). They use the convex formulation of neural networks introduced by Pilanci & Ergen (2020) nicely to avoid the difficulty of analyzing non-convex loss. Their analysis is limited to connectivity analysis. However, they do not consider hyperparameter settings for initialization and network scalings introduced by Yang & Hu (2020).

Convergence to Global minima Akiyama & Suzuki (2021) showed that with sparse regularization and norm-dependent gradient descent, an overparameterized two-layer ReLU student network can recover teacher network with high accuracy in noiseless teacher-student setting.

Training Dynamics and Hyperparameter. 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 (Chizat et al., 2019; Eilers et al., 2024). NTK (Jacot et al., 2018; Wang & Zhu, 2024) is an example. In the other regime, nerual networks learn features beyond their initialization (Frei et al., 2023; Dandi et al., 2024). A well-known example is mean-field regime (Mei et al., 2018; 2019). The training admits feature learning (Yang & Hu, 2020) or learns adaptively from samples (Williams et al., 2019). Yang & Hu (2020) found the conditions for feature learning and kernel regime under a non-trivial stable feature learning. They showed L-hidden layer MLP with tanh or σ-GELU (Hendrycks, 2016) for sufficiently small σ activation function either admits feature learning or is in kernel regime, but not both. (Corollary H.14 in Yang & Hu (2020)).

These training dynamics regimes relate to the performance of neural networks. The poor generalization performance of lazy training has been reported. (Chizat et al., 2019; Buffelli et al., 2024). Yang & Hu (2020) reported that feature learning is admitted on a part of the boundary of the kernel regime. They also showed that μP (Yang, 2019; Yang et al., 2023) is a vertex on the boundary and Neural Tangent Kernel (NTK) (Jacot et al., 2018) is in the kernel regime. Their numerical experiments also confirmed that the infinite-width and finite-width μP networks outperform the NTK limit on Word2Vec model (Mikolov et al., 2013).

The theoretical analysis on these learning behaviors often assumes infinite-width setting (Luo et al., 2021; Yang & Hu, 2020; Chen et al., 2025). Existing numerical experiments with large finite-width settings confirmed that the theoretical analysis in infinite-width agrees with large finite-width setting (Lee et al., 2019; Jacot et al., 2025), but not much work has been done on theoretical analysis with finite-width. In addition, the assumption that the loss being continuously differentiable in the prediction of the model is often assumed (Chizat et al., 2019). Therefore, although ℓ_2-regularization is crucial in practice (Papyan et al., 2020; D'Angelo et al., 2024), not much work about training dynamics has been done in the presence of ℓ_2-regularization.

Section 2.1 and 4 in (Bahri et al., 2020) shows analysis on chaotic signal propagation in random deep neural networks as a function of the weight variance and bias variance

3. Preliminary

Notations. Let $X = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ be input data and $Y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ be associated targets. We denote diagonal matrices as $D(S) = \mathrm{Diag}(\mathbf{1}[X \geq 0])$ and $D(S^c) = \mathrm{Diag}(\mathbf{1}[X \leq 0])$ where $\mathbf{1}[X \leq 0]$ is an indicator vector with $(\mathbf{1}[X \leq 0])_i = 1[x_i \leq 0]$. $X_S = \mathrm{Diag}(\mathbf{1}[X \geq 0])X$ and $X_{S^c} = \mathrm{Diag}(\mathbf{1}[X \leq 0])X$. By abuse of notation, we denote $\theta = (u_1, \dots, u_m, \omega_1, \dots, \omega_m)$ by $\theta = (u_i, \omega_i)_{i=1}^m$. We assume X has at least one positive element and one negative element (so $\|X_S\|_2 \neq 0$ and $\|X_{S^c}\|_2 \neq 0$), and that the training dataset X and Y are independent of m. To prevent trivial solutions, we assume $0 < \min\{|X_S^T Y|, |X_{S^C}^T Y|\}$.

Models. Our Neural Network (NN) model is

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

where the parameter of the model is $\theta = (U^T, W^T)^T \in \mathbb{R}^{2m}$ such that $U = (u_1, \dots, u_m)^T \in \mathbb{R}^m$ and $W = (\omega_1, \dots, \omega_m)^T \in \mathbb{R}^m$ are the weights of the first and second layers, m is the number of hidden neurons and $(\cdot)_+ = \max\{\cdot, 0\}$ is the ReLU activation, and $\alpha > 0$ is the scaling factor. The parameters are initialized independently by $u_i^0 \sim N(0, \tau_1^2)$ and $\omega_i^0 \sim N(0, \tau_2^2)$ where $\tau_1, \tau_2 \in \mathbb{R}_{>0}$ are hyperparameters. By abuse of notation, we write $f(X)$ to denote the output of the

training set $(f(x_1), \ldots, f(x_n))^T \in \mathbb{R}^n$.

Hyperparameters α, τ_1 and τ_2 are introduced to analyze global optima (Section 4) or compare the loss landscape of training dynamics (Section 5) for different choices of hyperparameters.

Remark 3.1. Our hyperparameter setting is in the abc-Parameterization class for L-hidden-layer perceptron defined in Yang & Hu (2020). We can write $L = 1$, $\alpha = m^{a_1+a_2}$, $\tau_1 = m^{-b_1}$, and $\tau_2 = m^{-b_2}$. For SGD learning rate ηm^{-c}, we may assume $c = 0$ thanks to the symmetry in abc-Parametrization discussed in Section 3.2 in Yang & Hu (2020). Under this hyperparameter setting, we write a weight coefficient β, which is later introduced for a loss function Eq.(5), as m^{-d}. (Remark 3.4)

Remark 3.2. The function set that the model (1) can represent is a set of piecewise linear functions $\mathcal{F} = \{a(x)_+ + b(-x)_+ : a, b \in \mathbb{R}\}$, which is independent of the number of hidden neurons m. If we consider general d-dimensional input (this includes a model with bias), $\mathcal{F}$ becomes larger as m becomes larger. As our focus is on the change of loss landscape and its effects on learning dynamics regime in the absence of the change in the number of effective parameters, we restrict our attention to one-dimensional input.

In addition to the effects of a choice of hyperparameters, we aim to analyze the effects of adding ℓ_2-regularization on the loss landscape of this model Eq.(1). For this purpose, we compare global optima for squared loss and ℓ_2-regularized squared loss. We consider the following two non-convex optimization problems. The first problem is

$$\min_{\theta \in \mathbb{R}^{2m}} l(\theta), \tag{2}$$

$$\text{where } l(\theta) = \frac{1}{2}\left\| \frac{1}{\alpha} \sum_{j=1}^{m} (Xu_j)_+ \omega_j - Y \right\|_2^2. \tag{3}$$

is a squared loss. The second problem is

$$\min_{\theta \in \mathbb{R}^{2m}} L(\theta), \tag{4}$$

$$\text{where } L(\theta) = \frac{1}{2}\left\| \frac{1}{\alpha} \sum_{j=1}^{m} (Xu_j)_+ \omega_j - Y \right\|_2^2 + \frac{\beta}{2} \sum_{j=1}^{m} (u_j^2 + \omega_j^2) \tag{5}$$

is a squared loss and $\beta > 0$ is the weight decay coefficient for ℓ_2-regularization.

Remark 3.3. In Section 4, we show that the condition $\beta > 0$ imposes further restrictions on the forms of globally optimal solutions (See Eq.(6) and Eq.(7)). Therefore, although Eq. (2) is the case of Eq. (4) with $\beta = 0$, we treat them as different problems.

Remark 3.4. In the same way as Remark 3.1, we write the weight decay coefficient as $\beta = m^{-d}$. If we use abc-Parameterization (Yang & Hu, 2020) as in Remark 3.1, this m-dependent choice for β is plausible. If $a_1 + a_2 > 0$ and β is a constant, $\Theta^*(m)$ collapses to a trivial solution $\{(0, 0)_{i=1}^m\}$ (See discussions about Assumption 4.3).

4. Geometric Descriptions of Global Optima

In this section, we quantitatively describe the geometrical properties of the sets of global optima $\varphi^*(m)$ and $\Theta^*(m)$ for Eq. (2) and Eq. (4), respectively, depending on the number of hidden neurons m. More specifically, we first provide the analytical solutions to the sets of global optima for one-dimensional input and then show their connectivity, dimensionality, and boundedness. Note that the connectivity of global optima $\Theta^*(m)$ has already been shown in Kim et al. (2025) for general d-dimensional input case, and we provide an alternative proof for one-dimensional input. The dimensionality and boundedness of $\Theta^*(m)$ are our findings in this paper. Figure 2 conceptually summarises the results from this section.

4.1. Global optimal solution for one-dimensional input

The explicit forms of $\varphi^*(m)$ and $\Theta^*(m)$ can be found for the one-dimensional input and output case. These forms are used to derive geometrical properties later in this Section.

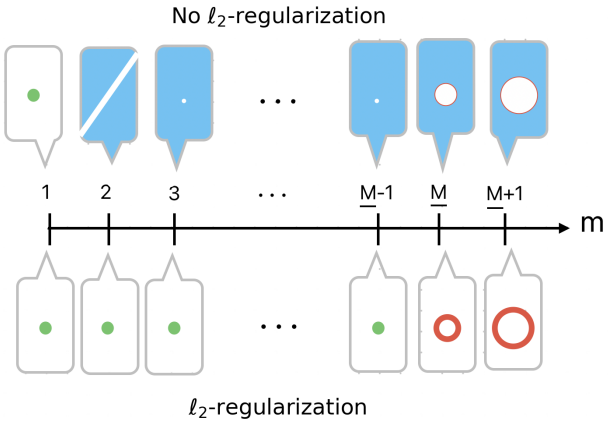


Figure 2. **A conceptual illustration of the Section 4 results.** This illustration shows geometrical changes in global optima as the number of neurons m changes. (above) **Geometrical properties of $\varphi^*(m)$.** The green dot depicts that $\{(0,0)\}_{i=1}^m$ is the only globally optimal parameter. The blue region depicts the region where global optima are densely scattered out. It is an unbounded region outside a sphere in $\mathbb{R}^{2m}$ (Proposition 4.12, 4.14). For $m=2$, global optima are exactly two connected components, and for $m \geq 3$, all the elements are connected (Theorem 4.4). (below) **Geometrical properties of $\theta^*(m)$.** $\underline{M}$ is a critical number. The red circle depicts that global optima are connected (Theorem 4.6) and have the same norm (Theorem 4.11). This conceptual illustration is inspired by Figure 1 in Kim et al. (2025).

Theorem 4.1. *The set of globally optimal parameters*

$$\varphi^*(m) = \{\theta \in \mathbb{R}^{2m} \ : \ \arg\min_{\theta \in \mathbb{R}^{2m}} l(\theta)\}$$

for squared loss Eq. (3) is

$$\varphi^*(m) := \Big\{ (u_j, \omega_j)_{j=1}^m \ \Big| \ \sum_{j:\, u_j \geq 0} u_j \omega_j = \frac{\alpha X_S^\top Y}{\|X_S\|_2^2}, \ \sum_{j:\, u_j \leq 0} u_j \omega_j = \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2} \Big\}. \tag{6}$$

Adding ℓ_2-regularization ($\beta > 0$) restricts the solution set because we need $|u_i| = |\omega_i| \ \forall i \in [m]$ for all $(u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$. (See the proof of Theorem A.3 in Appendix.) We quantitatively state the effects of adding ℓ_2-regularization on the set of global optima in Section 4.5.

Theorem 4.2. *The set of globally optimal parameters*

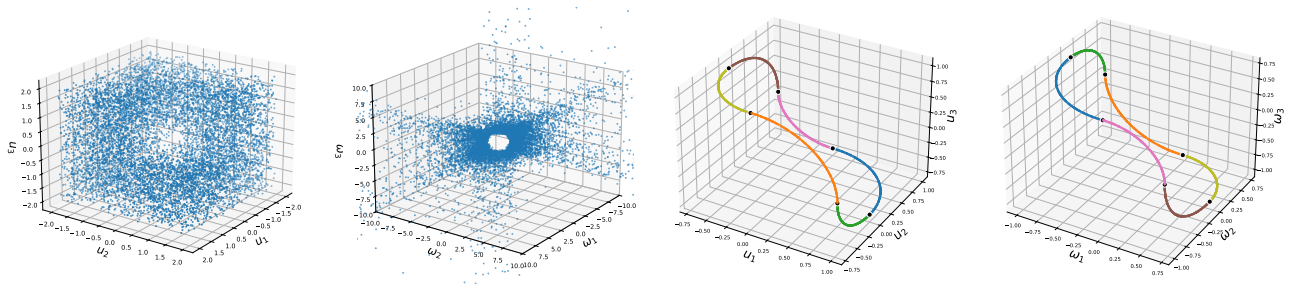
$$\Theta^*(m) = \{\theta \in \mathbb{R}^{2m} \ : \ \arg\min_{\theta \in \mathbb{R}^{2m}} L(\theta)\}$$

for ℓ_2-regularized squared loss Eq. (5)) is

$$\Theta^*(m) = \Big\{ (u_i, \omega_i)_{i=1}^m \ \Big| \ \sum_{i:\, u_i \geq 0} u_i^2 = |\gamma_P^*|, \ \sum_{i:\, u_i \leq 0} u_i^2 = |\gamma_N^*|, \ \omega_i = \begin{cases} \operatorname{sign}(\gamma_P^*)\, u_i & (u_i > 0), \\ \operatorname{sign}(\gamma_N^*)\, u_i & (u_i < 0), \\ 0 & (u_i = 0) \end{cases} \Big\} \tag{7}$$

where $\gamma_P^* = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_S^T Y)}{\|X_S\|_2^2}$ and $\gamma_N^* = \alpha \frac{\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y)}{\|X_{S^c}\|_2^2}$. ($\mathcal{S}_{\alpha\beta}(b) := \operatorname{sign}(b)\max(|b| - \alpha\beta, 0)$ is the soft-thresholding operator. $\mathcal{S}_{\alpha\beta}(b) = \operatorname{sign}(b)\max(|b| - m^{a_1+a_2-d}, 0)$ for m dependent notations in Remark 3.1.)

The Figure 3 show projections of globally optimal parameters onto u and ω coordinates when $X = [1, -1]^T$, $Y = [-\frac{2}{3}, \frac{1}{2}]^T$, $\alpha = 3^1$, $\beta = 3^{-2}$ ($\gamma_P^* = -1$, $\gamma_N^* = -0.5$). The Figure 3 (a) (b) show samples from $\varphi^*(3)$ and (c) (d) show all the points in



(a) *(b)* *(c)* *(d)*

Figure 3. Illustration of our theoretical findings. (a) (b) **Projections of samples from $\varphi^*(3)$ onto u and ω coordinates** (Theorem 4.1). 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 ω coordinate. (c) (d) **Projections of $\Theta^*(3)$ onto u and ω coordinates** (Theorem 4.2). 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.8. Different colors for u represent different sign patterns. For each u, corresponding ω has the same color in ω coordinate.

$\Theta^*(3)$. We can visually see that adding ℓ_2-regularization limits the set of globally optimal parameters. Visualizations for other datasets are provided in Appendix C.

If $a_1 + a_2 - d > 0$, then $\frac{\beta}{1/\alpha} = m^{a_1+a_2-d} \to \infty$ as $m \to \infty$, which implies $\gamma_P^*, \gamma_N^* \to 0$ as $m \to \infty$. Consequently, the optimal set $\Theta^*(m)$ collapses to a singleton, yielding the trivial solution $\{(0,0)\}_{i=1}^m$ for all sufficiently large m. In other words, when the weight decay parameter increases with m faster than the model scaling, the ℓ_2-regularizer drives all neurons to inactivity in the large-width regime. We denote the following standing assumption that avoids this situation.

Assumption 4.3. As we overparameterize the model, the weight decay coefficient β becomes much smaller compared to the scaling $1/\alpha$ of the model i.e.

$$\frac{\beta}{1/\alpha} \to 0 \text{ as } m \to \infty \iff d > a_1 + a_2.$$

4.2. Connectivity

In this section, we discuss how the connectivity of global optimal solutions $\varphi^*(m)$ and $\Theta^*(m)$ changes with respect to m.

We say $x, y \in \Theta^*(m)$ are **connected** in $\Theta^*(m)$ if $\exists$ a continuous function $f : [0, 1] \to S$ that satisfies $f(0) = x$ and $f(1) = y$. $\Theta^*(m)$ is **connected** if for any two points $x, y \in \Theta^*(m)$, x and y are connected in $\Theta^*(m)$.

There are three phases for the connectivity phase transition for global optimal parameters for squared loss.

Theorem 4.4. *We have the phase transitional behavior of the solution set.*
(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.

The connectivity result of global optimal parameters for ℓ_2-regularized squared loss with general d-dimensional input is described in five phases in Theorem 2 in (Kim et al., 2025). We find that, if we restrict our attention to one-dimensional input, there are only three phases for the connectivity phase transition. We additionally considered the case for $\Theta^*(m) = \emptyset$.

Theorem 4.5. *We have the 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.

We provide two proofs for Theorem 4.5. (See Theorem A.5 in Appendix.)

If we take the ratio $\frac{\beta}{1/\alpha}$ to be a constant independent of m, from Theorem 4.5, $\Theta^*(m)$ is always connected for $m \geq 3$ as long as $\max\{|X_S^T Y|,\ |X_{S^c}^T Y|\} > \alpha\beta$. Furthermore, we consider the effects of width-dependent hyperparameters (Remark 3.1). We find different critical widths and phase transition behaviors.

Theorem 4.6. *Assume Assumption 4.3. Define* $M^*(m) = 1[|X_S^T Y| > m^{a_1+a_2-d}] + 1[|X_{S^c}^T Y| > m^{a_1+a_2-d}]$. *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.*

Under Assumption 4.3, from Theorem 4.6, $\Theta^*(m)$ is always connected for sufficiently large m. Therefore, for both cases, the effects of ℓ_2-regularization on the connectivity results are negligible for a sufficiently overparameterized model.

From Theorem 4.5 and Theorem 4.6, we see the limiting connectivity behaviors for $\Theta^*(m) \neq \emptyset$; $\Theta^*(m)$ is a finite set, or all the elements in $\Theta^*(m)$ are connected. As Kim et al. (2025) showed, this is not always the case for general d-dimensional input data. We provide explanations for this limited connectivity result from a perspective on the roles of unnecessary neurons (Section 4.3) and a perspective on a convex formulation of neural networks introduced by Pilanci & Ergen (2020) (Section 4.4).

4.3. Roles of Permutation Symmetry and Inactive Neurons for Connectivity

Both of the two proofs for Theorem 4.5 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 (5). 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 A.9 in Appendix.) We define, by applying the concept of minimal optimal networks defined in Kim et al. (2025) to one-dimensional data, the set of Minimal Optimal Solutions as

$$\Theta^*_{\min}(m) := \Big\{(u_j, \omega_j)_{j=1}^m \ \Big|\ \forall\, p \neq q \in [m], \\ \omega_p \omega_q > 0 \Rightarrow u_p u_q < 0\Big\}. \tag{8}$$

This is the set of parameters that has least possible number of active neurons. We find that all elements in $\Theta^*_{min}(m)$ are permutations of each other. (See Appendix Lemma A.14). In Figure 3 (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 $\{(u_{\pi(i)}, v_{\pi(i)})_{i=1}^m \Big|\ \pi$ is a permutation on $\{1, ...\ .m\}\}$ are connected in $\Theta^*(m)$. (See Collorary A.18 in Appendix.) Additionally, a merging process, a process that is analogous to the merging process defined in Kim et al. (2025), proves that all the elements in $\Theta^*(m)$ are connected to a point in $\Theta^*_{min}(m)$. (See Lemma A.15). Paths created in merging process are illustrated by colored paths in Figure 3 (c) (d).

4.4. Connectivity and Convex Formulation

In this section, we provide an explanation of why the connectivity phase transition is restricted for one-dimensional input case (Theorem 4.5) from a perspective on a convex formulation introduced by Pilanci & Ergen (2020). Kim et al. (2025) 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 A.2.)

Firstly, we find the convex formulation of the non-convex problem (4). The convex formulaion of training problem is introduced by (Pilanci & Ergen, 2020). 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 4.7. *Consider the convex problem given as a cone-constrained group LASSO*

$$\min_{\substack{v_1,v_2,t_1,t_2\in\mathbb{R}\\ v_1,t_1\geq 0\\ v_2,t_2\leq 0}} L_{conv}(v_1, v_2, t_1, t_2), \tag{9}$$

$$\begin{aligned} &\text{where } L_{conv}(v_1, v_2, t_1, t_2) = \\ &\frac{1}{2}\left\|\Big((v_1 - t_1)D(S) + (v_2 - t_2)D(S^c)\Big)\frac{X}{\alpha} - Y\right\|_2^2 \\ &\quad + \beta\big(|v_1| + |v_2| + |t_1| + |t_2|\big). \end{aligned} \tag{10}$$

The convex problem (9) and the non-convex problem (7) 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 (9).*

Remark 4.8. M^* defined in Theorem 4.5 and M^* defined in Proposition 4.7 are the same. This is because γ_P^* and γ_N^* introduced in Theorem 4.2 can be written as $\gamma_P^* = v_1^* - t_1^*, \gamma_N^* = v_2^* - t_2^*$ by Proposition 4.9. We call M^* as the critical value.

Since the problem is convex, we can solve (9) directly to find optimal solutions that satisfy the constraints. It turns out that the solution set is a singleton.

Proposition 4.9. *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 (9) is $\mathcal{P}^* = \{(v_1^*, v_2^*, t_1^*, t_2^*)\}$ *where*

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

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

The staircase connectivity shown in (Kim et al., 2025) arises because we can relate the connectivity results of the cardinarity constraint set $\mathcal{P}^*(m) \subseteq \mathcal{P}^*$ defined as

$$\begin{aligned} \mathcal{P}^*(m) := \Big\{ (u_i, v_i)_{i=1}^P \;\Big|\; &(u_i, v_i)_{i=1}^P \in \mathcal{P}^*, \\ &\sum_{i\in[P]:v_i^*\neq 0} 1 + \sum_{i\in[P]:t_i^*\neq 0} 1 \leq m \Big\}. \end{aligned} \tag{11}$$

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 $m \geq M^*$ and we observe limited connectivity phase transitions.

4.5. Effects of ℓ_2-regularization for Overparameterized Model

In this section, we state the dimensionality and bounds for the set of optimal parameters. We describe them for sufficiently large m to see the effects of overparameterization. By comparing these properties for loss landscape with or without ℓ_2-regularization, we find that adding the ℓ_2-reguralization term reduces the dimension of the set of optimal parameters by m and imposes a bound on the set. For notations of hyperparameters, we use the abc-Parameteraization (Yang & Hu, 2020)) as in Remmark 3.1. 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).

There is always $(m+2)$-dimensional difference between $\Theta^*(m)$ and the parameter space. The dimensional difference grows with order m by overparameterization.

Proposition 4.10. *Under Assumption 4.3, for sufficiently large m,*

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

Theorem 4.11. *Under Assumption 4.3, for sufficiently large m, $\Theta^*(m)$ is bounded and $\forall\, \theta^* \in \Theta^*(m)$,*

$$\begin{aligned}\|\theta^*\|_2 &= \sqrt{2\alpha\left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2}\right)} \\ &= \Theta\big(m^{\frac{a_1+a_2}{2}}\big).\end{aligned}$$

If we omit ℓ_2-reguralization term, the global optimal prameters are more dense and spread out without a bound in a space outside the sphere.

Proposition 4.12. *Under Assumption 4.3, for sufficiently large m,*

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

Proposition 4.13. *For sufficiently large m,*

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

Proposition 4.14. *For sufficiently large m, $\varphi^*(m)$ is unbounded. Especially, $\forall a \geq \frac{a_1+a_2}{2}$,*

$$\exists \varphi_a^*(m) \subset \varphi^*(m)\ s.t.\ \forall \phi_a^* \in \varphi_a^*(m),\ \|\phi_a^*\|_2 = \Theta(m^a)$$

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

The dimensional difference between $\varphi^*(m)$ or $\varphi_a^*(m)$ and the parameter space is 2. This difference is independent of m. Thus, overparameterization does not change the dimentional difference.

5. Training Dynamics and Loss Landscape

Training dynamics phase transition is often studied in the context of dynamics analysis and in infinite-width. In this section, we provide an explanation for distinct learning behaviors from a perspective on loss landscape analysis on finite-width networks. For notations, we use the abc-Parametrization (Yang & Hu, 2020) as in Remark 3.1. Depending on the choice of hyperparameters, the locations of initialized parameters and the locations of globally optimal parameters are different. We compare them in a large finite-width setting, which provides a separation of hyperparameter space. We further showed that our separation is the same as that was found by Luo et al. (2021) based on learning dynamics phase transition in infinite-width setting. Figure 1 in Section 1 depicts this correspondence.

5.1. Main results

To minimize $L(\theta)$, we need $|u_i| = |\omega_i|$ due to ℓ_2-regularization. (Appendix Theorem A.3). Hence, it is plausible to consider the case $b_1 = b_2$.

Assumption 5.1. $b_1 = b_2$.

The norm of initialised model parameter θ_0 depends on b_1.

Proposition 5.2. *We denote the initialized point for model parameter as* $\theta_0 = (u_1^0, \ldots, u_m^0, \omega_1^0, \ldots, \omega_m^0)$. *Then, under Assumption 5.1,*

$$\mathbb{P}\Big(\frac{\sqrt{2}}{2}m^{\frac{1-2b_1}{2}} \leq \|\theta_0\| \leq \frac{3\sqrt{2}}{2}m^{\frac{1-2b_1}{2}}\Big) \geq 1 - 2\exp\left(-\frac{cm}{2\kappa^2}\right),$$

where $c > 0$ *is an absolute constant and* $\kappa = \|Z^2 - 1\|_{\psi_1}$ *is the sub-exponential norm of* $Z^2 - 1$ *where* $Z \sim N(0,1)$. *The left-hand side (LHS) tends to 1 as* m *tends to infinity, so* $\|\theta_0\|_2 = \Theta(m^{\frac{1-2b_1}{2}})$ *with high probability.*

Luo et al. (2021) analyzed the training dynamics behavior of parameters for two-layer ReLU neural networks for minimizing the squared loss $\frac{1}{2n}\|f_\theta(X) - Y\|_2^2$. They identified three learning regimes; the linear regime where the gradient flow of the model function is well approximated by gradient flow of its linearized model, the condensed regime where the relative change of model parameter tends to infinity, and the critical regime which serves as the boundary between the above two regimes. Luo et al. (2021) derived phase diagram that characterizes these dynamical regimes at infinite-width limit. We restate a part of their results in our notations: under Assumption 5.1, the training dynamics is linear if $a_1 + a_2 + 2b_1 < 1$, is critical if $a_1 + a_2 + 2b_1 = 1$, and is condensed if $a_1 + a_2 + 2b_1 > 1$. We deduce the same boundary conditions based on the behavior of the norm ratio between initialized parameters and global optima for ℓ_2-reguralized squared loss for large finite-width setting: under Assumption 5.1, for large m, with high probability, $\|\theta_0\|_2 \gg \|\theta^*\|_2$ if $a_1 + a_2 + 2b_1 < 1$, $\|\theta_0\|_2 \sim \|\theta^*\|_2$ if $a_1 + a_2 + 2b_1 = 1$, and $\|\theta_0\|_2 \ll \|\theta^*\|_2$ if $a_1 + a_2 + 2b_1 > 1$.

Theorem 5.3. *Assume Assumptions 4.3 and 5.1 hold and that* m *is sufficiently large. With probability at least* $1 - 2\exp(-Cm)$ *(*C *is a positive constant independent of* m*),*

$$\frac{\|\theta_0\|_2}{\|\theta^*\|_2} = \Theta(m^{\frac{1-a_1-a_2-2b_1}{2}}).$$

The behavior of the value $\frac{\|\theta_0\|_2}{\|\theta^*\|_2}$ *when overparameterizing (i.e. increasing* m*) depends on the sign of* $1 - a_1 - a_2 - 2b_1$.

- *If* $a_1 + a_2 + 2b_1 < 1$, $\frac{\|\theta_0\|_2}{\|\theta^*\|_2} \to \infty$ *with probability one as* $m \to \infty$.
- *If* $a_1 + a_2 + 2b_1 = 1$, $\frac{\|\theta_0\|_2}{\|\theta^*\|_2} = \Theta(1)$ *and is away from 0 with probability one as* $m \to \infty$.
- *If* $a_1 + a_2 + 2b_1 > 1$, $\frac{\|\theta_0\|_2}{\|\theta^*\|_2} \to 0$ *with probability one as* $m \to \infty$.

Remark 5.4. Although we use $m \to \infty$ in the last part, our critical quantity $\Theta(m^{\frac{1-a_1-a_2-2b_1}{2}})$ is derived from a finite-width setting.

From a context of dynamics analysis in infinite-width setting, Luo et al. (2021) proved that the maximum deviation of parameter values relative to the norm of initial parameter during training tends to 0 as the width tends to infinity in linear regime. From a context of loss landscape analysis in sufficiently large finite-width setting, we find that the set of global optima $\varphi^*(m)$ for squared loss is densely spread out outside the sphere on which $\Theta^*(m)$ exists (Theorem 4.1), and the initial starting point θ_0 is outside the sphere with high probability (Theorem 5.3) in the same hyperparameter setting as linear regime. Our analysis of the loss landscape provides an alternative explanation for learning dynamics to converge to a neighborhood of the initialized parameter for the linear regime. Theorem 5.5 further proves that there is a region where a global optimizer for squared loss is in the neighborhood of initialized parameter with high probability.

Theorem 5.5. *Under Assumptions 4.3 and 5.1, for sufficiently large* m*, with probability at least* $1 - 8\exp(-Cm)$,

$$\exists\, \phi^* \in \varphi^*(m) \; s.t. \; \|\theta_0 - \phi^*\|_2^2 = O(m^{2(a_1+a_2)-1+2b_1}),$$

where $C > 0$ *is an absolute constant. In particular, for* $a_1 + a_2 + b_1 \leq \frac{1}{2}$,

$$\exists\, \phi^* \in \varphi^*(m) \; s.t. \; \|\theta_0 - \phi^*\|_2^2 = O(1)$$

with high probability.

For $a_1 + a_2 \geq 0$, which can be assumed under stable abc-Paramtrization introduced by Yang & Hu (2020), the region $a_1 + a_2 + b_1 \leq \frac{1}{2}$ is within the linear regime region. The hyperparameter initialization for NTK (Jacot et al., 2018) ($a_1 + a_2 = \frac{1}{2}$, $b_1 = 0$) is in this region.

5.2. What are the implications of our theoretical results?

We hypothesize that hyperparameter setting in the critical regime does implicit ℓ_2-norm regularization.

Numerical experiments by Luo et al. (2021) show that relative change of parameter norm depends on hyperparameter settings. We focus on the case $\gamma = 0$, i,e, $b_1 = b_2$ in our notation. In the linear regime ($\gamma' < 1$), they show that the maximum deviation of parameter updates for squared loss tends to zero as we increase the width m (Figure 10 in Luo et al. (2021)). In the critical regime ($\gamma' = 1$), they show the maximum deviation tends to the same order as the norm of the initialized parameter as we increase the width m (Figure 10 in Luo et al. (2021)). If parameters converge to their nearest global minima, the convergence point for parameters initialized by the critical regime converges to a global minima that has a smaller order of norm with respect to m compared to the norm of a global minima for l_2-squared loss.

When $b_1 = b_2$, the critical regime corresponds to mean-field hyperparameter setting for two-layer neural network model. The smaller order of l_2-norm could explain good properties of mean field initialization such as μ-P parameterization in (Yang & Hu, 2020).

We hypothesize that adding ℓ_2-regularization will prevent lazy learning, which is known to lead to poor generalization performance.

6. Conclusion

In this paper, we aim to elucidate the effects of ℓ_2-regularization and hyperparameters on the loss landscape of Two-layer ReLU networks for one-dimensional Data. More specifically, we first obtained closed-form global optima of the model and analyzed their geometric properties including connectivity, boundedness, and dimensionality. These results allowed us to relate the norm of initialized parameters to that of the corresponding optima under squared loss, with and without ℓ_2-regularization, across different hyperparameter settings. We then use this relationship to explain the training dynamics phase transition of two-layer ReLU networks. It is important to note that, although our analysis is restricted to one-dimensional inputs and outputs, we showed that the phase transition in training dynamics can be explained from a loss-landscape perspective. Expanding our theoretical analysis to the d-dimensional input case remains as future work. This viewpoint also provides a path toward understanding training-dynamics behavior in finite-width settings.

Impact Statement

This work advances the theoretical understanding of the loss landscape and hyperparameter space, with potential applications in hyperparameter tuning. We do not identify any ethical concerns but recognize that implications may vary depending on the domain of application.

References

Achour, E. M., Malgouyres, F., and Gerchinovitz, S. The loss landscape of deep linear neural networks: a second-order analysis. *Journal of Machine Learning Research*, pp. 1–76, 2024.

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

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., and Ganguli, S. Statistical mechanics of deep learning. *Annual review of condensed matter physics*, 11(1):501–528, 2020.

Buffelli, D., McGowan, J., Xu, W., Cioba, A., Shiu, D.-s., Hennequin, G., and Bernacchia, A. Exact, tractable gauss-newton optimization in deep reversible architectures reveal poor generalization. *Advances in Neural Information Processing Systems*, 2024.

Chen, Z., Yang, G., Zhao, Q., and Gu, Q. Global convergence and rich feature learning in l-layer infinite-width neural networks under μ parametrization. In *International Conference on Machine Learning*, 2025.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in neural information processing systems*, 2019.

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

da Silva, M. F., Dangel, F., and Oore, S. Hide & seek: Transformer symmetries obscure sharpness & riemannian geometry finds it. In *International Conference on Machine Learning*, 2025.

Dandi, Y., Krzakala, F., Loureiro, B., Pesce, L., and Stephan, L. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, pp. 1–65, 2024.

D'Angelo, F., Andriushchenko, M., Varre, A. V., and Flammarion, N. Why do we need weight decay in modern deep learning? *Advances in Neural Information Processing Systems*, 2024.

Eilers, L., Memmesheimer, R.-M., and Goedeke, S. A generalized neural tangent kernel for surrogate gradient learning. *Advances in Neural Information Processing Systems*, 2024.

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

Fukumizu, K., Yamaguchi, S., Mototake, Y.-i., and Tanaka, M. Semi-flat minima and saddle points by embedding neural networks to overparameterization. *Advances in neural information processing systems*, 2019.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

Hendrycks, D. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 2018.

Jacot, A., Súkeník, P., Wang, Z., and Mondelli, M. Wide neural networks trained with weight decay provably exhibit neural collapse. In *International Conference on Learning Representations*, 2025.

Kim, S., Mishkin, A., and Pilanci, M. Exploring the loss landscape of regularized neural networks via convex duality. In *International Conference on Learning Representations*, 2025.

Kosson, A., Messmer, B., and Jaggi, M. Rotational equilibrium: How weight decay balances learning across neural networks. In *International Conference on Machine Learning*, 2024.

Kosson, A., Welborn, J., Liu, Y., Jaggi, M., and Chen, X. Weight decay may matter more than mup for learning rate transfer in practice. *arXiv preprint arXiv:2510.19093*, 2025.

Kunin, D., Raventos, A., Dominé, C. C. J., Chen, F., Klindt, D., Saxe, A. M., and Ganguli, S. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, 2024.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 2019.

Li, Y. and Yuan, Y. Convergence analysis of two-layer neural networks with relu activation. *Advances in neural information processing systems*, 2017.

Li, Y., Ma, T., and Zhang, H. R. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, 2020.

Luo, T., Xu, Z.-Q. J., Ma, Z., and Zhang, Y. Phase diagram for two-layer relu neural networks at infinite-width limit. *Journal of Machine Learning Research*, 2021.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 2018.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, 2019.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

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

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, 2018.

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

Tian, Y. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *International conference on machine learning*, 2017.

Wang, Z. and Zhu, Y. Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks. *The Annals of Applied Probability*, 34(2):1896–1947, 2024.

Williams, F., Trager, M., Panozzo, D., Silva, C., Zorin, D., and Bruna, J. Gradient dynamics of shallow univariate relu networks. *Advances in neural information processing systems*, 2019.

Wu, F. Z., Simsek, B., and Ged, F. G. Loss landscape of shallow reLU-like neural networks: Stationary points, saddle escape, and network embedding. In *International Conference on Learning Representations*, 2025.

Xu, X. and Zheng, L. Neural feature learning in function space. *Journal of Machine Learning Research*, 25(142):1–76, 2024.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

Yang, G. and Hu, E. J. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Yang, G., Simon, J. B., and Bernstein, J. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

A. Geormetric Descriptions of Global Optima: Proofs

Theorem A.1. *(Theorem 4.1 in main) The set of global optimal prameters*

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

for squared loss is

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

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

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

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 one-dimensional input case, we prove that the minimum value for original non-convex problem (4) equals to the minimum value for the convex problem (12) in Lemma A.2.

Lemma A.2. *Consider the optimization problem*

$$\min_{\gamma_P,\gamma_N\in\mathbb{R}} \frac{1}{2}\|\,\gamma_P\frac{X_S}{\alpha} + \gamma_N\frac{X_{S^c}}{\alpha} - Y\,\|_2^2 + \beta(|\gamma_P| + |\gamma_N|). \tag{12}$$

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

$$(\gamma_P^*, \gamma_N^*) = \left(\alpha\frac{\mathcal{S}_{\alpha\beta}(X_S^T Y)}{\|X_S\|_2^2}, \alpha\frac{\mathcal{S}_{\alpha\beta}(X_{S^c}^T Y)}{\|X_{S^c}\|_2^2}\right). \tag{13}$$

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

Proof. Note that the minimization problem (12) 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}. □

Theorem A.3. *(Theorem 4.2 in the main). The global optimum $\Theta^*(m) = \{\theta \in \mathbb{R}^{2m} : \arg\min_{\theta\in\mathbb{R}^{2m}} L(\theta)\}$ for L defined as (5) is*

$$\Theta^*(m) = \left\{(u_i, \omega_i)_{i=1}^m \;\middle|\; \sum_{i:\,u_i\geq 0} u_i^2 = |\gamma_P^*|,\ \sum_{i:\,u_i\leq 0} u_i^2 = |\gamma_N^*|,\ \omega_i = \begin{cases} \text{sign}(\gamma_P^*)\,u_i & (u_i > 0), \\ \text{sign}(\gamma_N^*)\,u_i & (u_i < 0), \\ 0 & (u_i = 0) \end{cases}\right\} \tag{14}$$

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

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

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,

$$\begin{aligned}
&\frac{\beta}{2}\sum_{j=1}^m (u_j^2 + \omega_j^2) \\
&\geq \frac{\beta}{2}\sum_{j:u_j>0}(u_j^2+\omega_j^2) + \frac{\beta}{2}\sum_{j:u_j<0}(u_j^2+\omega_j^2) \quad \text{(equality holds iff } u_j = 0 \Rightarrow \omega_j = 0) && (15)\\
&\geq \beta\sum_{j:u_j\geq 0}|u_j||\omega_j| + \beta\sum_{j:u_j\leq 0}|u_j||\omega_j| = \beta\sum_{j:u_j\geq 0}|u_j\omega_j| + \beta\sum_{j:u_j\leq 0}|u_j\omega_j| \quad \text{(equality holds iff } \forall\, j,\ |u_j| = |\omega_j|) && (16)\\
&\geq \beta\left|\sum_{j:u_j\geq 0} u_j\omega_j\right| + \beta\left|\sum_{j:u_j\leq 0} u_j\omega_j\right| = \beta|\gamma_P| + \beta|\gamma_N| \quad \text{(equality holds iff } \forall j, i \text{ s.t. the product } u_ju_i > 0,\ \text{sign}(u_j\omega_j) = \text{sign}(u_i\omega_i)). && (17)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\min_{\{u_i,\omega_i\}_{i=1}^m\in(\mathbb{R}\times\mathbb{R})^m} \frac{1}{2}\left\|\frac{1}{\alpha}\sum_{j=1}^m (Xu_j)_+\omega_j - Y\right\|_2^2 + \frac{\beta}{2}\sum_{j=1}^m (u_j^2+\omega_j^2) \\
&\geq \min_{\gamma_P,\gamma_N\in\mathbb{R}} \frac{1}{2}\|\,\gamma_P\frac{X_S}{\alpha} + \gamma_N\frac{X_{S^c}}{\alpha} - Y\,\|_2^2 + \beta(|\gamma_P| + |\gamma_N|) \qquad (18)\\
&= \frac{1}{2}\|\,\gamma_P^*\frac{X_S}{\alpha} + \gamma_N^*\frac{X_{S^c}}{\alpha} - Y\,\|_2^2 + \beta(|\gamma_P^*| + |\gamma_N^*|).
\end{aligned}$$

The equality in the last holds by Lemma A.2 where (γ_P^*, γ_N^*) is defined as (13).
By (15), (16) and (17),

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

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

$$|\gamma_P^*| = \left| \sum_{i:u_i^* \geq 0} u_i^* \omega_i^* \right| = \left| \sum_{i:u_i^* \geq 0} {u_i^*}^2 \right| = \sum_{i:u_i^* \geq 0} {u_i^*}^2, \tag{19}$$

$$|\gamma_N^*| = \left| \sum_{i:u_i^* \leq 0} u_i^* \omega_i^* \right| = \sum_{i:u_i^* \leq 0} {u_i^*}^2, \tag{20}$$

$$\omega_i^* = \begin{cases} \text{sign}(\gamma_P^*) u_i^* & (u_i^* > 0), \\ \text{sign}(\gamma_N^*) u_i^* & (u_i^* < 0), \\ 0 & (u_i^* = 0). \end{cases} \tag{21}$$

(16) and (17) 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 (19). Similarly, (16) and (17) imply (20).
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 (16) and (17), we need (21).
Hence, $(u_i^*, \omega_i^*)_{i=1}^m \in \Theta^*(m)$ if and only if it satisfies (19) (20) (21). □

A.1. Connectivity

Theorem A.4. *We have the phase transitional behavior of the solution set.*
(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 $A \coloneqq \frac{\alpha X_S^\top Y}{\|X_S\|_2^2}$ and $B \coloneqq \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}$.
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

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

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

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

and

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

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

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

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

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 prove the connectivity result (Theorem 4.5) of global optimal parameters $\Theta^*(m)$ in two different ways.

Theorem A.5. *(Theorem 4.5 in the main) We have 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 use the the explicit form of $\Theta^*(m)$. The second proof follows the principle ideas introduced by Kim et al. (2025). We provide the first proof in this section and the second proof is given in the next section (Appendix A.2).

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 (19) (20) (21),

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

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

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

   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

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

   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.

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

   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.

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

$\theta \in [0, 2\pi)$ 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 A.2.

Theorem A.6. *Assume Assumption 4.3. Define* $M^*(m) = 1[|X_S^T Y| > m^{a_1+a_2-d}] + 1[|X_{S^c}^T Y| > m^{a_1+a_2-d}]$. *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. $m^{a_1+a_2-d}$ is strictly decreasing and $m^{a_1+a_2-d} \to 0$ 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.

$$M^*(m) = 1[|X_S^T Y| > m^{a_1+a_2-d}] + 1[|X_{S^c}^T Y| > m^{a_1+a_2-d}] = 1[|\gamma_P^*| > 0] + 1[|\gamma_N^*| > 0].$$

Also,

$$M^*(m) = 0 \iff m < \underline{M}, \tag{22}$$
$$M^*(m) = 1 \iff \underline{M} \leq m < \overline{M}, \tag{23}$$
$$M^*(m) = 2 \iff \overline{M} \leq m. \tag{24}$$

(1) $m < \underline{M}$:
From (22), $M^*(m) = 0$, so the result from Theorem A.5 (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 (24), $M^(1) = 2$ 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 (23), $1 = m = M^*(1)$ iff $\underline{M} \leq 1 = m < \overline{M}$. By (24), $2 = m = M^*(2)$ 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 A.5 (4), $\Theta^*(m)$ is connected. □

A.2. Connectivity and Convex Formulation

In general, convex proplems are easier to deal with than non-convex problems. Kim et al. (2025) showed that the staircase connectivity of global optimal solutions to the original non-convex training loss minimization problem can be found by analysing 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. (2025).

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 (14) throughout Section A.2.

Fistly, we find the convex formulation of our non-convex problem (4).

Proposition A.7. *(Proposition 4.7 in main) Consider the convex problem given as a cone-constrained group LASSO*

$$\min_{v_1,v_2,t_1,t_2} \frac{1}{2}\left\|\Big((v_1-t_1)D(S)+(v_2-t_2)D(S^c)\Big)\frac{X}{\alpha}-Y\right\|_2^2 \quad +\beta\big(|v_1|+|v_2|+|t_1|+|t_2|\big) \\ \text{s.t.} \quad v_1\geq 0,\ t_1\geq 0,\ v_2\leq 0,\ t_2\leq 0 \tag{25}$$

where $D(S)=\mathrm{Diag}(\mathbf{1}[X\geq 0])$ *and* $D(S^c)=\mathrm{Diag}(\mathbf{1}[X\leq 0])$.
The convex problem (25) and the non-convex problem (4) 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 (25).*

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

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

(5) can be equivalently written as

$$L(\theta)=\frac{1}{2}\|\sum_{j=1}^m(\frac{X}{\alpha}u_j)_+\omega_j-Y\|_2^2+\frac{\beta}{2}\sum_{j=1}^m(u_j^2+\omega_j^2) \tag{26}$$

Consider the following convex problem

$$\min_{\{v_i,t_i\}_{i=1}^3}\frac{1}{2}\left\|\Big((v_1-t_1)D(S)+(v_2-t_2)D(S^c)+(v_3-t_3)I_n\Big)\frac{X}{\alpha}-Y\right\|_2^2 \quad +\beta\sum_{i=1}^3\big(|v_i|+|t_i|\big) \\ \text{s.t.}\quad (2D_i-I_n)Xv_i\geq 0,\quad (2D_i-I_n)Xt_i\geq 0,\ \forall i\in[3].\ \ (D_1=D(S),\ D_2=D(S^c),\ D_3=I_n) \tag{27}$$

By applying Theorem 1 in Pilanci & Ergen (2020) to our problem with $d=1$ and input matrix $\frac{X}{\alpha}$, the convex problem (27) and the non-convex problem (4) 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 (25). As X has both positive and negative elements,

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

Rewriting (27) with this condition gives (25). □

To find the connectivity of $\Theta^*(m)$, we use this convex formulation. We find the solution set $\mathcal{P}^*$ to the convex problem (25) is a singleton.

Proposition A.8. *(Proposition 4.9 in main) 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 (25) is $\mathcal{P}^* = \{(v_1^*, v_2^*, t_1^*, t_2^*)\}$ *where*

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

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

Proof.

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

By the KKT condition, at an optimal,

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

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

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

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 A.9. $\forall (u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$, $|u_i| = |\omega_i|$ *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 (1). 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, $r^2u_i^2 + \frac{\omega_i^2}{r^2} \geq 2|u_i||\omega_i|$ with equality iff $r^2u_i^2 = \frac{\omega_i^2}{r^2}$.
For $r > 0$, $r^2u_i^2 = \frac{\omega_i^2}{r^2}$ 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 A.5.

1. Construct functions that map elements in $\mathcal{P}^*$ and elements in $\Theta^*(m)$. (Definition A.10 and Definition A.11)
2. Introduce the notion of Minimal Optimal Solution. (Definition A.13)
3. Prove that Minimal Optimal Solutions are permutations of each other. (Lemma A.14)
4. Prove that any optimal solution is connected to a Minimal Optimal Solution. (Lemma A.15)
5. For $m = M^*$, prove that all the optimal solutions are Minimal Optimal Solutions. (Lemma A.16)
6. For $m > M^*$, prove that all the permutation solutions are connected in $\Theta^*(m)$. (Lemma A.17)

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 A.10. 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 A.11. Suppose $m \geq M^*$. Define $\Phi : \Theta^*(m) \to \mathcal{P}^*$ as

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

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 A.12. *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 A.7, 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 A.9.
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 A.7, 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 A.13. 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 A.14 proves that all the Minimal Optimal Solutions are permutations of each other.

Lemma A.14. *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 A.10, $v_1^*, t_1^* \geq 0$ 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 A.13 implies that $\mathcal{I}$ introduced in Definition A.11 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 A.9, $P+Q = \sum_{i=1}^m 1(u_i\omega_i \neq 0)$, 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 A.9, $\forall j,\ \sqrt{|u_j\omega_j|} = |\omega_j|$. 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 A.15. *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 $u_1u_2 > 0$. Denote $\mathrm{sign}(\omega_1) = \mathrm{sign}(\omega_2)$ by s. Define a path

$$C(t) = \left(\frac{u_1|\omega_1| + tu_2|\omega_2|}{\sqrt{|u_1\omega_1 + tu_2\omega_2|}}, \sqrt{|u_1\omega_1 + tu_2\omega_2|}s\right) \oplus \left(\sqrt{1-t}\frac{u_2|\omega_2|}{\sqrt{|u_2\omega_2|}}, \sqrt{(1-t)|u_2\omega_2|}s\right) \oplus (u_j,\omega_j)_{j=3}^m$$

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 $\mathrm{sign}(u_1\omega_1) = \mathrm{sign}(u_2\omega_2) \neq 0$. 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 A.9. 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

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

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

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

   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 A.16. $\Theta^*(M^*) = \Theta^*_{min}(M^*)$

Proof. Assume that there exists $A \in \Theta^*(M^*) \setminus \Theta^*_{min}(M^*)$. By applying the merging process defined in Lemma A.15 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 A.14, 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 A.17. *For $m \geq M^* + 1$, all permutation solutions are connected.*

The idea of the proof is inspired by Kim et al. (2025). We prove it using notions of group theory similarly to our other proofs (e.g. the proof for Theorem A.4 (3), the first proof for Theorem A.5).

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 A.14 and Lemma A.15, 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 A.14, $m > M^* = \sum_{i=1}^m 1(u_i\omega_i \neq 0)$ 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

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

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 A.9 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.

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

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

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

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 A.18.

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

Proof. This is a corollary from the proof of Lemma A.17. □

The second proof for Theorem A.5

Proof. (The second proof)

$$M^*=1[\mathcal{S}_{\alpha\beta}(X_S^TY)\neq0)+1[\mathcal{S}_{\alpha\beta}(X_{S^c}^TY)\neq0]=1[|\gamma_P^*|\neq0]+1[|\gamma_N^*|\neq0].$$

(1) $M^*=0$:
$M^*=0$, implies $|\gamma_P^*|=0$ and $|\gamma_N^*|=0$. For $(u_i,\omega_1)_{i=1}^m\in\Theta^*(m)$, $\sum_{i:\,u_i\geq0}u_i^2=0$ and $\sum_{i:\,u_i\leq0}u_i^2=0$, 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 $\sum_{i:\,u_i\geq0}u_i^2=|\gamma_P^*|$ and $\sum_{i:\,u_i\leq0}u_i^2=|\gamma_N^*|$ to hold.

(3) $m=M^*$:
As a set of permutations of finite number of neurons is a finite set, Lemma A.14 implies that $\Theta_{min}^*(m)$ is a finite set for all m. By this and Lemma A.16, $\Theta^*(M^*)$ is a finite set.

(4) $m>M^*$:
By Lemma A.14 and Lemma A.17, all the elements in $\Theta_{min}^*(m)$ are connected in $\Theta^*(m)$. By this and Lemmma A.15, all elemets in $\Theta^*(m)$ are connected, so $\Theta^*(m)$ is connected. □

A.3. Effects of ℓ_2-regularization for Overparameterized model

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).

Proposition A.19. *Under Assumption 4.3, for sufficiently large m,*

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

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 $m > 1$,

$$\begin{aligned} & \mathcal{S}_{\alpha\beta}(X_S^\mathsf{T} Y) \neq 0 \wedge \mathcal{S}_{\alpha\beta}(X_{S^c}^\mathsf{T} Y) \neq 0 \\ \iff & \beta < \frac{1}{\alpha} \min\{|X_S^T Y|,\ |X_{S^C}^T Y|\} \iff m^{-d} < m^{-(a_1+a_2)} C \\ \iff & -d < -a_1 - a_2 + \log_m C \iff a_1 + a_2 - d < \log_m C. \end{aligned}$$

$\log_m C \to 0$ as $m \to \infty$. Under Assumption 4.3, $a_1 + a_2 - d < \log_m C$ holds for sufficiently learge m. □

Proposition A.20. *(Proposition 4.10 in main) Under Assumption 4.3, for sufficiently large m,*

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

Proof. By the explicit form of $\Theta^*(m)$ as in (7), 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 $\mathcal{S}_{\alpha\beta}(X_S^\mathsf{T} Y)$ (resp. $\mathcal{S}_{\alpha\beta}(X_{S^c}^\mathsf{T} Y)$), we have

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

By Proposition A.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. □

Theorem A.21. *(Theorem 4.11 in main) Under Assumption 4.3, for sufficiently large m, $\Theta^*(m)$ is bounded and*

$$\begin{aligned} \forall\, \theta^* \in \Theta^*(m), \quad \|\theta^*\|_2 &= \sqrt{2\alpha \left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2} \right)} \\ &= \Theta\big(m^{\frac{a_1+a_2}{2}}\big). \end{aligned}$$

Proof. By the explicit form of $\Theta^*(m)$ in (7), $\forall\ \theta^* = (u_i, \omega_i)_{i=1}^m \in \Theta^*(m)$,

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

$\|\theta^*\|_2^2 = \Theta(m^{a_1+a_2}) \Rightarrow \|\theta^*\|_2 = \Theta(m^{\frac{a_1+a_2}{2}})$. □

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

Proposition A.22. *(Proposition 4.12 in main) Under Assumption 4.3, for sufficiently large m,*

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

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

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

□

Proposition A.23. *(Proposition 4.13 in main) For sufficiently large m,*

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

Proof. For notations, write

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

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

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

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

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

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

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

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 A.24. *(Proposition 4.14 in main) For sufficiently large m, $\varphi^*(m)$ is unbounded. Especially, $\forall a \geq \frac{a_1+a_2}{2}$, $\exists \varphi_a^*(m) \subset \varphi^*(m)$ 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 $b_1 \coloneqq \frac{X_S^\top Y}{\|X_S\|_2^2} \neq 0,\ b_2 \coloneqq \frac{X_{S^c}^\top Y}{\|X_{S^c}\|_2^2} \neq 0$ and write $c_1(m) \coloneqq \alpha b_1,\ c_2(m) \coloneqq \alpha b_2$ where $\alpha = m^{a_1+a_2}$. 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 A.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

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

We set $P = \{1\}$ and $N = \{2,\ \dots, m\}$ and denote $\mathcal{U}_{P,N}$ by $\mathcal{U}$. From the proof for Proposition A.23, $\dim\big(\varphi^*(m) \cap \mathcal{U}\big) = 2m-2$.
Pick any $a \geq \frac{a_1+a_2}{2}$. Introduce notations $\beta \coloneqq \min\{|b_1|, |b_2|\} > 0$ and $\kappa \coloneqq \sqrt{\frac{\beta}{2}}$.
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 $j = 3, \dots, m$,

$$u_j \in \Big(-\kappa\, m^{\frac{a_1+a_2-1}{2}},\ -\frac{\kappa}{2}\, m^{\frac{a_1+a_2-1}{2}}\Big), \qquad \omega_j \in \Big(-\kappa\, m^{\frac{a_1+a_2-1}{2}},\ \kappa\, m^{\frac{a_1+a_2-1}{2}}\Big),$$

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 A.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, $|\omega_1| = \left|\frac{c_1(m)}{u_1}\right| = \Theta\Big(\frac{m^{a_1+a_2}}{m^a}\Big) = \Theta\big(m^{a_1+a_2-a}\big)$. For $j \geq 3$ we have $|u_j\omega_j| \leq \kappa^2 m^{a_1+a_2-1}$, hence

$$\left|\sum_{j=3}^{m} u_j\omega_j\right| \leq (m-2)\kappa^2 m^{a_1+a_2-1} \leq \kappa^2 m^{a_1+a_2}.$$

Therefore, for all sufficiently large m,

$$\left|c_2(m) - \sum_{j=3}^{m} u_j\omega_j\right| \geq |c_2(m)| - \left|\sum_{j=3}^{m} u_j\omega_j\right| \geq |c_2(m)| - \kappa^2 m^{a_1+a_2} \geq \left(|b_2| - \kappa^2\right) m^{a_1+a_2} \geq \frac{\beta}{2} m^{a_1+a_2}.$$

Also,

$$\left|c_2(m) - \sum_{j=3}^{m} u_j\omega_j\right| \leq |c_2(m)| + \left|\sum_{j=3}^{m} u_j\omega_j\right| \leq |c_2(m)| + \kappa^2 m^{a_1+a_2} \leq \left(|b_2| + \kappa^2\right) m^{a_1+a_2} \leq \frac{3|b_2|}{2} m^{a_1+a_2}.$$

Hence, $\left|c_2(m) - \sum_{j=3}^{m} u_j\omega_j\right| = \Theta(m^{a_1+a_2})$. By using $|u_2| = \Theta(m^a)$, we get $|\omega_2| = \Theta\Big(\frac{m^{a_1+a_2}}{m^a}\Big) = \Theta\big(m^{a_1+a_2-a}\big)$. Moreover, by construction, $|u_j| = |\omega_j| = \Theta\Big(m^{\frac{a_1+a_2-1}{2}}\Big) \quad (j \geq 3)$. Hence,

$$\sum_{j=3}^{m}(u_j^2 + \omega_j^2) = O\big(2(m-2)m^{a_1+a_2-1}\big) = O\big(m^{a_1+a_2}\big).$$

Their total contribution satisfies $\left(\sum_{j=3}^{m}(u_j^2 + \omega_j^2)\right)^{1/2} = O\Big(m^{\frac{a_1+a_2}{2}}\Big)$. Since $a \geq \frac{a_1+a_2}{2}$, we have $m^{\frac{a_1+a_2}{2}} \leq m^a$, so $\left(\sum_{j=3}^{m}(u_j^2 + \omega_j^2)\right)^{1/2} = O(m^a)$. Also, $a \geq \frac{a_1+a_2}{2}$ implies $a_1 + a_2 - a \leq a$, 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. □

B. Training Dynamics and Loss Landscape

Proposition B.1. *(Proposition 5.2 in main) We denote the initialized point for model parameter as* $\theta_0 = (u_1^0, \ldots, u_m^0, \omega_1^0, \ldots, \omega_m^0)$. *Then, under Assumption 5.1,*

$$\mathbb{P}\big(\frac{\sqrt{2}}{2} m^{\frac{1-2b_1}{2}} \leq \|\theta_0\| \leq \frac{3\sqrt{2}}{2} m^{\frac{1-2b_1}{2}}\big) \geq 1 - 2\exp\left(-\frac{cm}{2\kappa^2}\right)$$

where $c > 0$ *is an absolute constant and* $\kappa = \|Z^2 - 1\|_{\psi_1}$ *is the sub-exponential norm of* $Z^2 - 1$ *where* $Z \sim N(0,1)$. *The left-hand side (LHS) tends to 1 as* m *tends to infinity, so* $\|\theta_0\|_2 = \Theta(m^{\frac{1-2b_1}{2}})$ *with high probability.*

Proof. Denote $\tau \coloneqq m^{-b_1}$. u_i's and ω_i's are sub-gaussian random variables, so u_i^2's and ω_i^2's are sub-exponential random variables. $\mathbb{E}(u_i^2) = \mathbb{E}(\omega_i^2) = \tau^2$. By the centering property, $u_i^2 - \tau^2$'s and $\omega_i^2 - \tau^2$'s are independent mean-zero sub-exponential random variables with $\|u_i^2 - \tau^2\|_{\psi_1} = \tau^2\|(u_i/\tau)^2 - 1\|_{\psi_1} = \tau^2\kappa$ where $Z \sim N(0,1)$ and $\kappa \coloneqq \|Z^2 - 1\|_{\psi_1}$ is finite.
By the Sub-exponential Bernstein inequality,

$$\forall t \geq 0,\ \mathbb{P}\big(|\|\theta\|_2^2 - 2m\tau^2| \geq t\big) = \mathbb{P}\Big(\Big|\sum_{i=1}^m (u_i^2 - \tau^2) + \sum_{i=1}^m (\omega_i - \tau^2)\Big| \geq t\Big) \leq 2\exp\Big[-c\min\{\frac{t^2}{2m\tau^4\kappa^2},\ \frac{t}{\tau^2\kappa}\}\Big]$$

where $c > 0$ is an absolute constant.
By substiuting back $\tau = m^{-b_1}$,

$$\forall t \geq 0,\ \mathbb{P}\big(|\|\theta\|_2^2 - 2m^{1-2b_1}| \geq t\big) \leq 2\exp\Big[-c\min\{\frac{t^2 m^{4b_1-1}}{2\kappa^2},\ \frac{tm^{2b_1}}{\kappa}\}\Big] \tag{28}$$

For any $s \geq 0$,

$$\begin{aligned}
&\mathbb{P}\big(|\|\theta\|_2 - \sqrt{2}m^{\frac{1-2b_1}{2}}| > s\big) \\
=&\mathbb{P}\big(\sqrt{2}m^{\frac{1-2b_1}{2}}|\|\theta\|_2 - \sqrt{2}m^{\frac{1-2b_1}{2}}| > \sqrt{2}m^{\frac{1-2b_1}{2}}s\big) \\
\leq&\mathbb{P}\big(|\|\theta\|_2 + \sqrt{2}m^{\frac{1-2b_1}{2}}|\|\theta\|_2 - \sqrt{2}m^{\frac{1-2b_1}{2}}| > \sqrt{2}m^{\frac{1-2b_1}{2}}s\big) \\
=&\mathbb{P}\big(|\|\theta\|_2^2 - 2m^{1-2b_1}| > \sqrt{2}m^{\frac{1-2b_1}{2}}s\big).
\end{aligned} \tag{29}$$

By substituting $t = \sqrt{2}m^{\frac{1-2b_1}{2}}s$ into (28),

$$\mathbb{P}\big(|\|\theta\|_2^2 - 2m^{1-2b_1}| > \sqrt{2}m^{\frac{1-2b_1}{2}}s\big) \leq 2\exp\Big[-c\min\{\frac{m^{2b_1}s^2}{\kappa^2},\ \frac{\sqrt{2}m^{1/2+b_1}s}{\kappa}\}\Big].$$

Hence, by comparing this inequality with (29),

$$\forall s \geq 0,\ \mathbb{P}\big(|\|\theta\|_2 - \sqrt{2}m^{\frac{1-2b_1}{2}}| > s\big) \leq 2\exp\Big[-c\min\{\frac{m^{2b_1}s^2}{\kappa^2},\ \frac{\sqrt{2}m^{1/2+b_1}s}{\kappa}\}\Big].$$

By setting $s = \frac{\sqrt{2}}{2}m^{\frac{1-2b_1}{2}} > 0$,

$$\begin{aligned}
\mathbb{P}\big(|\|\theta\|_2 - \sqrt{2}m^{\frac{1-2b_1}{2}}| > \frac{\sqrt{2}}{2}m^{\frac{1-2b_1}{2}}\big) &\leq 2\exp\Big[-c\min\{\frac{m}{2\kappa^2},\ \frac{m}{\kappa}\}\Big] \\
&\leq 2\exp\Big(-\frac{cm}{2\kappa^2}\Big) \quad \text{(because } \kappa > 1/2\text{)}.
\end{aligned}$$

Hence,

$$\mathbb{P}\big(\frac{\sqrt{2}}{2}m^{\frac{1-2b_1}{2}} \leq \|\theta\| \leq \frac{3\sqrt{2}}{2}m^{\frac{1-2b_1}{2}}\big) = 1 - \mathbb{P}\big(|\|\theta\|_2 - \sqrt{2}m^{\frac{1-2b_1}{2}}| > \frac{\sqrt{2}}{2}m^{\frac{1-2b_1}{2}}\big) \geq 1 - 2\exp\Big(-\frac{cm}{2\kappa^2}\Big).$$

□

Theorem B.2. *(Theorem 5.3 in main) Assume Assumptions 4.3 and 5.1 hold and that m is sufficiently large. With probability at least $1 - 2\exp(-Cm)$ (C is a positive constant independent of m),*

$$\frac{\|\theta_0\|_2}{\|\theta^*\|_2} = \Theta\big(m^{\frac{1-a_1-a_2-2b_1}{2}}\big).$$

The behavior of the value $\frac{\|\theta_0\|_2}{\|\theta^*\|_2}$ when overparameterizing (i.e. increasing m) depends on the sign of $1 - a_1 - a_2 - 2b_1$.

- *If* $a_1 + a_2 + 2b_1 < 1$, $\frac{\|\theta_0\|_2}{\|\theta^*\|_2} \to \infty$ *with probability one as* $m \to \infty$.
- *If* $a_1 + a_2 + 2b_1 = 1$, $\frac{\|\theta_0\|_2}{\|\theta^*\|_2} = \Theta(1)$ *and is away from 0 with probability one as* $m \to \infty$.
- *If* $a_1 + a_2 + 2b_1 > 1$, $\frac{\|\theta_0\|_2}{\|\theta^*\|_2} \to 0$ *with probability one as* $m \to \infty$.

Proof. Under our Assumptions, $\|\theta^*\|_2 = \Theta(m^{\frac{a_1+a_2}{2}})$ by Theorem A.21, and $\|\theta_0\|_2 = \Theta(m^{\frac{1-2b_1}{2}})$ with probability at least $1 - 2\exp(-Cm)$ ($C = \frac{c}{2\kappa^2}$ is a positive constant independent of m) by Proposition B.1. They imply that, $\exists k_1, k_2, m_0$ that are independent of m such that $\forall m \geq m_0$, the event

$$E_m = \{k_1 m^{\frac{1-a_1-a_2-2b_1}{2}} \leq \frac{\|\theta_0\|_2}{\|\theta^*\|_2} \leq k_2 m^{\frac{1-a_1-a_2-2b_1}{2}}\}$$

satisfies $\mathbb{P}(E_m) \geq 1 - 2\exp(-Cm)$. This proves the first part.
Hence, $\sum_{m=m_0}^{\infty} \mathbb{P}(E_m^c) \leq \sum_{m=m_0}^{\infty} 2\exp(-Cm) < \infty$. By Borel-Cantelli lemma, $\mathbb{P}(E_m^c \text{ happens infinitely often}) = 0$. Thus, $\mathbb{P}(E_m \text{ happens eventually}) = 1$. Also, we can take

$$\begin{aligned} k_1 &= \frac{\sqrt{2}}{2\sqrt{2\left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2}\right)}} \\ &= \frac{1}{2\sqrt{\left(\frac{|\mathcal{S}_{\alpha\beta}(X_S^\top Y)|}{\|X_S\|_2^2} + \frac{|\mathcal{S}_{\alpha\beta}(X_{S^c}^\top Y)|}{\|X_{S^c}\|_2^2}\right)}} \\ &> 0. \end{aligned}$$

k_1 is found by the proofs of Theorem A.21 and Proposition B.1. This proves the latter part.

□

Theorem B.3. *(Theorem 5.5 in main) Under Assumptions 4.3 and 5.1, for sufficiently large* m*, with probability at least* $1 - 8\exp(-Cm)$*,*

$$\exists\, \phi^* \in \varphi^*(m)\ s.t.\ \|\theta_0 - \phi^*\|_2^2 = O(m^{2(a_1+a_2)-1+2b_1}),$$

where $C > 0$ *is an absolute constant.*

Proof. For notations, write

$$B_+ \coloneqq \frac{\alpha X_S^\top Y}{\|X_S\|_2^2}, \qquad B_- \coloneqq \frac{\alpha X_{S^c}^\top Y}{\|X_{S^c}\|_2^2}.$$

Let the random index partition induced by the initialization be

$$P \coloneqq \{j \in [m] : u_j^0 \geq 0\}, \qquad N \coloneqq \{j \in [m] : u_j^0 \leq 0\}.$$

As $u_j \sim N(0, \tau_1^2)$ (a continuous distribution), $\mathbb{P}(u_j^0 = 0) = 0$, hence $P \cup N = [m]$ and $P \cap N = \emptyset$ almost surely. Introduce the partial sums and partial norms

$$S_+^0 \coloneqq \sum_{j\in P} u_j^0 \omega_j^0, \qquad S_-^0 \coloneqq \sum_{j\in N} u_j^0 \omega_j^0,$$

$$U_+^0 \coloneqq \Big(\sum_{j\in P} (u_j^0)^2\Big)^{1/2}, \qquad U_-^0 \coloneqq \Big(\sum_{j\in N} (u_j^0)^2\Big)^{1/2}.$$

$(u_j^0)_{j=1}^m$ are i.i.d. mean-zero sub-Gaussian with $\|(u_j^0)\|_{\psi_2} \lesssim \tau_1$. Then, the nonnegative random variables

$$X_j \coloneqq (u_j^0)^2 \mathbf{1}\{u_j^0 \geq 0\}, \qquad Y_j \coloneqq (u_j^0)^2 \mathbf{1}\{u_j^0 \leq 0\}$$

are i.i.d. sub-exponential with

$$\|X_j\|_{\psi_1} = \|(u_j^0)^2 1[u_j^0 \geq 0]\|_{\psi_1} \leq \|(u_j^0)^2\|_{\psi_1} = \|(u_j^0)\|_{\psi_2}^2 \lesssim \tau_1^2, \quad \|Y_j\|_{\psi_1} = \lesssim \tau_1^2.$$

By symmetry, $\mathbb{E}[X_j] = \mathbb{E}[Y_j] = \frac{1}{2}\mathbb{E}[(u_j^0)^2] = \frac{1}{2}\tau_1^2$. By the Sub-exponential Bernstein inequality,

$$\forall t \geq 0,\ \mathbb{P}\Big(\left| \sum_{i=1}^m X_j - \frac{m}{2}\tau_1^2 \right| \geq t \Big) = \mathbb{P}\Big(\left| \sum_{i=1}^m (X_j - \frac{1}{2}\tau_1^2) \right| \geq t \Big) \leq 2\exp\left[-c\min\{ \frac{t^2}{m\|X_1\|_{\psi_1}^2},\ \frac{t}{\|X_1\|_{\psi_1}} \} \right] \tag{30}$$

where $c > 0$ is an absolute constant.
As $\tau_1 = m^{-b_1}$,

$$\forall t \geq 0,\ \mathbb{P}\Big(\left| \sum_{i=1}^m X_j - \frac{m^{1-2b_1}}{2} \right| \geq t \Big) \leq 2\exp\left[-c\min\{ \frac{t^2}{m\|X_1\|_{\psi_1}^2},\ \frac{t}{\|X_1\|_{\psi_1}} \} \right].$$

$$\|X_1\|_{\psi_1} \lesssim \tau_1^2 = m^{-2b_1} \Rightarrow \begin{cases} \frac{t^2}{m\|X_1\|_{\psi_1}^2} \gtrsim t^2 m^{4b_1 - 1}, \\ \frac{t}{\|X_1\|_{\psi_1}} \gtrsim t m^{2b_1}. \end{cases}$$

By taking $t = \frac{m^{1-2b_1}}{4}$, $t^2 m^{4b_1-1} \gtrsim m$ and $t m^{2b_1} \gtrsim m$. Hence, for $t = \frac{m^{1-2b_1}}{4}$,

$$\begin{aligned}
&\min\{ \frac{t^2}{m\|X_1\|_{\psi_1}^2},\ \frac{t}{\|X_1\|_{\psi_1}} \} \gtrsim m \\
\Rightarrow &- c\min\{ \frac{t^2}{m\|X_1\|_{\psi_1}^2},\ \frac{t}{\|X_1\|_{\psi_1}} \} \lesssim -m \\
\Rightarrow &\exp\left[-c\min\{ \frac{t^2}{m\|X_1\|_{\psi_1}^2},\ \frac{t}{\|X_1\|_{\psi_1}} \} \right] \leq \exp(-Cm) \\
\Rightarrow &\mathbb{P}\Big(\left| \sum_{i=1}^m X_j - \frac{m^{1-2b_1}}{2} \right| \geq t \Big) \leq 2\exp(-Cm) \qquad \text{(by (30))}
\end{aligned}$$

where $C > 0$ is an absolute constant. By substituting $t = \frac{m^{1-2b_1}}{4}$,

$$\mathbb{P}\Big(\left| \sum_{i=1}^m X_j - \frac{m^{1-2b_1}}{2} \right| \geq \frac{m^{1-2b_1}}{4} \Big) \leq 2\exp(-Cm)$$

In the same way, $\mathbb{P}\Big(\left| \sum_{i=1}^m Y_j - \frac{m^{1-2b_1}}{2} \right| \geq \frac{m^{1-2b_1}}{4} \Big) \leq 2\exp(-Cm)$.
As $\sum_{i=1}^m X_j = \sum_{j\in P}(u_j^0)^2$ and $\sum_{i=1}^m Y_j = \sum_{j\in N}(u_j^0)^2$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{m^{1-2b_1}}{4} \leq \sum_{j\in P}(u_j^0)^2,\ \sum_{j\in N}(u_j^0)^2 \leq \frac{3m^{1-2b_1}}{4} \right) \\
=&1 - \mathbb{P}\left(\left| \sum_{j\in P}(u_j^0)^2 - \frac{m^{1-2b_1}}{2} \right| \geq \frac{m^{1-2b_1}}{4} \ \vee \ \left| \sum_{j\in N}(u_j^0)^2 - \frac{m^{1-2b_1}}{2} \right| \geq \frac{m^{1-2b_1}}{4} \right) \\
\geq&1 - \mathbb{P}\left(\left| \sum_{j\in P}(u_j^0)^2 - \frac{m^{1-2b_1}}{2} \right| \geq \frac{m^{1-2b_1}}{4} \right) - \mathbb{P}\left(\left| \sum_{j\in N}(u_j^0)^2 - \frac{m^{1-2b_1}}{2} \right| \geq \frac{m^{1-2b_1}}{4} \right) \\
\geq&1 - 4\exp(-Cm).
\end{aligned}$$

Hence, there exist absolute constants $c_1, C_1 > 0$ ($c_1 = \frac{1}{4}, C_1 = \frac{3}{4}$) such that with probability at least $1 - 4\exp(-Cm)$,

$$c_1\, m^{1-2b_1} \le \sum_{j\in P}(u_j^0)^2,\ \sum_{j\in N}(u_j^0)^2 \le C_1\, m^{1-2b_1} \tag{31}$$

Taking square roots and using $\tau_1 = m^{-b_1}$ yields: with the same probability,

$$c_2\, m^{\frac{1-2b_1}{2}} \le U_+^0,\ U_-^0 \le C_2\, m^{\frac{1-2b_1}{2}}, \tag{32}$$

for some absolute constants $c_2, C_2 > 0$.
$(\omega_j^0)_{j=1}^m$ are i.i.d. mean-zero sub-Gaussian with $\|(\omega_j^0)\|_{\psi_2} \lesssim \tau_2 = \tau_1$ (under Assumption 5.1). For each j, u_j and ω_j are independent, so $\mathbb{E}[u_j^0\omega_j^0] = 0$. Hence, $u_j^0\omega_j^0$ are i.i.d. mean-zero sub-exponential with

$$\|u_j^0\omega_j^0\|_{\psi_1} \ \le\ \|u_j^0\|_{\psi_2}\|\omega_j^0\|_{\psi_2} \lesssim \tau_1^2.$$

Similarly to the case for $\sum_{i=1}^m X_j$ and $\sum_{i=1}^m Y_j$, by the Sub-exponential Bernstein inequality,

$$\forall t \ge 0,\ \mathbb{P}\Big(\Big|\sum_{j\in P} u_j^0\omega_j\Big| \ge t\Big) \le 2\exp\left[-c\min\{\frac{t^2}{m\|u_1\omega_1\|_{\psi_1}^2},\ \frac{t}{\|u_1\omega_1\|_{\psi_1}}\}\right]$$

where $c > 0$ is an absolute constant.
By substituting $t = \frac{m^{1-2b_1}}{4}$,

$$\mathbb{P}\Big(\,|S_+^0| \ge \frac{m^{1-2b_1}}{4}\Big) = \mathbb{P}\Big(\Big|\sum_{j\in P} u_j^0\omega_j^0\Big| \ge \frac{m^{1-2b_1}}{4}\Big) \le 2\exp\left(-C'm\right)$$
$$\mathbb{P}\Big(\,|S_-^0| \ge \frac{m^{1-2b_1}}{4}\Big) = \mathbb{P}\Big(\Big|\sum_{j\in N} u_j^0\omega_j^0\Big| \ge \frac{m^{1-2b_1}}{4}\Big) \le 2\exp\left(-C'm\right)$$

Thus,

$$\begin{aligned}
&\mathbb{P}\Big(\,|S_+^0| + |S_-^0| \le \frac{m^{1-2b_1}}{2}\Big)\\
\ge&\mathbb{P}\Big(\,|S_+^0| \le \frac{m^{1-2b_1}}{4}\ \wedge\ |S_-^0| \le \frac{m^{1-2b_1}}{4}\Big)\\
=&1 - \mathbb{P}\Big(\,|S_+^0| \ge \frac{m^{1-2b_1}}{4}\ \vee\ |S_-^0| \ge \frac{m^{1-2b_1}}{4}\Big)\\
\ge&1 - \mathbb{P}\Big(\,|S_+^0| \ge \frac{m^{1-2b_1}}{4}\Big) - \mathbb{P}\left(|S_-^0| \ge \frac{m^{1-2b_1}}{4}\right)\\
\ge&1 - 4\exp(-C'm).
\end{aligned}$$

Hence, there exists absolute constant $C_3 > 0$ such that with probability at least $1 - 4\exp(-C'm)$,

$$|S_+^0| + |S_-^0| \le C_3\, m^{1-2b_1}. \tag{33}$$

Define u^* by $u_j^* \coloneqq u_j^0$ for all j, so the sign partition induced by u^* is exactly (P, N). Introduce notations $\lambda_+ \coloneqq \frac{B_+ - S_+^0}{(U_+^0)^2}$ and $\lambda_- \coloneqq \frac{B_- - S_-^0}{(U_-^0)^2}$, and define ω^* coordinatewise by

$$\omega_j^* \coloneqq \begin{cases}\omega_j^0 + \lambda_+ u_j^0, & j \in P,\\ \omega_j^0 + \lambda_- u_j^0, & j \in N.\end{cases}$$

Then, $\sum_{j:\, u_j^* \geq 0} u_j^* \omega_j^* = \sum_{j \in P} u_j^0 (\omega_j^0 + \lambda_+ u_j^0) = S_+^0 + \lambda_+ (U_+^0)^2 = B_+$. Similarly, $\sum_{j:\, u_j^* \leq 0} u_j^* \omega_j^* = B_-$. Therefore, by Theorem 4.1, $\phi^* := (u^*, \omega^*) \in \varphi^*(m)$.
Since $u^* = u^0$, we have

$$\begin{aligned}\|\theta_0 - \phi^*\|_2^2 &= \sum_{j=1}^{m} (\omega_j^* - \omega_j^0)^2 = \sum_{j \in P} (\lambda_+ u_j^0)^2 + \sum_{j \in N} (\lambda_- u_j^0)^2 \\ &= \lambda_+^2 (U_+^0)^2 + \lambda_-^2 (U_-^0)^2 = \frac{(B_+ - S_+^0)^2}{(U_+^0)^2} + \frac{(B_- - S_-^0)^2}{(U_-^0)^2}. \end{aligned} \tag{34}$$

Define an event $\mathcal{E} = \{c_2\, m^{\frac{1-2b_1}{2}} \leq U_+^0,\ U_-^0 \leq C_2\, m^{\frac{1-2b_1}{2}}\} \cap \{|S_+^0| + |S_-^0| \leq C_3\, m^{1-2b_1}\}$.

$$\begin{aligned}\mathbb{P}(\mathcal{E}) =& 1 - \mathbb{P}(\{c_2\, m^{\frac{1-2b_1}{2}} \leq U_+^0,\ U_-^0 \leq C_2\, m^{\frac{1-2b_1}{2}}\}^C \vee \{|S_+^0| + |S_-^0| \leq C_3\, m^{1-2b_1}\}^C) \\ \geq& 1 - \mathbb{P}(\{c_2\, m^{\frac{1-2b_1}{2}} \leq U_+^0,\ U_-^0 \leq C_2\, m^{\frac{1-2b_1}{2}}\}^C) - \mathbb{P}(\{|S_+^0| + |S_-^0| \leq C_3\, m^{1-2b_1}\}^C) \\ \geq& 1 - 4\exp(-Cm) - 4\exp(-C'm) \\ \geq& 1 - 8\exp(-C''m)\end{aligned}$$

where $C'' = \max\{C, C''\} > 0$ is an absolute constant. On $\mathcal{E}$, the denominators satisfy $(U_\pm^0)^2 = \Theta(m^{1-2b_1})$. For the numerators, $|B_+| + |B_-| = \alpha \left(\left| \frac{X_S^\top Y}{\|X_S\|_2^2} \right| + \left| \frac{X_S^\top Y}{\|X_S\|_2^2} \right| \right) = \Theta(\alpha) = \Theta(m^{a_1 + a_2})$
and $|S_+^0| + |S_-^0| \leq C_3 m^{1-2b_1}$. Therefore, on $\mathcal{E}$,

$$|B_+ - S_+^0| + |B_- - S_-^0| \leq (|B_+| + |B_-|) + (|S_+^0| + |S_-^0|) \leq C_4 m^{a_1+a_2} + C_3 m^{1-2b_1}.$$

for some absolute constants $C_3, C_4 > 0$. Plugging into (34) and using the lower bound on $(U_\pm^0)^2$ gives

$$\begin{aligned}\|\theta_0 - \phi^*\|_2^2 &\leq \frac{C_5}{m^{1-2b_1}} \left\{ (B_+ - S_+^0)^2 + (B_- - S_-^0)^2 \right\} \\ &\leq \frac{C_5}{m^{1-2b_1}} \left(|B_+ - S_+^0| + |B_- - S_-^0| \right)^2 \\ &\leq \frac{C_5}{m^{1-2b_1}} \left(C_4 m^{a_1+a_2} + C_3 m^{1-2b_1} \right)^2 \\ &\leq C_6 m^{2(a_1+a_2)-1+2b_1} + C_7 m^{1-2b_1} + C_8 m^{a_1+a_2} \\ &= O(m^{\max\{2(a_1+a_2)-1+2b_1,\ 1-2b_1,\ a_1+a_2\}}) \\ &= O(m^{2(a_1+a_2)-1+2b_1}).\end{aligned}$$

Thus, with probability at least $1 - 8\exp(-C''m)$, $\|\theta_0 - \phi^*\|_2^2 = O(m^{2(a_1+a_2)-1+2b_1})$.
$\|\theta_0 - \phi^*\|_2^2 = O(m^{2(a_1+a_2)-1+2b_1})$ implies that, for $a_1 + a_2 + b_1 \leq \frac{1}{2}$, $\|\theta_0 - \phi^*\|_2^2 = O(1)$. □

C. Visualizations

We provide visualizations of $\Theta^*(m)$ for several dataset examples.

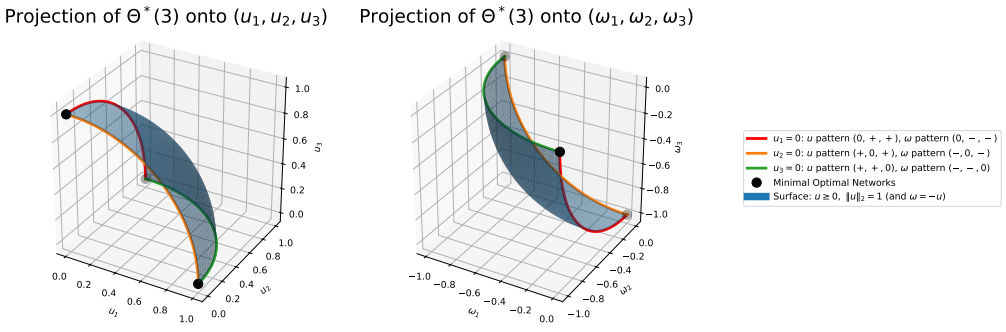


Figure 4. The projections of $\Theta^*(3)$ onto u and ω coordinates for $X = [1, -1]^T$, $Y = [-\frac{4}{9}, \frac{1}{18}]^T$, $\alpha = 3^1$, $\beta = 3^{-3}$ ($\gamma_P^* = -1, \gamma_N^* = 0$). The black dots correspond to Minimal Optimal Solutions defined in Eq.8. Both the surface and the colored boundaries are globally optimal parameters. For each u, corresponding ω has the same color in ω coordinate.

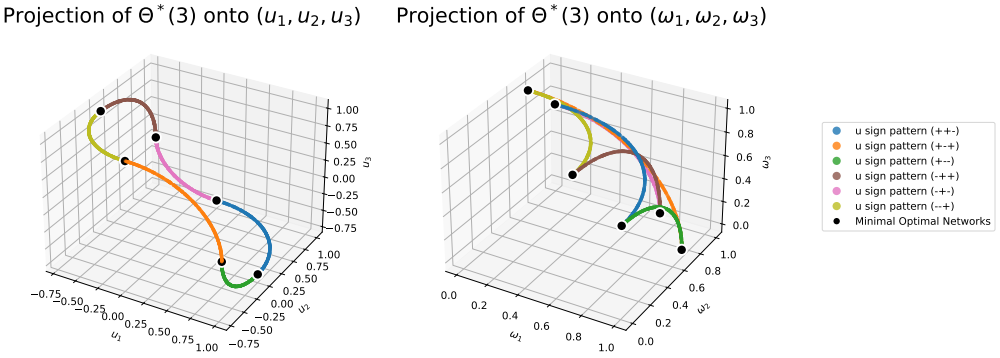


Figure 5. The projections of $\Theta^*(3)$ onto u and ω coordinates for $X = [1, -1]^T$, $Y = [\frac{2}{3}, \frac{1}{2}]^T$, $\alpha = 3^1$, $\beta = 3^{-2}$ ($\gamma_P^* = 1, \gamma_N^* = -0.5$). The black dots correspond to Minimal Optimal Solutions defined in Eq.8. For each u, corresponding ω has the same color in ω coordinate.

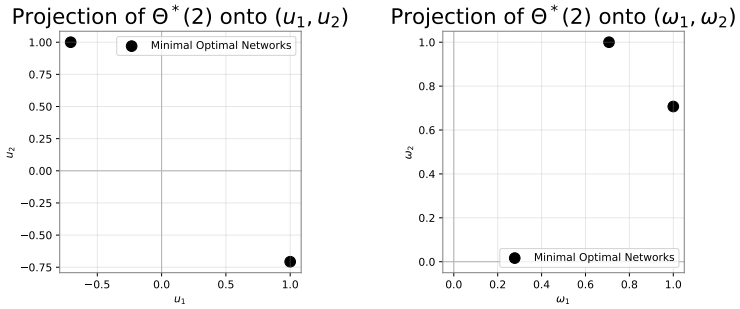


Figure 6. The projections of $\Theta^*(2)$ onto u and ω coordinates for $X = [1, -1]^T$, $Y = [1, \frac{4}{3}]^T$, $\alpha = 2^1$, $\beta = 2^{-2}$ ($\gamma_P^* = 1, \gamma_N^* = -0.5$). The black dots correspond to Minimal Optimal Solutions defined in Eq.8. As $m = M^*$, $\Theta^*(m)$ is a finite set.