Overparameterized ReLU Neural Networks Learn the Simplest Model: Neural Isometry and Phase Transitions

Yifei Wang[1], Yixuan Hua[2], Emmanuel J. Candès[3], Mert Pilanci[1] *Member, IEEE*

Abstract

The practice of deep learning has shown that neural networks generalize remarkably well even with an extreme number of learned parameters. This appears to contradict traditional statistical wisdom, in which a trade-off between model complexity and fit to the data is essential. We aim to address this discrepancy by adopting a convex optimization and sparse recovery perspective. We consider the training and generalization properties of two-layer ReLU networks with standard weight decay regularization. Under certain regularity assumptions on the data, we show that ReLU networks with an arbitrary number of parameters learn only simple models that explain the data. This is analogous to the recovery of the sparsest linear model in compressed sensing. For ReLU networks and their variants with skip connections or normalization layers, we present isometry conditions that ensure the exact recovery of planted neurons. For randomly generated data, we show the existence of a phase transition in recovering planted neural network models, which is easy to describe: whenever the ratio between the number of samples and the dimension exceeds a numerical threshold, the recovery succeeds with high probability; otherwise, it fails with high probability. Surprisingly, ReLU networks learn simple and sparse models that generalize well even when the labels are noisy. The phase transition phenomenon is confirmed through numerical experiments.

Index Terms

neural networks, deep learning, convex optimization, sparse recovery, compressed sensing, ℓ_1 minimization, Lasso.

I. Introduction

Recent work has shown that neural networks (NNs) exhibit extraordinary generalization abilities in many machine learning tasks. Although NNs employed in practice are often over-parameterized, meaning the number of parameters exceeds the sample size, they generalize to unseen data and perform well. In this work, we study the problem of generalization in such over-parameterized models from a convex optimization and sparse recovery perspective. We are interested in the setting where an NN with an arbitrary number of neurons is trained on datasets with a simple structure, for instance, when the label vector is the output of a simple NN, which may have additive noise. A natural question arises: under which conditions, a neural network with an arbitrary number of neurons, in which the number of trainable parameters can be quite large, achieves perfect generalization? We uncover a sharp phase transition in the behavior of NNs in the recovery of simple planted models. Our results imply that the weight decay regularization solely controls whether the NN recovers the underlying simple model planted in the data or fails by overfitting a more complex model, regardless of the number of parameters in the NN.

Our findings are close in spirit to classical results on sparse recovery and compressed sensing. It is known that there exists a sharp phase transition in recovering a sparse planted vector from random linear measurements. Specifically, the probability of successful recovery will be close to one when the sample number exceeds a certain threshold. Otherwise, the probability of recovery is close to zero. This can be shown by analyzing the intersection probability of a convex cone with a random subspace, which undergoes a sharp phase transition as the statistical dimension of the cone changes with respect to the ambient dimension [1]. By leveraging recently discovered connections between ReLU NNs and Group Lasso models [2, 3, 4, 5, 6], we show that a calculation involving statistical dimensions of convex cones implies a phase transition in two-layer ReLU NNs for recovering simple planted models.

In addition, we consider deterministic training data and derive analogs of the *irrepresentability condition* [7] and *Restricted Isometry Property* [8, 9], which play an important role in the recovery of sparse linear models. We develop the notion of *Neural Isometry Conditions* to characterize non-random training data that allow exact recovery of planted neurons. We further show that random i.i.d. Gaussian, sub-Gaussian, and Haar distributed random matrices satisfy Neural Isometry Conditions with high probability when the number of samples is sufficiently high. The random Gaussian data assumption is widely used in practice. For instance, the subproblem in training diffusion models [10, 11] is equivalent to learn the function over Gaussian random data.

[1] Yifei Wang and Mert Pilanci are with the Department of Electrical Engineering, Stanford University, Stanford, CA USA. E-mail: {wangyf18,pilanci}@stanford.edu

[2] Yixuan Hua is with the Department of Operations Research & Financial Engineering, Princeton University, Princeton, NJ USA.

[3] Emmanuel J. Candès is with the Department of Statistics and Department of Mathematics, Stanford University, Stanford, CA USA.

Model	Data	Success		Failure	Results
		strong recovery	weak recovery		
linear	Haar	$n > 7.613d$	$n > 2d$	$n < 2d$	Thm. 7 Thm. 9
	Gaussian	$n > O(d\log(n))$	$n > 2d$	$n < 2d$	Thm. 8, Thm. 5
ReLU-normalized	Gaussian	-	$n > 2d$	-	Thm. 11
ReLU	Gaussian	-	$n \to \infty$	-	Prop. 9
two ReLUs-normalized	Gaussian	-	$n \to \infty$	-	Prop. 8
k ReLUs	-	-	Neural Isometry Condition (NIC-k)	-	Prop. 4
k ReLUs-normalized	-	-	-	-	Prop. 5
noisy linear model	sub-Gaussian	-	$n > O(d\log(n))$	-	Thm. 10

TABLE I: Summary of our results for the success or failure of the recovery of planted neurons. Gaussian data refers to training matrices with i.i.d. standard Gaussian entries, Haar data refers to uniformly drawn training matrices from the set of orthogonal matrices. Strong recovery refers to the recovery of all possible planted neurons, whereas weak recovery refers to the recovery of fixed neurons.

Although neural networks lead to non-convex optimization problems which are challenging to analyze, a recent line of work [2, 3, 4] showed that regularized training problems of multilayer ReLU networks can be reformulated as convex programs. Based on the convex optimization formulations, [6] further gives the exact characterizations of all global optima in two-layer ReLU networks. More precisely, it was shown that all globally optimal solutions of the nonconvex training problem are given by the solution set of a simple convex program up to permutation and splitting. In other words, we can find the set of optimal NNs for the regularized training problem by solving a convex optimization problem. The convex optimization formulations of NNs were also extended to NNs with batch normalization layers [12], convolutional NNs (CNNs) [13, 4], polynomial activation networks [14], transformers with self-attention layers [15] and Generative Adversarial Networks (GANs) [16].

We study the recovery properties of optimal two-layer ReLU neural networks by considering their equivalent convex formulations and leveraging connections to sparse recovery and compressed sensing. We also consider variants of the basic two-layer architecture with skip connections and normalization layers, which are basic building blocks of modern DNNs such as ResNets [17]. We show the existence of a sharp phase transition in the recovery of simple models via ReLU NNs. To be more specific, for a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, there exists a critical threshold for the ratio n/d, above which a planted network with few neurons will be the unique optimal solution of the convex program of two-layer networks with probability close to 1, as long as n and d are moderately high. Otherwise, this probability will be close to 0. The same conclusion applies to the non-convex training of the two-layer network with any number of neurons, up to permutation and neuron splitting (see Appendix B). Moreover, our results highlight the importance of skip-connections and normalization layers in the success of recovery as we show in Section III-C. We also provide deterministic isometry conditions that guarantee the recovery of an arbitrary number of ReLU neurons. We summarize these results in Table I.

A. Related works

Recent works have investigated linearized neural network models trained with gradient descent from a kernel-based learning perspective [18, 19, 20, 21]. When the width of the neural network approaches infinity, it is known that NNs can fit all the training data [22, 23, 24]. However, in this regime, the analysis shows that almost no hidden neurons move from their initial values to learn features [25]. Further experiments also confirm that infinite width limits and linearized kernel approximations are insufficient to give an adequate explanation for the success of non-convex neural network models as their width goes to infinity [26]. Due to the non-linear structure of neural networks and non-convexity of the training problem, only a few works consider the role of over-parameterization when the width of the neural network is finite [27, 28, 21].

It was conjectured that models trained with simple iterative methods such as Stochastic Gradient Descent (SGD) approach flat local minima [29, 30, 31] that generalize well. However, it was also shown that the behavior of the trained model heavily depends on the choice of the specific optimization algorithm and its hyper-parameters [32, 33, 30, 34].

Developing algorithms for the recovery of planted two-layer neural networks has been studied in the literature. In [35, 36], the authors design spectral methods for learning the weight matrices of a planted two-layer ReLU network with k neurons. In contrast, our work studies the recovery properties of over-parameterized NNs with arbitrarily many neurons that minimize the training objective and shed light on the optimization landscape. [37, 38, 39, 40] analyze the recovery of two-layer ReLU neural networks using the gradient descent method, while [41] extends the analysis to other activation functions, including leaky ReLU. In comparison, our results apply to the case where neural networks have many more neurons than the planted model.

B. Notation

We introduce notations used throughout the paper. We use the notation $[n]$ to represent the set $\{1,\ldots,n\}$. We use the notation $\mathbb{I}(\cdot)$ for the 0-1 valued indicator function which takes the value 1 when its argument is a true logical statement and 0 otherwise. We reserve boldcase lower-case letters for vectors, boldcase upper-case letters for matrices and plain lower-case letters for scalars. For a vector $\mathbf{w}\in\mathbb{R}^d$, we use $\|\mathbf{w}\|_p := \left(\sum_{i=1}^d |w_i|^p\right)^{1/p}$ to represent its ℓ_p norm. The notation $\cos\angle(\mathbf{w},\mathbf{v}) := \frac{\mathbf{w}^T\mathbf{v}}{\|\mathbf{w}\|_2\|\mathbf{v}\|_2}\in[-1,1]$ represents the cosine angle between two vectors $\mathbf{w}\in\mathbb{R}^d$, and $\mathbf{v}\in\mathbb{R}^d$. For a matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$, we use $\|\mathbf{X}\|_2$ to denote its operator norm. We use $\|\mathbf{X}\|_{\infty,\infty}$ for the matrix infinity norm, i.e., the elementwise maximum absolute value. We use $\mathbf{I}_n\in\mathbb{R}^{n\times n}$ to denote the identity matrix with size $n\times n$. We denote $\lambda_{\max}(\mathbf{X})$ for the maximum eigenvalue of a symmetric matrix $\mathbf{X}$. Similarly, the notion $\mathrm{eigmax}(\mathbf{X})$ represents the subspace spanned by the eigenvectors corresponding to the maximum eigenvalue for a symmetric matrix $\mathbf{X}$. We use the shorthand $\mathbf{diag}(x_1,...,x_n)$ to represent a diagonal matrix with entries $x_1,...,x_n$ on the diagonal. The notation $\mathbf{x}_i^{\text{row}}\in\mathbb{R}^d$ and $\mathbf{x}_j^{\text{col}}\in\mathbb{R}^n$ denoted the i-th row and j-th column of an $n\times d$ matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$ respectively. We use the notation $\mathrm{poly}(n)$ to denote a polynomial function of the variable n. We call a zero-mean random variable X sub-Gaussian with variance proxy $\sigma^2>0$ if it holds that $\mathbb{E}[e^{sX}\le e^{\frac{1}{2}\sigma^2 s^2}]$ for any $s\in\mathbb{R}$.

C. Organization

In Section II, we present a preview of our results for the special case of linear neuron recovery with ReLU NNs. We introduce variants of the ReLU NN architecture in Section III. In Section IV we develop deterministic conditions, termed **Neural Isometry Conditions**, for the recovery of linear and non-linear neurons using different NN architectures. In Section V, we investigate random ensembles of training data matrices, for which we show the existence of a sharp phase transition in satisfying these deterministic conditions via non-asymptotic probabilistic bounds. In Section VI, we develop asymptotic results on the recovery probability for the case of multiple ReLU neurons. We present numerical simulations to corroborate our theoretical results in Section VII. We present our conclusions in Section VIII.

II. A preview: an exact characterization of linear neuron recovery

In this section, we present a preview of our results on two-layer ReLU networks with skip connection in the special case of linear neuron recovery. We generalize our result to the recovery of nonlinear neurons using different architectures later in Sections IV-VI. We invite the reader to refer to Appendix A for the background on the isometry conditions in compressed sensing, which share important parallels with our analysis. We start with the following two-layer ReLU network model with a linear skip connection.

$$f(\mathbf{X};\Theta) = \mathbf{X}\mathbf{w}_1 v_1 + \sum_{i=2}^{m}(\mathbf{X}\mathbf{w}_i)_+ v_i,$$

where $\Theta=\{\mathbf{W},\mathbf{v}\}$ denotes trainable parameters including first layer weights $\mathbf{W}\in\mathbb{R}^{d\times m}$ and second layer weights $\mathbf{v}\in\mathbb{R}^m$. We will first consider the minimum norm interpolation problem

$$\min_{\Theta}\underbrace{\|\mathbf{W}\|_F^2+\|\mathbf{v}\|_2^2}_{\|\Theta\|_F^2},\ \text{s.t. } f(\mathbf{X};\Theta)=\mathbf{y}. \tag{1}$$

where the objective $\|\Theta\|_F^2$ stands for the weight-decay regularization term.

A. Hyperplane Arrangements

We now introduce an important concept from combinatorial geometry called hyperplane arrangement patterns, in order to introduce convex optimization formulations of ReLU network training problems.

Definition 1 (Diagonal Arrangement Patterns) We define $\{0,1\}$ valued diagonal matrices $\mathbf{D}_1,\ldots,\mathbf{D}_p$ that contain an enumeration of the set of hyperplane arrangement patterns of the training data matrix $\mathbf{X}$ as follows. Let us define

$$\mathcal{H} := \Big\{\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}\ge 0))\ \big|\ \mathbf{h}\in\mathbb{R}^d,\mathbf{h}\neq 0\Big\}. \tag{2}$$

We call $\mathbf{D}_1,\ldots,\mathbf{D}_p\in\mathcal{H}$, an enumeration of the elements of $\mathcal{H}$ in an arbitrary fixed order, *diagonal arrangement patterns* associated with the training data matrix $\mathbf{X}$.

The $\{0,1\}$ valued patterns on the diagonals of $\mathbf{D}_1,\ldots,\mathbf{D}_p$ encodes a partition of $\mathbb{R}^d$ by hyperplanes passing through the origin that are perpendicular to the rows of $\mathbf{X}$. The number of such distinct patterns is the cardinality of the set $\mathcal{H}$ and is bounded as follows

$$p:=|\mathcal{H}|\le 2\sum_{k=0}^{r-1}\binom{n-1}{k}\le 2r\Big(\frac{e(n-1)}{r}\Big)^r,$$

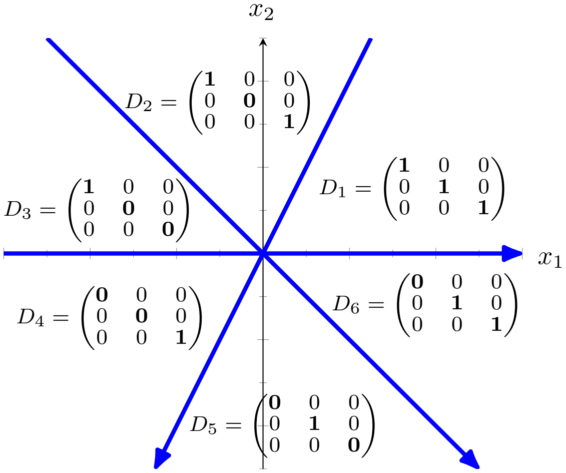

x_2
$D_2 = \begin{pmatrix} \mathbf{1} & 0 & 0 \\ 0 & \mathbf{0} & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}$
$D_1 = \begin{pmatrix} \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}$
$D_3 = \begin{pmatrix} \mathbf{1} & 0 & 0 \\ 0 & \mathbf{0} & 0 \\ 0 & 0 & \mathbf{0} \end{pmatrix}$
x_1
$D_4 = \begin{pmatrix} \mathbf{0} & 0 & 0 \\ 0 & \mathbf{0} & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}$
$D_6 = \begin{pmatrix} \mathbf{0} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} \end{pmatrix}$
$D_5 = \begin{pmatrix} \mathbf{0} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{0} \end{pmatrix}$

Fig. 1: A 2-dimensional example of diagonal arrangement patterns.

where $r = \operatorname{rank}(\mathbf{X})$, see [42]. A 2-dimensional example of diagonal arrangement patterns with three hyperplanes, i.e., $n = 3$, is presented in Figure 1. Note that there are $p = 2\left(\binom{2}{0} + \binom{2}{1}\right) = 6$ regions and corresponding patterns associated with this configuration. For every fixed dimension d (or rank r), the number of patterns p is bounded by $\operatorname{poly}(n)$. In [13], it was shown that convolutional neural networks (CNNs) have a small fixed rank. For instance, a typical convolutional layer of size $3 \times 3 \times 512$, e.g. 512 filters of size 3×3 implies $r \leq 3 \times 3 = 9$.

B. Convex Reformulations

The non-convex optimization problem (1) of ReLU networks with skip connection is equivalent[1] to a convex program:

$$\begin{aligned} \min_{\mathbf{w}_0, \left(\mathbf{w}_j, \mathbf{w}_j'\right)_{j=1}^p} \quad & \sum_{j=1}^p \left(\|\mathbf{w}_j\|_2 + \|\mathbf{w}_j'\|_2 \right) \\ \text{s.t.} \quad & \mathbf{X}\mathbf{w}_0 + \sum_{j=1}^p \mathbf{D}_j \mathbf{X} \left(\mathbf{w}_j - \mathbf{w}_j'\right) = \mathbf{y}, \\ & (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{w}_j \geq 0, (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{w}_j' \geq 0, j \in [p]. \end{aligned} \tag{3}$$

Here $\mathbf{w}_j$ corresponds to the ReLU neuron associated with positive second layer weight and $\mathbf{w}_j'$ corresponds to the ReLU neuron associated with negative second layer weight. These variables can be considered as local linearizations of ReLU neurons in each arrangement.

In practice, it is unnecessary to enumerate all p hyperplane arrangements. Indeed, in [43], it is proven that with $m = O((n/d)\log(n))$ randomly sampled hyperplane arrangements, the relative optimality gap between the convex relaxation and the nonconvex training problem is bounded by $O(\sqrt{\log(n)})$. This implies that the convex parameterization can be solved in polynomial time while the training objective is proportional to the optimal objective (e.g., they both tend to zero as $n \to \infty$). Furthermore, by taking a small number of arrangements (e.g., in the order of thousands) the convex reparameterization can outperform the nonconvex training in standard benchmarks, as shown in [44]. The global optimal set of the non-convex program (1) can be characterized by the optimal solutions of the convex program (3). The next result is an extension of the earlier results [5, 6, 3] to NNs with a skip connection.

Lemma 1 *All globally optimal solutions of the non-convex problem* (1) *of ReLU networks with skip connection can be found (up to splitting and permutation) via the optimal solution set of the convex program* (3) *when* $m \geq m^*$. *Here* $m^* := \mathbb{I}[\tilde{\mathbf{w}}_0 \neq 0] + \sum_{j=1}^p \mathbb{I}[\tilde{\mathbf{w}}_j \neq 0] + \mathbb{I}[\tilde{\mathbf{w}}_j' \neq 0]$ *is the number of non-zero neurons in the minimum norm optimal solution* $\{\tilde{\mathbf{w}}_0, \left(\tilde{\mathbf{w}}_j, \tilde{\mathbf{w}}_j'\right)_{j=1}^p\}$ *of* (3).

C. Linear Neural Isometry Condition

We now consider the case when the labels are generated by a planted linear model, i.e., $\mathbf{y} = \mathbf{X}\mathbf{w}^*$ and ask the following question:

Can ReLU NNs with a linear skip connection learn a planted linear relation as effectively as a linear model?

[1] Under the condition that the NN has sufficiently many neurons (see e.g., Theorem 1 in [2])

In order to prove that the linear model can be recovered by solving the non-convex problem (1) or its convex reformulation (3), we introduce the following isometry condition on the training data.

Definition 2 (Linear Neural Isometry Condition) The linear neural isometry condition for recovering the linear model $\mathbf{y} = \mathbf{X}\mathbf{w}^*$ from (12) is given by:

$$\left\| \mathbf{X}^T \mathbf{D}_j \mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \hat{\mathbf{w}}^* \right\|_2 < 1, \forall j \in [p], \tag{NIC-L}$$

where $\hat{\mathbf{w}}^* := \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$. Note that (NIC-L) holds uniformly for all $\mathbf{w}^*$ only if

$$\left\| \mathbf{I}_d - \sum_{k:\mathbf{x}_k^T \mathbf{w} < 0} \mathbf{x}_k \mathbf{x}_k^T \left(\sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \right\|_2 < 1, \forall \mathbf{w} : \mathbf{w} \neq 0. \tag{4}$$

The above is a spectral isometry condition on the empirical covariance $\hat{\mathbf{\Sigma}} := \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T$ and its subsampled version $\hat{\mathbf{\Sigma}}_{\mathbf{w}} := \sum_{k:\mathbf{x}_k^T \mathbf{w} < 0} \mathbf{x}_k \mathbf{x}_k^T$ that excludes samples activated by a ReLU neuron. Intuitively, for the above condition to hold, the empirical covariance should be relatively stable when samples that lie on a halfspace are removed, and consequently $\left\| \mathbf{I}_d - \hat{\mathbf{\Sigma}}_{\mathbf{w}} \hat{\mathbf{\Sigma}}^{-1} \right\|_2 = \left\| (\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_{\mathbf{w}}) \hat{\mathbf{\Sigma}}^{-1} \right\|_2 < 1$.

In the following proposition, we show that the linear neural isometry condition implies the recovery of the planted linear model by solving (3).

Proposition 1 *Suppose that $n > d$ and the neural network contains an arbitrary number of neurons, i.e., $m \geq 1$. Let $\mathbf{y} = \mathbf{X}\mathbf{w}^*$. Suppose that the linear neural isometry condition* (NIC-L) *holds. Then, the unique optimal solution to* (1) *and* (3) *(up to permutation and splitting[2]) is given by the planted linear model, i.e., $\tilde{\mathbf{W}} = \left\{ \tilde{\mathbf{w}}_0, \{ \tilde{\mathbf{w}}_j, \tilde{\mathbf{w}}'_j \}_{j=1}^p \right\}$, where $\tilde{\mathbf{w}}_0 = \mathbf{w}^*$ and $\tilde{\mathbf{w}}_j = \tilde{\mathbf{w}}'_j = 0$ for $j \in [p]$.*

The above result implies that minimizing the ℓ_2^2 objective subject to the interpolation condition uniquely recovers the ground truth model by setting all the neurons except the skip connection to zero, regardless of the number of neurons in the NN. Remarkably, an NN with an arbitrary number of neurons, i.e., containing arbitrarily many parameters, optimizing the criteria (1) achieves perfect generalization when the ground truth is the linear model. The exact recovery follows from the equivalence of the problem (1) to the group sparsity minimization problem (3), however, this sparsity inducing regularization is hidden in the typical non-convex formulation with weight decay regularization, i.e., $\|\Theta\|_F^2$.

D. Sharp Phase Transition

Our second main result in this paper is that there exists a sharp phase transition in training ReLU NNs when the data is generated by a random matrix ensemble and the observations are produced by a planted linear model. We precisely identify the relation between the number of samples n and the feature dimension d under i.i.d. training data and planted model assumptions. We first summarize our results in this section informally and then present detailed theorems in later sections.

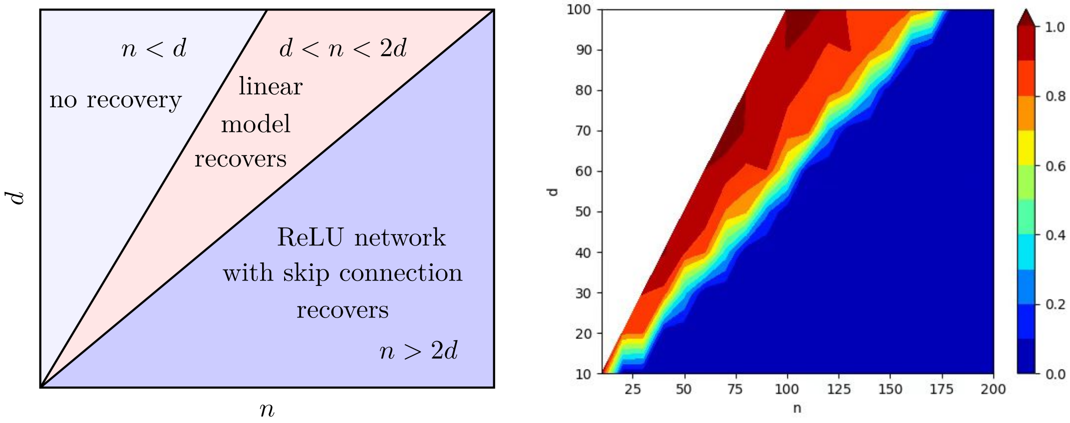
Theorem 1 (informal) *Suppose that the training data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is i.i.d. Gaussian, and $f(\mathbf{X}; \Theta)$ is a two-layer ReLU network containing arbitrarily many neurons with skip connection. Assume that the response is a noiseless linear model $\mathbf{y} = \mathbf{X}\mathbf{w}^*$. The condition $n > 2d$ is sufficient for ReLU networks with skip connections or normalization layers to recover the planted model exactly with high probability. Furthermore, when $n < 2d$, the recovery fails with high probability.*

Interestingly, in the regime where $n \in (d, 2d)$, fitting a simple linear model instead of an NN recovers the planted linear model. In contrast, the ReLU network fails to recover the linear model due to the richness of the model class. We formalize this observation as follows:

Corollary 1 *Suppose that the observations are given by $\mathbf{y} = \mathbf{X}\mathbf{w}^*$, where $\mathbf{w}^* \in \mathbb{R}^d$ is a fixed, unknown parameter. If $n \in (d, 2d)$, a simple linear model $f^{\text{lin}}(\mathbf{X}; \mathbf{w}) = \mathbf{X}\mathbf{w}$ recovers the planted linear neuron $\mathbf{w}^*$ exactly for any $\mathbf{w}^*$, while the ReLU network with skip connection and $m \geq 1$ neurons fitted via either* (5) *with any $\beta \geq 0$ or* (1) *fails with high probability for all $\mathbf{w}^*$ such that $\mathbf{X}\mathbf{w}^* \geq 0$.*

This corollary is illustrated in Figure 2 as a phase diagram in the (n, d) plane. This result clearly shows that the model complexity of ReLU networks can hurt generalization compared to simpler linear models when the number of samples is limited, but information-theoretically sufficient for exact recovery. On the other hand, ReLU networks learn the true model and close the gap in generalization when twice as many samples are available, i.e., $n > 2d$.

[2] Please see Appendix B for a precise definition of the notion of permutation and splitting.



(b)

Fig. 2: Phase transition in recovering a linear neuron. Left: when $n \in (d, 2d)$, ReLU network fails to recover a planted linear model, while a simple linear model succeeds in recovery. Right: Empirical generalization error in recovering a linear neuron by solving the convex program (3) numerically.

E. Noisy Observations

We also develop theoretical results when the observation vector $\mathbf{y}$ is noisy and the following regularized version of the training problem is solved. We consider the regularized training problem

$$\min_{\Theta} \frac{1}{2}\|f(\mathbf{X};\Theta) - \mathbf{y}\|_2^2 + \frac{\beta}{2}R(\Theta), \tag{5}$$

where $R(\Theta) = \|\Theta\|_F^2$ is the weight-decay regularization term.

Theorem 2 (informal) *Suppose that the training data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is with i.i.d. sub-Gaussian entries, and $f(\mathbf{X};\Theta)$ is a two-layer ReLU network with m neurons, skip connection and a normalization layer applied before the ReLU layer. Assume that the observation $\mathbf{y}$ is the sum of a linear neuron and an arbitrary disturbance term $\mathbf{z}$, i.e., $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{z}$. Then, there exists a range of values for the regularization parameter β, such that the non-convex optimization problem in* (5) *and its corresponding convex formulation (see* (24) *in Section V) exactly recovers a linear model with high probability when n is sufficiently large for all values of the number of neurons $m \geq 1$. Additionally, the ℓ_2 distance between the learned linear neuron $\mathbf{w}$ and the planted neuron is bounded by*

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \mathcal{O}(\beta) + \mathcal{O}(\|\mathbf{z}\|_2).$$

Considering the non-convex optimization problem (6) with weight decay, i.e., ℓ_2-regularization, with a proper choice of the regularization parameter β, the optimal NN with an arbitrary number of neurons consists of only a linear weight. The key to understanding this result involves techniques from sparse recovery combined with the convex re-formulation of the regularized training problem (5) as group ℓ_1 regularization.

The above result shows the bias of two-layer ReLU networks towards simple models even when the relation between the data and labels is not exactly linear. Our proof also shows that skip connection and normalization layers in these two-layer networks are critical for learning compact models, as they are the building blocks of the ResNets popularly used in practice. The details of Theorem 2 can be found in Section V-D.

Corollary 2 *By combining the above proposition with Lemma 1, we note that the global optima of the nonconvex problem* (7) *consists of exactly the planted linear neuron up to permutation and splitting.*

III. Convex Formulations of ReLU NNs with Skip Connection and Normalization

In this section, we introduce important variants of the simple ReLU architecture and their corresponding convex reformulations. As will be shown next, architectural choices of these models, such as an addition of a normalization layer play a significant role in their recovery properties.

Suppose that $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the training data matrix and $\mathbf{y} \in \mathbb{R}^d$ is the label vector. We will focus on the following two-layer ReLU NNs:

- plain ReLU networks:
$$f^{\mathrm{ReLU}}(\mathbf{X};\Theta) = (\mathbf{X}\mathbf{W})_+\mathbf{v} = \sum_{i=1}^{m}(\mathbf{X}\mathbf{w}_i)_+v_i, \quad \Theta = (\mathbf{W},\mathbf{v}),$$
where $\mathbf{W} \in \mathbb{R}^{d\times m}$ and $\mathbf{v} \in \mathbb{R}^m$.
- ReLU networks with skip connections:
$$f^{\mathrm{ReLU-skip}}(\mathbf{X};\Theta) = \mathbf{X}\mathbf{w}_1v_1 + \sum_{i=2}^{m}(\mathbf{X}\mathbf{w}_i)_+v_i,$$
where $\Theta = (\mathbf{W},\mathbf{v})$, $\mathbf{W} \in \mathbb{R}^{d\times m}$ and $\mathbf{v} \in \mathbb{R}^m$.
- ReLU networks with normalization layers:
$$f^{\mathrm{ReLU-norm}}(\mathbf{X};\Theta) = \sum_{i=1}^{m}\mathrm{N}_{\alpha_i}((\mathbf{X}\mathbf{w}_i)_+)v_i,$$
where $\Theta = (\mathbf{W},\mathbf{v},\boldsymbol{\alpha})$ and the normalization operator $\mathrm{N}_\alpha(\mathbf{z})$ is defined by
$$\mathrm{N}_\alpha(\mathbf{z}) = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}\alpha, \mathbf{z} \in \mathbb{R}^n, \alpha \in \mathbb{R}.$$

We note that the normalization layer employed in the model $f^{\mathrm{ReLU-norm}}$ is a variant of the well-known batch normalization (BN) with a trainable normalization-correction variable. However, convex formulations of ReLU NNs with exact Batch Normalization layers can be derived as shown in [12]. We now focus on the regularized training problem
$$\min_{\Theta}\frac{1}{2}\|f(\mathbf{X};\Theta)-\mathbf{y}\|_2^2 + \frac{\beta}{2}R(\Theta). \tag{6}$$
For plain ReLU networks and ReLU networks with skip connections, we consider weight decay regularization. This is given by $R(\Theta) = \|\Theta\|_F^2 = (\|\mathbf{W}\|_F^2 + \|\mathbf{v}\|_2^2)$, while for ReLU networks with normalization layers, we have $R(\Theta) = (\|\mathbf{W}\|_F^2 + \|\mathbf{v}\|_2^2 + \|\boldsymbol{\alpha}\|_2^2)$. When $\beta \to 0$, the optimal solution of the above problem approaches the following minimum norm interpolation problem:
$$\min_{\Theta} R(\Theta) \ \text{ s.t. } f(\mathbf{X};\Theta) = \mathbf{y}. \tag{7}$$

A. Convex formulations for plain ReLU NNs

According to the convex optimization formulation of two-layer ReLU networks in [2], the minimum norm problem (7) of plain ReLU networks, i.e., the model $f^{\mathrm{ReLU}}(\mathbf{X};\Theta)$, is equivalent to a convex program:
$$\begin{aligned} \min_{\{\mathbf{w}_j,\mathbf{w}'_j\}_{j=1}^p} \quad & \sum_{j=1}^{p}\left(\|\mathbf{w}_j\|_2 + \|\mathbf{w}'_j\|_2\right) \\ \text{s.t.} \quad & \sum_{j=1}^{p}\mathbf{D}_j\mathbf{X}\left(\mathbf{w}_j - \mathbf{w}'_j\right) = \mathbf{y}, \\ & (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{w}_j \geq 0, (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{w}'_j \geq 0, j \in [p]. \end{aligned} \tag{8}$$
Here the matrices $\mathbf{D}_1,\dots,\mathbf{D}_p$ are diagonal arrangement patterns as defined in Definition 1. Compared to the formulation in [2], we exclude the hyperplane arrangement induced by the zero vector without loss of generality as justified in Appendix D. Analogous to the results in [2, 6], the global optimal set of the non-convex program (7) can be characterized by the optimal solutions of the convex program (8).

Theorem 3 *All globally optimal solutions of the non-convex problem* (7) *of ReLU networks can be computed via the optimal solutions of the convex program* (8) *up to splitting and permutation as soon as the number of neurons m exceeds a critical threshold m^*.*

The inequality constraints in (8) render the analysis of the uniqueness of the optimal solution directly via (8) difficult. However, we show that we can instead work with relaxation without any loss of generality. By dropping all inequality constraints, we obtain a relaxation of the convex program (8) which reduces to the following group ℓ_1-minimization problem:
$$\min_{\{\mathbf{w}_j\}_{j=1}^p} \quad \sum_{j=1}^{p}\|\mathbf{w}_j\|_2 \quad \text{s.t.} \quad \sum_{j=1}^{p}\mathbf{D}_j\mathbf{X}\mathbf{w}_j = \mathbf{y}. \tag{9}$$

The above form corresponds to the well-known Group Lasso model [45] studied in high-dimensional variable selection and compressed sensing. We note that a certain unique solution to the relaxation in (9) satisfying the constraint in the original problem (8) implies that the original problem (8) also has the same unique solution.

Another interesting observation is that the above group ℓ_1-minimization problem corresponds to the minimum norm interpolation problem with gated ReLU activation, for which the NN model is given by

$$f^{\mathrm{gReLU}}(\mathbf{X};\Theta)=\sum_{i=1}^{m}\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_i\geq 0))\mathbf{X}\mathbf{w}_i v_i, \tag{10}$$

where $\Theta=(\mathbf{W},\mathbf{v},\mathbf{H})$. This derivation is provided in Appendix E.

B. Convex formulations for ReLU networks with skip connection

The minimal problem (7) of ReLU networks with skip connection, i.e., the model $f^{\mathrm{ReLU-skip}}(\mathbf{X};\Theta)$, is equivalent to a convex program:

$$\begin{aligned}\min_{\mathbf{w}_0,\left(\mathbf{w}_j,\mathbf{w}_j'\right)_{j=1}^{p}}\quad & \sum_{j=1}^{p}\left(\|\mathbf{w}_j\|_2+\left\|\mathbf{w}_j'\right\|_2\right)\\ \text{s.t.}\quad & \mathbf{X}\mathbf{w}_0+\sum_{j=1}^{p}\mathbf{D}_j\mathbf{X}\left(\mathbf{w}_j-\mathbf{w}_j'\right)=\mathbf{y},\\ & (2\mathbf{D}_j-\mathbf{I}_n)\mathbf{X}\mathbf{w}_j\geq 0,(2\mathbf{D}_j-\mathbf{I}_n)\mathbf{X}\mathbf{w}_j'\geq 0, j\in[p].\end{aligned} \tag{11}$$

By dropping all inequality constraints, the convex program (3) reduces to the following group ℓ_1-minimization problem:

$$\min_{\{\mathbf{w}_j\}_{j=0}^{p}}\quad \sum_{j=0}^{p}\|\mathbf{w}_j\|_2\quad \text{s.t.}\qquad \mathbf{X}\mathbf{w}_0+\sum_{j=0}^{p}\mathbf{D}_j\mathbf{X}\mathbf{w}_j=\mathbf{y}. \tag{12}$$

Similarly, this group ℓ_1-minimization problem is equivalent to the minimum norm interpolation problem of gated ReLU networks with skip connection.

C. Convex formulations for ReLU networks with normalization layer

The minimum norm problem (7) of ReLU networks with normalization layer, i.e., the model $f^{\mathrm{ReLU-norm}}(\mathbf{X};\Theta)$. is equivalent to the following convex program:

$$\begin{aligned}\min_{\left\{\mathbf{w}_j,\mathbf{w}_j'\right\}_{j=1}^{p}}\quad & \sum_{j=1}^{p}\left(\|\mathbf{w}_j\|_2+\left\|\mathbf{w}_j'\right\|_2\right)\\ \text{s.t.}\quad & \sum_{j=1}^{p}\mathbf{U}_j\left(\mathbf{w}_j-\mathbf{w}_j'\right)=\mathbf{y},\\ & (2\mathbf{D}_j-\mathbf{I}_n)\mathbf{X}\mathbf{V}_j^T\mathbf{\Sigma}_j^{-1}\mathbf{w}_j\geq 0,(2\mathbf{D}_j-\mathbf{I}_n)\mathbf{X}\mathbf{V}_j^T\mathbf{\Sigma}_j^{-1}\mathbf{w}_j'\geq 0, j\in[p],\end{aligned} \tag{13}$$

where $\mathbf{D}_j\mathbf{X}=\mathbf{U}_j\mathbf{\Sigma}_j\mathbf{V}_j$ is the compact singular value decomposition (SVD) of $\mathbf{D}_j\mathbf{X}$ for $j\in[p]$. Again, the global optima of the non-convex program (7) can be characterized by the optimal solutions of the convex program (13).

Theorem 4 *Suppose that $n>d$. All globally optimal solutions of the non-convex problem* (7) *of ReLU networks with normalization layer can be found by the optimal solutions of the convex program* (13) *(up to splitting and permutation) when m is greater than some threshold m^*.*

By dropping all inequality constraints, the convex program (13) reduces to a group ℓ_1-minimization problem:

$$\min_{\{\mathbf{w}_j\}_{j=1}^{p}}\quad \sum_{j=1}^{p}\|\mathbf{w}_j\|_2\,,\text{s.t.}\qquad \sum_{j=1}^{p}\mathbf{U}_j\mathbf{w}_j=\mathbf{y}. \tag{14}$$

Analogously, the above group ℓ_1-minimization problem corresponds to the minimum norm interpolation problem of a gated ReLU network with the normalization layer.

IV. Neural isometry conditions and recovery of nonlinear neurons

In this section, we introduce conditions on the training data, called neural isometry, that guarantee the recovery of planted models via solving the non-convex problem (7) or the convex problems (9, 12, 14).

A. Recovery of a single ReLU neuron using plain ReLU NNs

In this section, we present recovery results on ReLU networks. Suppose that $\mathbf{y} = (\mathbf{X}\mathbf{w}^*)_+$ is the output of a planted ReLU neuron[3], where $\mathbf{w}^* \neq 0$. Let $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}^* \geq 0))$ with $i^* \in [p]$ be the diagonal arrangement pattern corresponding to the planted neuron. Based on these diagonal arrangement patterns, we introduce a regularity condition on the data which is analogous to the irrepresentability condition of sparse recovery, called the *Neural Isometry Condition* (NIC). For the recovery of a single neuron, the NIC is defined as follows.

Definition 3 (Neural Isometry Condition for a single ReLU neuron) A sufficient condition for recovering a single neuron $\mathbf{y} = (\mathbf{X}\mathbf{w}^*)_+$ via the problem (9) is given by:

$$\left\|\mathbf{X}^T\mathbf{D}_j\mathbf{D}_{i^*}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X}\right)^{-1}\hat{\mathbf{w}}^*\right\|_2 < 1, \forall j \neq i^*, \tag{NIC-1}$$

where $\hat{\mathbf{w}}^* = \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$.

We assume that the matrix $\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X} = \sum_{k:\,\mathbf{x}_k^T\mathbf{w}^*\geq 0}\mathbf{x}_k\mathbf{x}_k^T$ is invertible whenever the above condition holds. The condition above can be equivalently stated as

$$\left\|\sum_{k:\mathbf{x}_k^T\mathbf{w}^*\geq 0,\,\mathbf{x}_k^T\mathbf{w}\geq 0}\mathbf{x}_k\mathbf{x}_k^T\Big(\sum_{k:\,\mathbf{x}_k^T\mathbf{w}^*\geq 0}\mathbf{x}_k\mathbf{x}_k^T\Big)^{-1}\hat{\mathbf{w}}^*\right\|_2 < 1, \forall \mathbf{w} : \mathbb{I}(\mathbf{X}\mathbf{w} \geq 0) \neq \mathbb{I}(\mathbf{X}\mathbf{w}^* \geq 0). \tag{15}$$

In the following proposition, we show that the neural isometry condition (NIC-1) implies the recovery of the planted model by solving the convex reformulation (8), or equivalently the non-convex problem (7).

Proposition 2 *Suppose that $n \geq d$. Let $\mathbf{y} = (\mathbf{X}\mathbf{w}^*)_+$. Suppose that* (NIC-1) *holds. Then, the unique optimal solution to* (8) *is given by the planted ReLU neuron w^*, i.e, $\hat{\mathbf{W}} = \{(\hat{\mathbf{w}}_j, \hat{\mathbf{w}}'_j)|j \in [p]\}$, where $\hat{\mathbf{w}}_{i^*} = \mathbf{w}^*$, $\hat{\mathbf{w}}'_{i^*} = 0$ and $\hat{\mathbf{w}}_j = \hat{\mathbf{w}}'_j = 0$ for $j \neq i^*$.*

Corollary 3 *By combining the above proposition with Theorem 3, we note that all global optima of the nonconvex problem* (7) *consist of permuted and split versions of the planted ReLU neuron when the condition* (NIC-1) *holds.*

B. Recovery of a single ReLU neuron using ReLU networks with normalization layer

We now consider the case where $\mathbf{y} = \frac{(\mathbf{X}\mathbf{w}^*)_+}{\|(\mathbf{X}\mathbf{w}^*)_+\|_2}$, which is the output of a single-neuron ReLU network followed by a normalization layer. Let $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}^* \geq 0))$ and denote $\tilde{\mathbf{w}}^* = \frac{\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*}{\|\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*\|_2}$. Here, $\mathbf{U}_{i^*}, \mathbf{\Sigma}_{i^*}, \mathbf{V}_{i^*}$ are the SVD factors of $\mathbf{D}_{i^*}\mathbf{X}$ as defined in subsection III-C. We note the simplified expression

$$\mathbf{U}_{i^*}\tilde{\mathbf{w}}^* = \frac{\mathbf{U}_{i^*}\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*}{\|\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*\|_2} = \frac{(\mathbf{X}\mathbf{w}^*)_+}{\|(\mathbf{X}\mathbf{w}^*)_+\|_2} = \mathbf{y}.$$

We introduce the following normalized Neural Isometry Condition.

Definition 4 (Normalized Neural Isometry Condition) The normalized neural isometry condition for recovering the planted model $\mathbf{y} = \frac{(\mathbf{X}\mathbf{w}^*)_+}{\|(\mathbf{X}\mathbf{w}^*)_+\|_2}$ from (14) is given by:

$$\left\|\mathbf{U}_j^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\right\|_2 < 1, \forall j \in [p], j \neq i^*. \tag{NNIC-1}$$

Similarly, the normalized neural isometry condition implies the the recovery of the planted model via solving the convex formulation for ReLU NNs with normalization layer given in (13) or the corresponding non-convex problem (7).

Proposition 3 *Let $\mathbf{y} = \frac{(\mathbf{X}\mathbf{w}^*)_+}{\|(\mathbf{X}\mathbf{w}^*)_+\|_2}$. Suppose that the* NNIC-1 *given in* (NNIC-1) *holds. Then, the unique optimal solution to* (13) *is given by the planted normalized ReLU neuron, i.,e., $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_j, \hat{\mathbf{w}}'_j)_{j=1}^p$, where $\hat{\mathbf{w}}_{i^*} = \frac{\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*}{\|\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*\|_2}$, $\mathbf{w}'_{i^*} = 0$ and $\hat{\mathbf{w}}_j = \hat{\mathbf{w}}'_j = 0$ for $j \neq i^*$.*

Similarly, by combining the above proposition with Theorem 4, we note that all global optima of the nonconvex problem (7) consist of split and permuted versions of the planted neuron.

Remark 1 Our analysis reveals that normalization layers play a key role in the recovery conditions. Note that in (NIC-1), the matrices $\{\mathbf{D}_1\mathbf{X}, \ldots, \mathbf{D}_p\mathbf{X}\}$ are replaced by their whitened versions, effectively canceling the matrix inverse $\left(\sum_{k:\,\mathbf{x}_k^T\mathbf{w}^*\geq 0}\mathbf{x}_k\mathbf{x}_k^T\right)^{-1}$ in (NNIC-1). Therefore, the conditioning of the matrices is improved by the addition of a normalization layer, which applies implicit whitening to the data blocks $\{\mathbf{D}_1\mathbf{X}, \ldots, \mathbf{D}_p\mathbf{X}\}$. As a result, it can be deduced that normalization layers help NNs learn simple models more efficiently from the data.

[3] Here we ignore the second layer weight without loss of generality. Positive weights can be absorbed into the first layer weight, while for negative weights we can consider using $-y$ as the label.

C. Recovering Multiple ReLU neurons using plain ReLU NNs

We now extend the Neural Isometry Condition to the recovery of $k > 1$ ReLU neurons, starting with plain ReLU NNs. Suppose that the label vector is given by

$$\mathbf{y} = \sum_{i=1}^{k} (\mathbf{X}\mathbf{w}_i^*)_+ r_i^*,$$

where $\mathbf{w}_i^* \in \mathbb{R}^d$, $r_i^* \in \{-1, +1\}$ for $i \in [k]$, and $\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}_i^* \geq 0))$ are distinct for $i \in [k]$. Suppose that an enumeration of the diagonal arrangement patterns corresponding to the planted neurons is given by $\mathbf{D}_1, ..., \mathbf{D}_p$. We denote $\mathbf{D}_{s_i} = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}_i^* \geq 0))$ for $i \in [k]$, where $S = \{s_1, \dots, s_k\} \subseteq [p]$ contain the indices of planted neurons in the enumeration of arrangement patterns $\{1, ..., p\}$ according to any fixed order.

Definition 5 (Neural Isometry Condition for k neurons) For recovering k ReLU neurons from the observations $\mathbf{y} = \sum_{i=1}^{k} (\mathbf{X}\mathbf{w}_i^*)_+ r_i^*$ via the optimization problem (13), we introduce the multi-neuron Neural Isometry Condition:

$$\left\| \mathbf{X}^T \mathbf{D}_j \begin{bmatrix} \mathbf{X}^T \mathbf{D}_{s_1} \\ \vdots \\ \mathbf{X}^T \mathbf{D}_{s_k} \end{bmatrix}^{\dagger} \begin{bmatrix} \hat{\mathbf{w}}_1 \\ \vdots \\ \hat{\mathbf{w}}_k \end{bmatrix} \right\|_2 < 1, \forall j \in [p], j \notin S, \tag{NIC-k}$$

where $\hat{\mathbf{w}}_i := r_i^* \mathbf{w}_i^* / \|\mathbf{w}_i^*\|_2 \, \forall i \in [k]$.

In the following proposition, we show that (NIC-k) implies the recovery of the planted model with k neurons by solving the non-convex problem (7) or its convex reformulation (8).

Proposition 4 *Suppose that* (NIC-k) *is satisfied. Then, the unique optimal solution to* (8) *is given by the planted neurons, i.e.,* $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_j, \hat{\mathbf{w}}'_j)_{j=1}^{p}$, *where we let* $\hat{\mathbf{w}}_{s_i} = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2}$, $\mathbf{w}'_{s_i} = 0$ *for* $r_i^* = 1$, $\mathbf{w}_{s_i} = 0$, $\hat{\mathbf{w}}'_{s_i} = \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2}$, *for* $r_i^* = -1$ *and* $\hat{\mathbf{w}}_j = \hat{\mathbf{w}}'_j = 0$ *for* $j \neq i^*$.

As a corollary, by combining the above proposition with Theorem 3, we deduce that all globally optimal solutions of the non-convex problem (7) with plain ReLU NNs consist of the planted model up to permutation and splitting.

D. Recovering Multiple ReLU neurons using ReLU NNs with Normalization Layer

Next, we consider ReLU NNs with normalization layers. Suppose that the label vector takes the form

$$\mathbf{y} = \sum_{i=1}^{k} \frac{(\mathbf{X}\mathbf{w}_i^*)_+}{\|(\mathbf{X}\mathbf{w}_i^*)_+\|_2} r_i^*,$$

where $\mathbf{w}_i^* \in \mathbb{R}^d$ are first layer weights, $r_i^* \in \mathbb{R}$ are second layer weights for $i \in [k]$ and $\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}_i^* \geq 0))$ are distinct for each $i \in [k]$. For simplicity, denote $\mathbf{D}_{s_i} = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}_i^* \geq 0))$ for $i \in [k]$, where $S = \{s_1, \dots, s_k\} \subseteq [p]$. Let $\mathbf{D}_i \mathbf{X} = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i$ be the singular value decomposition for $i \in [p]$.

Definition 6 (Normalized Neural Isometry Condition for k neurons) For recovering k normalized ReLU neurons from the observations $\mathbf{y} = \sum_{i=1}^{k} \frac{(\mathbf{X}\mathbf{w}_i^*)_+}{\|(\mathbf{X}\mathbf{w}_i^*)_+\|_2} r_i^*$ via the optimization problem (13), the normalized neural isometry condition is given by

$$\left\| \mathbf{U}_j^T \begin{bmatrix} \mathbf{U}_{s_1}^T \\ \vdots \\ \mathbf{U}_{s_k}^T \end{bmatrix}^{\dagger} \begin{bmatrix} \tilde{\mathbf{w}}_1 \\ \vdots \\ \tilde{\mathbf{w}}_k \end{bmatrix} \right\|_2 < 1, \forall j \in [p], j \notin S, \tag{NNIC-k}$$

where $\tilde{\mathbf{w}}_i := r_i^* \mathbf{\Sigma}_{s_i} \mathbf{V}_{s_i} \mathbf{w}_i^* / \|\mathbf{\Sigma}_{s_i} \mathbf{V}_{s_i} \mathbf{w}_i^*\|_2 \, \forall i \in [k]$.

Suppose that $\mathbf{D}_{s_1}, \dots, \mathbf{D}_{s_j}$ further satisfy that $\mathbf{D}_{s_i} \mathbf{D}_{s_j} = 0$ for $i \neq j$. Then, we can simplify the neural isometry condition to the following form

$$\left\| \mathbf{U}_j^T \sum_{i=1}^{k} \frac{(\mathbf{X}\mathbf{w}_i^*)_+}{\|(\mathbf{X}\mathbf{w}_i^*)_+\|_2} r_i^* \right\|_2 = \left\| \mathbf{U}_j^T \mathbf{y} \right\|_2 < 1, \forall j \in [p]/S. \tag{16}$$

In the following proposition, we show that (NNIC-k) implies the recovery of the k planted normalized ReLU neurons via the optimization problem (13).

Proposition 5 *Suppose that* (NNIC-k) *is satisfied. Then, the unique optimal solution to* (13) *is given by the planted neurons, i.e.,* $\hat{\mathbf{W}} = (\hat{\mathbf{w}}_j, \hat{\mathbf{w}}'_j)_{j=1}^{p}$, *where* $\hat{\mathbf{w}}_{s_i} = \frac{\mathbf{\Sigma}_{s_i} \mathbf{V}_{s_i} \mathbf{w}_i}{\|\mathbf{\Sigma}_{s_i} \mathbf{V}_{s_i} \mathbf{w}_i\|_2}$, $\hat{\mathbf{w}}'_{s_i} = 0$ *for* $r_i^* = 1$, $\hat{\mathbf{w}}_{s_i} = 0$, $\hat{\mathbf{w}}'_{s_i} = \frac{\mathbf{\Sigma}_{s_i} \mathbf{V}_{s_i} \mathbf{w}_i}{\|\mathbf{\Sigma}_{s_i} \mathbf{V}_{s_i} \mathbf{w}_i\|_2}$ *for* $r_i^* = -1$ *and* $\hat{\mathbf{w}}_j = \hat{\mathbf{w}}'_j = 0$ *for* $j \neq i^*$.

Corollary 4 *By combining the above proposition with Theorem 3, we note that all globally optimal solutions of the non-convex problem* (7) *consist of the split and permuted version of the planted model.*

We explain why we do not use a ReLU NN to recover a linear model in Appendix C

V. Sharp Phase Transitions

One of our major results in this paper is that there exists a sharp phase transition in the success probability of recovering planted neurons in the (n, d) plane for certain random ensembles for the training data matrix. We start illustrating this phenomenon with the case of recovering linear neurons through an application of the kinematic formula for convex cones.

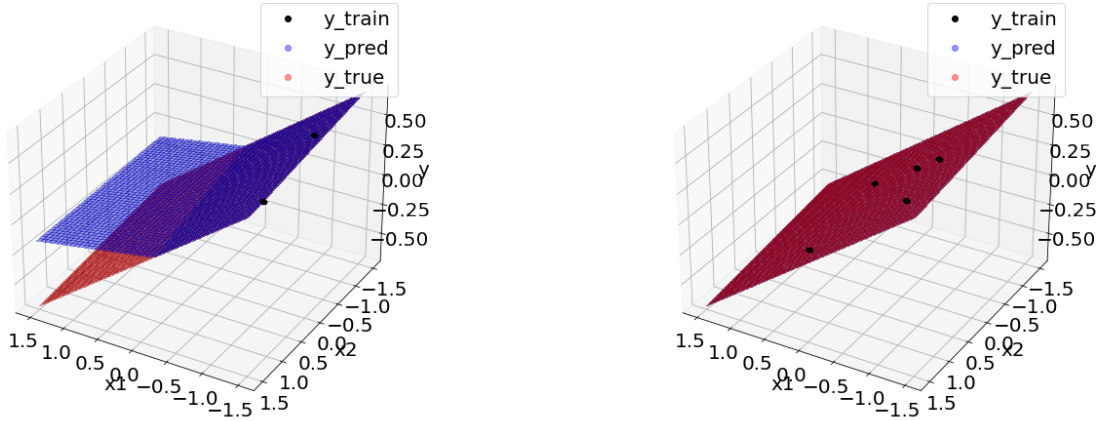


Fig. 3: Optimal ReLU NNs found via the convex program (3). Left: A ReLU neuron is fitted to the observations generated from a linear model when $n = 2, d = 2$. Right: Only a linear neuron is fitted to the observations generated from a linear model when $n = 5, d = 2$.

A. Kinematic formula

We introduce an important result from [1] that will be used in proving the phase transitions in the recovery of planted neurons via ReLU NNs.

Lemma 2 *For a convex cone $K \subseteq \mathbb{R}^n$, define the statistical dimension of the cone K by*

$$\delta(K) = \mathbb{E}\left[\|\Pi_K(\mathbf{g})\|_2^2\right], \tag{17}$$

where $\mathbf{g}$ is a standard Gaussian random vector and Π_K is the Euclidean projection onto the cone K. Define $\alpha := \frac{(n-\delta(K)-d)^2}{64n^2}$. Then, we have

$$P(\exists \mathbf{w} \neq 0 : \mathbf{X}\mathbf{w} \in K) \begin{cases} \leq 4e^{-n\alpha}, & \delta(K) + d < n, \\ \geq 1 - 4e^{-n\alpha}, & \delta(K) + d > n. \end{cases} \tag{18}$$

Consider the positive orthant $K = \mathbb{R}_+^n$, which is a convex cone. It can be easily computed that

$$\delta(K) = \mathbb{E}[\|\pi_K(\mathbf{g})\|^2] = \mathbb{E}\left[\sum_{i=1}^n \max\{0, g_i\}^2\right] = \frac{n}{2}. \tag{19}$$

A direct corollary of the kinematic formula is as follows.

Proposition 6 *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix whose entries are i.i.d. random variables following $\mathcal{N}(0, 1)$. Then, we have*

$$P(\exists \mathbf{h} \neq 0 : \mathbf{X}\mathbf{h} \geq 0) \begin{cases} \leq 4e^{-n\alpha}, & d < n/2, \\ \geq 1 - 4e^{-n\alpha}, & d > n/2, \end{cases}$$

where $\alpha = \frac{(n/2-d)^2}{64n^2}$.

This implies that for $n > 2d$, the probability that the set of diagonal arrangement patterns contain the identity matrix, i.e., a vector $\mathbf{h} \neq 0$ such that $\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0) = \mathbf{1}$, is close to 1, while for $n < 2d$, this probability is close to 0.

B. Recovering Linear Models

1) Gaussian random data: For Gaussian random data, our next result shows that the ratio n/d controls whether the Linear Neural Isometry Condition for linear neuron recovery holds or fails with high probability. Therefore, neural networks do not overfit when the number of samples is above a critical threshold regardless of the number of neurons.

Theorem 5 (Phase transition in Gaussian Data: Success) *Suppose that each entry $x_{i,j}$ of the data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is an i.i.d. random variable following the Gaussian distribution $\mathcal{N}(0, 1/n)$. For $n > 2d$,* NIC-L *shown in* (NIC-L) *holds with probability at least $1 - \exp(-\alpha n)$ where $\alpha = \frac{(n/2-d)^2}{64n^2}$. Consequently, the unique optimal ReLU NN with skip connection found via the convex reformulation* (11) *uniquely recovers planted linear models up to permutation and splitting.*

Remark 2 This implies that when the planted model is $\mathbf{y} = \mathbf{X}\mathbf{w}^*$, the linear term $\mathbf{w}_0 = \mathbf{w}^*$ is the unique optimal solution to the convex program (3). It is also the unique globally optimal solution to the non-convex problem (7).

We next show that when the ratio n/d is below a certain critical threshold, $\mathbf{W} = (\mathbf{w}^*, 0, \dots, 0)$ is not the unique optimal solution to the convex problem (11) (or equivalently to the corresponding non-convex problem (7)). Consequently, neural networks overfit when the number of samples is below this critical threshold.

Theorem 6 (Phase transition in Gaussian Data: Failure) *Suppose that $n < 2d$. Assume that each entry of $\mathbf{X}$ is an i.i.d. random variable following the Gaussian distribution $\mathcal{N}(0, 1/n)$. Suppose that the planted parameter $\mathbf{w}^*$ satisfies $\mathbf{X}\mathbf{w}^* \geq 0$. Then, with probability 1, $\mathbf{W} = (\mathbf{w}^*, 0, \dots, 0)$ is not the unique optimal solution to the minimum norm interpolation problem of ReLU NNs with skip connection* (7) *or its convex reformulation* (11).

Proof When $n < 2d$, according to the kinematic formula, we have

$$P(\exists \mathbf{h} \in \mathbb{R}^d : \mathbf{X}\mathbf{h} \geq 0, \mathbf{h} \neq 0) \geq 1 - e^{-n\alpha}, \tag{20}$$

where α is as defined in Proposition 6. Suppose that there exists $\mathbf{h} \in \mathbb{R}^d$ such that $\mathbf{h} \neq 0$, $\mathbf{X}\mathbf{h} \geq 0$. This implies that the identity matrix is among the diagonal arrangement patterns, i.e., there exists $j \in [p]$ such that $\mathbf{D}_j = \mathbf{I}_n$. Assume that the planted neuron $\mathbf{w}^*$ further satisfies that $\mathbf{X}\mathbf{w}^* \geq 0$. In this case, let $\tilde{\mathbf{w}}_i = \mathbf{w}'_i = 0$ if $i \neq j$, $\tilde{\mathbf{w}}_j = \mathbf{w}^*$ and $\tilde{\mathbf{w}}'_j = 0$. Then, $\mathbf{W}' = (0, \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}'_1, \dots, \tilde{\mathbf{w}}_p, \tilde{\mathbf{w}}'_p)$ is also a feasible solution to (12). This implies that for $n < 2d$, with probability close to 1, there exists an optimal solution which consist of at least one non-zero ReLU neuron. ■

However, if the condition $\mathbf{X}\mathbf{w}^* \geq 0$ is not satisfied, we next show that $\mathbf{W} = (\mathbf{w}^*, 0, \dots, 0)$ is the unique optimal solution to the minimum norm interpolation problem (7) even when $\mathbf{I}_n \in H$.

Proposition 7 *Suppose that $d < n < 2d$. Assume that each entry of $\mathbf{X}$ is an i.i.d. random variable following the Gaussian distribution $\mathcal{N}(0, 1/n)$. Suppose that $\mathbf{X}\mathbf{w}^* \geq 0$ does not hold. Then, with probability 1, $\mathbf{W} = (\mathbf{w}^*, 0, \dots, 0)$ is the unique optimal solution to the minimum norm interpolation problem* (7) *for ReLU NNs with skip connection, or the corresponding convex reformulation* (11).

In Figure 4, we numerically verify the phase transition by solving the convex program (11) and its relaxation given by the group ℓ_1-minimization problem (12), which drops the linear inequality constraints in (3).

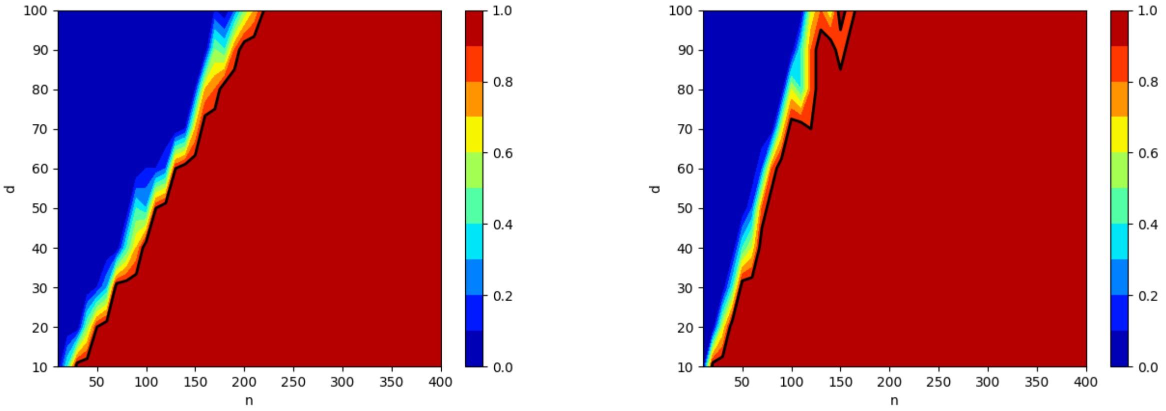


Fig. 4: The empirical probability of successful recovery of the planted linear neuron estimated over 5 independent trials. Left panel: solving the group ℓ_1-minimization problem (12) as a relaxation of (3). Right panel: solving the convex NN program (3). Red (blue) region shows the region where exact recovery probability is close to one (zero). The black lines represent the boundaries of successful recovery with probability 1.

2) Haar random data: We now investigate the case where the training data $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a Haar distributed random matrix. More precisely, $\mathbf{X}$ is uniformly sampled from the set of column orthogonal matrices $\left\{\mathbf{X} \in \mathbb{R}^{n\times d} \,|\, \mathbf{X}^T\mathbf{X} = \mathbf{I}_d\right\}$ for $n \geq d$. In this case, since $(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{I}_d$, (NIC-L) for the recovery of a linear neuron reduces to

$$\max_{\mathbf{h}\in\mathbb{R}^d,\, \mathbf{h}\neq 0} \left\|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{Xh}\geq 0))\mathbf{X}\hat{\mathbf{w}}\right\|_2 < 1, \tag{orth-NIC-L}$$

where $\hat{\mathbf{w}} := \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$.

Based on the simpler form of (orth-NIC-L), we derive the phase transition results on the recovery of a planted linear neuron. Moreover, this form enables the analysis of a stronger recovery condition that ensures the simultaneous recovery of all planted linear neurons.

Theorem 7 (Phase transition in Haar data) *Suppose that the planted neuron $\mathbf{w}^* \in \mathbb{R}^d$, $\mathbf{w}^* \neq 0$ is fixed and $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a Haar distributed random matrix. Denote $\mathbf{W}^* = (\mathbf{w}^*, 0, \dots, 0)$. For $n > 2d$, with probability at least $1 - \exp(-\alpha n)$, $\mathbf{W}^*$ is the unique optimal solution to* (11)*, where $\alpha = \frac{(n/2-d)^2}{64n^2}$. For $n < 2d$, with probability at least $1 - \exp(-\alpha n)$, $\mathbf{W}^*$ is not the unique optimal solution to* (11)*.*

Similar to our previous analysis for the Gaussian distribution, if $\mathbf{Xw}^* \geq 0$, then $\mathbf{W}^* = (\mathbf{w}^*, 0, \dots, 0)$ is not the unique optimal solution to (3). However, if $\mathbf{Xw}^* \geq 0$ is not satisfied, for $n > d$, $\mathbf{W}^*$ can still be the unique optimal solution of (3).

3) Relation between Haar data and normalization layers: Consider the minimum norm interpolation problem of two-layer ReLU networks with skip connection and normalization layer (before ReLU)

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{v},\boldsymbol{\alpha}} \left(\|\mathbf{W}\|_F^2 + \|\mathbf{v}\|_2^2 + \|\boldsymbol{\alpha}\|_2^2\right), \\ &\text{s.t.} \quad \mathbf{y} = \mathrm{N}_{\alpha_1}\mathbf{X}\mathbf{w}_1 v_1 + \sum_{i=2}^m (\mathrm{N}_{\alpha_i}(\mathbf{X}\mathbf{w}_i)_+ v_i. \end{aligned} \tag{21}$$

The above problem can be reformulated as a convex program [12]:

$$\begin{aligned} &\min_{\mathbf{w}_0, \{\mathbf{w}_j, \mathbf{w}_j'\}_{j=1}^p} \left(\|\mathbf{w}_0\|_2 + \sum_{j=1}^p (\|\mathbf{w}_j\|_2 + \|\mathbf{w}_j'\|_2)\right), \\ &\text{s.t. } \mathbf{U}\mathbf{w}_0 + \sum_{j=1}^p \mathbf{D}_j\mathbf{U}(\mathbf{w}_j - \mathbf{w}_j') = \mathbf{y}, \\ &(2\mathbf{D}_j - \mathbf{I}_n)\mathbf{U}\mathbf{w}_j \geq 0, (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{U}\mathbf{w}_j' \geq 0, j \in [p], \end{aligned} \tag{22}$$

where the matrix $\mathbf{U}$ is computed by the compact SVD $\mathbf{X} = \mathbf{U\Sigma V}$. Therefore, the training problem with Gaussian random data with an additional normalization layer before the ReLU activation is equivalent to the training problem with Haar data with no normalization layers.

4) Sub-Gaussian random data: Our previous results on Gaussian and Haar data can be extended to a broad class of random matrices with independent entries. In the following result, we show that for i.i.d., sub-Gaussian data distributions, a ReLU neural network with a skip connection recover a planted linear model exactly with sufficiently many samples.

Theorem 8 *Suppose that each entry $x_{i,j}$ of the data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is an i.i.d. symmetric random variable following a mean-zero sub-Gaussian distribution with variance proxy σ^2 and $\mathbb{E}[x_{i,j}^2] = \frac{1}{n}$. Then,* (NIC-L) *holds when $n \geq C_1 d\log(n)$ with probability at least $1-4\exp(-C_2 n)-2\exp(d)$ for sufficiently large n, where C_1, C_2 are absolute constants depending only on σ^2.*

We note that there exists an additional logarithmic term in the scaling $n \geq C_1 d\log(n)$ provided in Theorem 8, compared to our results on Gaussian and Haar data matrices.

5) General datasets: We also provide two examples of the recovery of a linear model for general datasets. The condition (NIC-L) can hold when X is an identity matrix and $\mathbf{w}^*$ has no zero entries. This is because $\|X^T D_j X(X^TX)^{-1}\hat{\mathbf{w}}^*\|_2 = \|\mathbf{D}_j\hat{w}^*\|_2 < \|\hat{\mathbf{w}}^*\|_2 = 1$, as $\hat{\mathbf{w}} = \mathbf{w}^*/\|\mathbf{w}^*\|_2$ also has no zero entries. On another case, suppose that we have $\mathbf{X}^T\mathbf{X} = \mathbf{I}_d$ and $\mathbf{w}^T\mathbf{x}_j \neq 0$ for all $j \in [n]$. Let $\mathbf{D}_j = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w} \geq 0))$. Then, the condition (NIC-L) holds as

$$\|\mathbf{X}^T\mathbf{D}_j X(\mathbf{X}^T\mathbf{X})^{-1}\hat{\mathbf{w}}^*\|_2 = \|\hat{\mathbf{w}}^* - \sum_{k:\mathbf{x}_k^T\mathbf{w}<0}\mathbf{x}_k\mathbf{x}_k^T\hat{\mathbf{w}}^*\|_2 < 1,$$

C. Uniform (Strong) Recovery of All Linear Models

Thus far, we considered the recovery of a fixed planted neuron and the associated exact recovery probability. To ensure the recovery of all possible planted neurons $\mathbf{w}^*$ simultaneously, we introduce the following stronger form of NIC-L specialized to column orthogonal matrices by requiring (orth-NIC-L) to hold for every $\mathbf{w}^*$:

$$\max_{\mathbf{h}\in\mathbb{R}^d,\,\mathbf{h}\neq 0}\left\|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0))\mathbf{X}\right\|_2 < 1. \qquad \text{(ortho-SNIC-L)}$$

For a diagonal arrangement pattern $\mathbf{D}_j$, $\|\mathbf{X}^T\mathbf{D}_j\mathbf{X}\|_2 < 1$ is equivalent to $\mathbf{X}^T\mathbf{D}_j\mathbf{X} \prec \mathbf{I}_d$ or, equivalently, $\mathbf{X}^T(\mathbf{I}_n - \mathbf{D}_j)\mathbf{X} \succ 0$. As $\mathbf{X}$ is column orthonormal, this is also equivalent to $\mathrm{tr}(\mathbf{I}_n - \mathbf{D}_j) \geq d$, or equivalently, $\mathrm{tr}(\mathbf{D}_j) \leq n-d$. Therefore, (ortho-SNIC-L) can be simplified for column orthogonal matrices as

$$\text{(orth-SNIC-L)} \qquad \max_{\mathbf{h}\in\mathbb{R}^d,\,\mathbf{h}\neq 0}\mathrm{tr}(\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0))) \leq n-d. \tag{23}$$

Theorem 9 *Suppose that $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a Haar distributed random matrix. Let $\theta^* > 0$ be the unique solution of the scalar equation $\theta + \frac{1}{2}\int_{F_{\chi^2}^{-1}(1-2\theta)}^{\infty} r dF_{\chi^2}(r) = \frac{1}{2}$, where $F_{\chi^2}(r)$ is the cumulative distribution function (CDF) of a χ^2-random variable with 1 degree of freedom. Then, for sufficiently large n, d satisfying $\frac{d}{n} < \theta^*$, the* orth-SNIC-L *given in* (23) *holds with high probability.*

Remark 3 It can be computed that $\theta^* \approx 0.1314$ and $(\theta^*)^{-1} \approx 7.613$ by numerically solving the scalar equation $\theta + \frac{1}{2}\int_{F_{\chi^2}^{-1}(1-2\theta)}^{\infty} r dF_{\chi^2}(r) = \frac{1}{2}$. This implies that when $n > 7.613d$, the orth-SNIC-L shown in (23) holds w.h.p.

D. Inexact/Noisy Linear Recovery

In this section, we consider inexact or noisy observations $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^n$ is an arbitrary disturbance component. We will show that the optimal NN only learns a linear model when the regularization parameter β is chosen from an appropriate interval. This can be understood intuitively as follows: when β is too small, the NN trained with insufficient norm penalty overfits the noise. On the other hand, when β is too large, the NN underfits the observations as the norms of the parameters are over-penalized.

For two-layer ReLU networks with skip connection and normalization layer before the ReLU, the regularized training problem (6) can be equivalently cast as the convex program (see e.g., [2, 6]):

$$\begin{aligned}\min_{\mathbf{w}_0,\{\mathbf{w}_j,\mathbf{w}_j'\}_{j=1}^p} \quad & \left\|\mathbf{U}\mathbf{w}_0 + \sum_{j=1}^p \mathbf{D}_j\mathbf{U}\left(\mathbf{w}_j - \mathbf{w}_j'\right) - \mathbf{y}\right\|_2^2 + \beta\Big(\|\mathbf{w}_0\|_2 + \sum_{j=1}^p\Big(\|\mathbf{w}_j\|_2 + \|\mathbf{w}_j'\|_2\Big)\Big) \\ \text{s.t.} \quad & (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{U}\mathbf{w}_j \geq 0, j\in[p], (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{U}\mathbf{w}_j' \geq 0, j \in [p],\end{aligned} \tag{24}$$

where $\mathbf{U}$ is computed from the compact SVD of $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

As a direct corollary of Theorem 1 in [6], the global optima of the nonconvex regularized training problem (6) are given by the optimal solutions of (24) up to permutation and splitting (see Appendix B for details). By dropping all inequality constraints, the convex program (24) reduces to the following group-Lasso problem:

$$\min_{\{\mathbf{w}_j\}_{j=0}^p} \quad \Big\| \sum_{j=0}^p \mathbf{D}_j \mathbf{U} \mathbf{w}_j - \mathbf{y} \Big\|_2^2 + \beta \sum_{j=0}^p \|\mathbf{w}_j\|_2 \,. \tag{25}$$

As we show next, the convex NN objective (24) inherits recovery properties from its group-Lasso relaxation above. The following two theorems show that for sufficiently large n, with a suitable choice of the regularization parameter β, the optimal neural network learns only a linear component. In addition, the ℓ_2 distance between the optimal solution and the embedded neuron can be bounded by a linear function of β.

Theorem 10 *Suppose that the entries of the data matrix $\mathbf{X}$ is with i.i.d. sub-Gaussian entries with variance proxy σ^2 as in Theorem 8 and the noisy observation takes the form $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{z}$. Assume that the weight decay regularization parameter satisfies $\beta \in \left[\|\mathbf{z}\|_2 \frac{7(\eta - \|\mathbf{z}\|_2)}{\eta - 7\|\mathbf{z}\|_2}, \eta - \|\mathbf{z}\|_2\right]$, the norm of the noise component satisfies $\|\mathbf{z}\|_2 \leq \frac{1}{14}\eta$, where $\eta \triangleq \left\|\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}^*\right\|_2$ and $n \geq \max\{4000\sigma^2 d \log(54n), 1024d\}$. Then, with probability at least $1 - 4\exp(-n/8000\sigma^2) - 2\exp(-d)$, the optimal solutions to convex programs (24) and (25) consist of only a linear neuron and no ReLU neurons, i.e., there exists $\mathbf{w}$ such that $\mathbf{W} = (\mathbf{w}, 0, \ldots, 0)$ is strictly optimal. As a consequence, the non-convex weight decay regularized objective (6) has the same strictly optimal solutions up to permutation and splitting. Furthermore, we have the ℓ_2 distance upper bound*

$$\left\|\mathbf{w} - \mathbf{\Sigma}\mathbf{V}^T\mathbf{w}^*\right\|_2 \leq \frac{\beta\eta}{\eta - \|\mathbf{z}\|_2} + \|\mathbf{z}\|_2.$$

Here, η satisfies $(1 - 1/16)\|\mathbf{w}^*\|_2 \leq \eta \leq (1 + 1/16)\|\mathbf{w}^*\|_2$ with probability $1 - 2\exp(-d)$. Moreover, the optimal weights of the neural network $f(\mathbf{X}; \Theta) = \mathrm{N}_{\alpha_1}(\mathbf{X}\mathbf{w}_{1,1})\mathbf{w}_{2,1} + \sum_{i=2}^m \left(\mathrm{N}_{\alpha_i}(\mathbf{X}\mathbf{w}_{1,i})\right)_+ \mathbf{w}_{2,i}$ are given by $\mathbf{w}_{1,1} = \mu\mathbf{w}, \alpha_1 = \mathbf{w}_{2,1} = \sqrt{\|\mathbf{X}\mathbf{w}\|_2}$ and $\mathbf{w}_{1,j} = \mathbf{0}, \alpha_j = \mathbf{w}_{2,j} = 0$ for $2 \leq j \leq n$, where $\mu > 0$ is an arbitrary constant.

Remark 4 (Estimating the regularization coefficient) In practice, it may be appropriate to assume the noise component $\mathbf{z}$ follows an i.i.d. sub-Gaussian distribution with zero mean. Under this assumption, we can control the term $\|\mathbf{z}\|_2$ with high probability using classical concentration bounds. From the triangle inequality, we note that $\|\mathbf{y}\|_2 \in [\eta - \|\mathbf{z}\|_2, \eta + \|\mathbf{z}\|_2]$. Thus, we can estimate η using the interval $\eta \in \left[\left(\|\mathbf{y}\|_2 - \|\mathbf{z}\|_2\right)_+, \|\mathbf{y}\|_2 + \|\mathbf{z}\|_2\right]$, and use η to compute the required inverval for β in Theorem 10. Knowing that a large interval for β exists, we can run a hold-out or cross-validation scheme to select an appropriate value of β.

We note that the sample complexity required for noisy recovery in Theorem 10 is near-optimal ignoring the $\log(n)$ term and the constant factor. This follows from standard minimax lower-bounds (e.g., see the book [46]) showing that $n \geq \Omega(\sigma^2 d)$ is required for stable estimation in the linear model, which can be achieved by fitting a linear model. The additional $\log(n)$ factor in our bounds is due to the fitting of the more complex ReLU NN model. However, our result shows that there is no disadvantage to using a ReLU NN with skip connection even when the ground truth is linear if the sample size is only a factor of $\Omega(\log n)$ larger.

E. Recovering a ReLU Neuron

Now we focus on recovering a single planted ReLU neuron. We consider the case where the training data matrix $\mathbf{X}$ is composed of i.i.d. standard Gaussian variables $\mathcal{N}(0, 1/n)$. The following theorem illustrates that when $n > 2d$, the NNIC-1 will hold with high probability. In other words, ReLU NNs with normalization layer optimizing the objective (9) uniquely recovers the ReLU neuron with high probability.

Theorem 11 *Let $\mathbf{w}^* \in \mathbb{R}^d$ is a fixed unit-norm vector. Suppose that each entry of $\mathbf{X}$ are i.i.d. random variables following the Gaussian distribution $\mathcal{N}(0, 1/n)$. Let $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}^* \geq 0))$. Then, when $n > 2d$, the* NNIC-1 *given in* (NNIC-k) *holds with probability at least $1 - \exp\left(-\frac{1}{6}\left(\frac{n-2d}{n}\right)^2 n\right)$.*

The above result shows that a single normalized ReLU neuron is uniquely recovered by ReLU NN with normalization layer via (9) (up to permutation and splitting) and its convex reformulation (13), regardless of the number of neurons in the NN. In other words the set of global optimum of (9) only consists of networks with only a single non-zero ReLU neuron along with permutations and split versions of this neuron.

Remark 5 We note that the recovery condition for a single ReLU neuron and a linear neuron is the same using ReLU NNs with normalization layer and is given by $\frac{n}{d} \geq 2$.

VI. Asymptotic analysis

In this section, we present an asymptotic analysis of the Neural Isometry Conditions when n goes to infinity while d is fixed. While our analysis in Section V provides sharp estimates on the recovery threshold for a single linear or ReLU neuron, the asymptotic analysis in this section proves the recovery of multiple ReLU neurons.

We begin with the case of two planted ReLU neurons in the asymptotic setting. For $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, recall our notation for the cosine angle between $\mathbf{w}$ and $\mathbf{v}$ given by $\cos\angle(\mathbf{w}, \mathbf{v}) = \frac{\mathbf{w}^T\mathbf{v}}{\|\mathbf{w}\|_2\|\mathbf{v}\|_2}$. In order to show that the NNIC-2 in (NNIC-k) holds, and consequently two-neuron recovery succeeds via ReLU NNs with normalization layer, we calculate the asymptotic limit of the left-hand-side in (NNIC-k) when $k = 2$, which is given by

$$T := \left\| \mathbf{U}_j^T \begin{bmatrix} \mathbf{U}_{s_1} & \mathbf{U}_{s_2} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{s_1}^T\mathbf{U}_{s_1} & \mathbf{U}_{s_1}^T\mathbf{U}_{s_2} \\ \mathbf{U}_{s_2}^T\mathbf{U}_{s_1} & \mathbf{U}_{s_2}^T\mathbf{U}_{s_2} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\mathbf{w}}_1 \\ \tilde{\mathbf{w}}_2 \end{bmatrix} \right\|_2 . \tag{26}$$

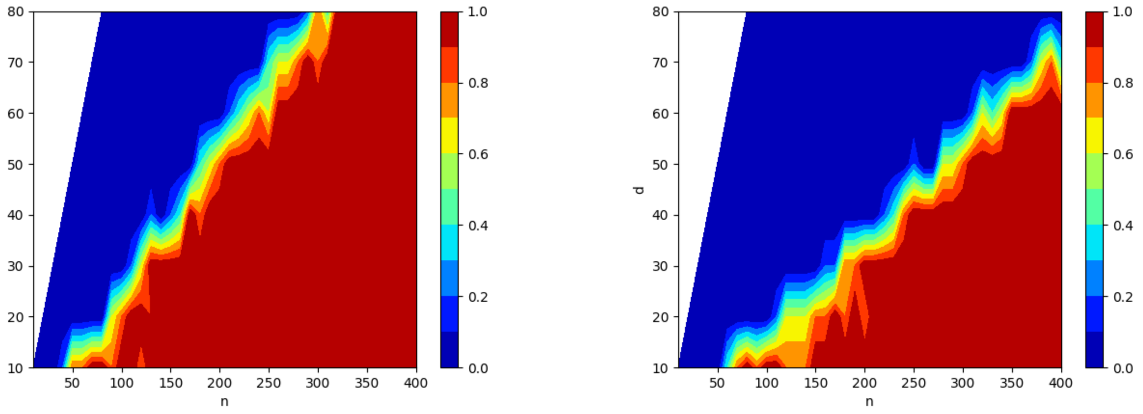
We consider the case of $\mathbf{w}_1, \mathbf{w}_2$ satisfying $\cos\angle(\mathbf{w}_1, \mathbf{w}_2) = -1$ and $\cos\angle(\mathbf{w}_1, \mathbf{w}_2) = 0$ for simplicity.

Proposition 8 *Suppose that each entry of $\mathbf{X} \in \mathbb{R}^{n\times d}$ is an i.i.d. random variable following the normal distribution $\mathcal{N}(0, 1/n)$. Suppose that $\mathbf{y} = \frac{(\mathbf{X}\mathbf{w}_1)_+}{\|(\mathbf{X}\mathbf{w}_1)_+\|_2} + \frac{(\mathbf{X}\mathbf{w}_2)_+}{\|(\mathbf{X}\mathbf{w}_2)_+\|_2}$ is the planted model, where $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$. Let $\mathbf{D}_1 = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}_1 \geq 0))$ and $\mathbf{D}_2 = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}_2 \geq 0))$. Consider any diagonal arrangement pattern $\mathbf{D}_j = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_j \geq 0))$. Consider the random variable T defined in* (26)*.*

- *Suppose that $\cos\angle(\mathbf{w}_1, \mathbf{w}_2) = -1$. Let $\gamma = \cos\angle(\mathbf{w}_1, \mathbf{h}_j)$. As $n \to \infty$, T converges in probability to a univariate function $g_1(\gamma)$. Here $g_1(\gamma) \leq 1$ and the equality holds if and only $\gamma = 1$ or $\gamma = -1$.*
- *Suppose that $\cos\angle(\mathbf{w}_1, \mathbf{w}_2) = 0$. Let $\gamma_1 = \cos\angle(\mathbf{w}_1, \mathbf{h}_j)$ and $\gamma_2 = \cos\angle(\mathbf{w}_2, \mathbf{h}_j)$. As $n \to \infty$, T converges in probability to a bivariate function $g_2(\gamma_1, \gamma_2)$. Here $g_2(\gamma_1, \gamma_2) \leq 1$ and the equality holds if and only $(\gamma_1, \gamma_2) = (1, 0)$ or $(\gamma_1, \gamma_2) = (0, 1)$.*

Therefore, in both of the above cases, we have $T{<}1$ as $n \to \infty$ and consequently the NNIC-2 *holds. Moreover, the planted two-neuron NN is the unique optimal solution to* (7) *(up to permutation and splitting) and its convex reformulation* (13)*.*

We validate this asymptotic behavior in Figure 5. It can be observed that the recovery threshold for two ReLU neurons is approximately $n \geq 4d$ when $\mathbf{w}_1 = e_1$ and $\mathbf{w}_2 = e_2$, i.e., $\cos\angle(\mathbf{w}_1, \mathbf{w}_2) = 0$ in Figure 5(a). In addition, it can be observed that the recovery threshold for three ReLU neurons is approximately $n \geq 6d$.



(a) $k = 2,\ \mathbf{w}_1^* = \mathbf{e}_1,\ \mathbf{w}_2^* = \mathbf{e}_2,$

(b) $k = 3,\ \mathbf{w}_i^* = \mathbf{e}_i (i = 1, 2, 3)$

Fig. 5: The empirical probability of successful recovery of the planted normalized ReLU neurons by solving the group ℓ_1-minimization problem (14) over 5 independent trials. The label vector $\mathbf{y}$ is generated by a planted ReLU NN with k neurons.

Similar to the ReLU networks with the normalization layer, we present asymptotic analysis of plain ReLU networks.

Proposition 9 *Suppose that each entry of $\mathbf{X} \in \mathbb{R}^{n\times d}$ is an i.i.d. random variable following the normal distribution $\mathcal{N}(0, 1/n)$. Let $\mathbf{D}_i = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}^* \geq 0))$. Consider any hyperplane arrangement $\mathbf{D}_j = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_j \geq 0))$ such that* $\cos\angle(\mathbf{w}^*, \mathbf{h}_j) =: \gamma < 1$. *Define*

$$R := \left\| \mathbf{X}^T\mathbf{D}_j\mathbf{D}_i\mathbf{X} \left(\mathbf{X}^T\mathbf{D}_i\mathbf{X}\right)^{-1} \hat{\mathbf{w}}^* \right\|_2 , \tag{27}$$

where $\hat{\mathbf{w}}^* = \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$. Then, as $n \to \infty$, R converges in probability to $g(\gamma)$. Here the function $g : [-1,1] \to \mathbb{R}$ monotonically increases on $[-1,1]$ and $g(1) = 1$.

This implies that asymptotically, as $n \to \infty$, the NIC-1 given in (NIC-1) holds, and plain ReLU NNs recover a single planted ReLU neuron.

VII. Numerical experiments

In this section, we present numerical experiments on ReLU networks with skip connection and normalization layer to validate our theoretical results on phase transitions in different NN architectures. We provide the illustration of our main results in this section and provide additional numerical results for various settings in Appendix M. The code is available at https://github.com/pilancilab/Neural-recovery

Numerical results are divided into three parts: the first part consists of phase transition graphs for the recovery rate when the observation is noiseless by solving convex NN problems in (12) and (14). The second and third parts consist of phase transition graphs for certain types of distance measures when the observation is noisy by solving convex NN problems and the regularized training problem (convex and non-convex), respectively.

A. ReLU networks with skip connection

We start with phase transition graphs for successful recovery of the planted neuron by solving the convex optimization problem (12). We compute the recovery rate for d ranging from 10 to 100 and n ranging from 10 to 400. For each pair of (n, d), we generate 5 realizations of random training data matrices and solve the convex problem (12) on each dataset. We test for four types of randomly generated data matrices:

- Gaussian: each entry $x_{i,j}$ of $\mathbf{X} \in \mathbb{R}^{n\times d}$ is an i.i.d. random variable following the normal distribution $\mathcal{N}(0, 1/n)$.
- cubic Gaussian: each element $x_{i,j}$ of $\mathbf{X} \in \mathbb{R}^{n\times d}$ satisfies $x_{i,j} = z_{i,j}^3$, where $z_{i,j}$ are i.i.d. random variable following $\mathcal{N}(0, 1/n)$.
- Haar: $\mathbf{X} \in \mathbb{R}^{n\times d}$ is drawn uniformly random from the set of column orthonormal matrices. We note that a Haar matrix can be generated by sampling an i.i.d. Gaussian matrix as above and extracting its d left singular vectors of dimension n.
- whitened cubic Gaussian: $\mathbf{X} \in \mathbb{R}^{n\times d}$ is drawn non-uniformly from the set of column orthonormal matrices the matrix of left singular vectors of $\mathbf{X}'$ if $n > d$ and the matrix of right singular vectors of $\mathbf{X}'$ if $n < d$. Here $\mathbf{X}' \in \mathbb{R}^{n\times d}$ is a cubic Gaussian data matrix.

In each recovery problem, the planted neuron $\mathbf{w}^*$ is either a random vector following $\mathcal{N}(0, \mathbf{I}_d)$ or chosen as the smallest right singular vector of $\mathbf{X}$ as specified. In numerical experiments, we use a random subset $\mathcal{H}'$ of the set $\mathcal{H}$ of all possible hyperplane arrangements to approximate the solution of the convex program. Here $\mathcal{H}'$ is generated by

$$\mathcal{H}' = \{\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_i))|\mathbf{h}_i \in \mathbb{R}^d, i \in [\tilde{n}]\},$$

where $\mathbf{h}_i$ is an i.i.d. random vector following the standard normal distribution $\mathcal{N}(0, \mathbf{I}_d)$ and $\tilde{n} = \max(n, 50)$. On the other hand, from our theoretical analysis, the recovery will fail if there exists an all-ones hyperplane arrangement, i.e., $\mathbf{I}_n \in \mathcal{H}$. However, as $\mathcal{H}'$ is a random subset of $\mathcal{H}$, it might not be easy to validate $\mathbf{I}_n \in \mathcal{H}$ by examining whether $\mathbf{I}_n \in \mathcal{H}'$ is satisfied. Therefore, we solve the following feasibility problem before solving the convex problem (12).

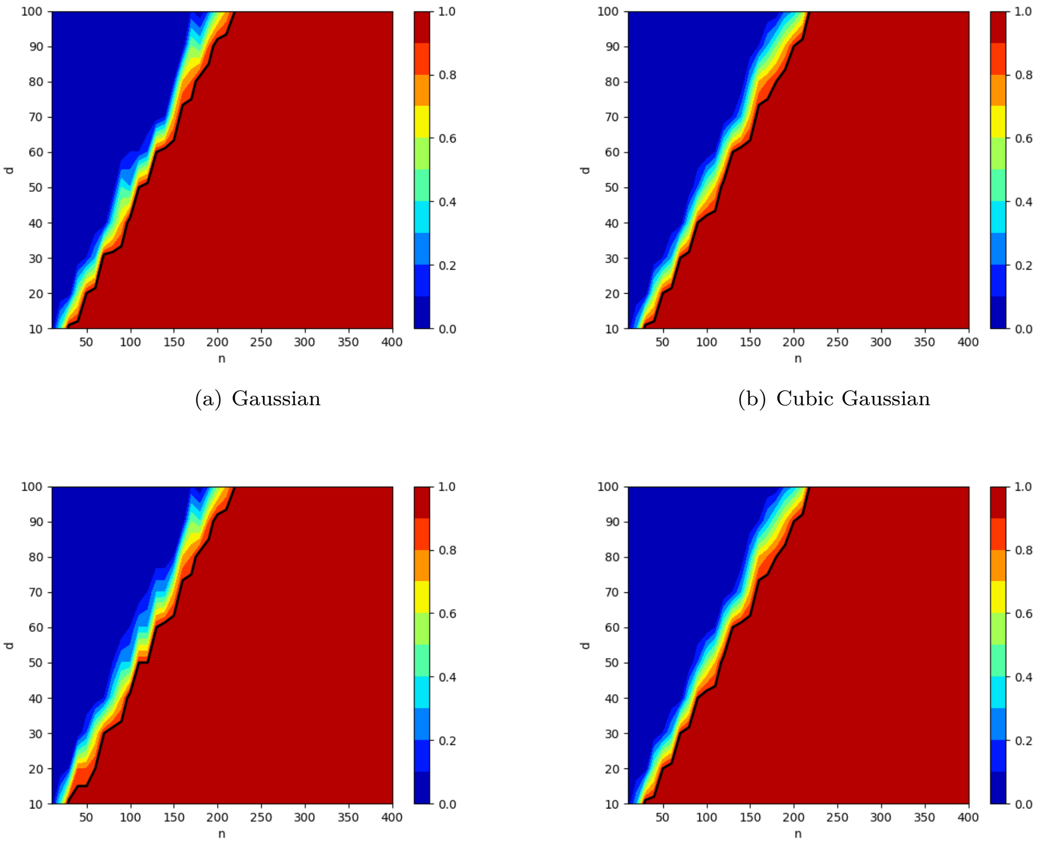
$$\begin{aligned} \max_{\mathbf{w}\in\mathbb{R}^d, t\in\mathbb{R}} \quad & t \\ \text{s.t. } \|\mathbf{w}\|_2 \le 1, \quad & \mathbf{X}w \ge t 1_n\,. \end{aligned} \tag{28}$$

If the optimal value of the above problem is strictly greater than zero, there must exist an all-ones hyperplane arrangement, i.e., $\exists w : \mathbf{X}w > 0$. In this case, we add $\mathbf{I}_n$ to the subset $\mathcal{H}'$. Otherwise, such an arrangement pattern does not exist.

In Figure 6, we present the phase transition graph for the probability of successful recovery when the planted neuron $\mathbf{w}^*$ is randomly generated from $\mathcal{N}(0, \mathbf{I}_d)$. The boundaries indicate a phase transition between $n = 2d$ and $n = 3d$. In Appendix M-A, we will show similar phase transition graphs when the planted neuron $\mathbf{w}^*$ is the smallest right singular vector of $\mathbf{X}$ in Figure 21.

The second part is phase transition under noisy observation, i.e., $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2/n)$. We still focus on the convex problem (12), i.e., the convex optimization formulation of gated ReLU networks with skip connection. Here we focus on Gaussian data and choose $\mathbf{w}^*$ as the smallest right singular vector of $\mathbf{X}$. We define the following two types of distance for the solution of the convex program to evaluate the performance.

- Absolute distance: the ℓ_2 distance between the linear term $\mathbf{w}_0$ and $\mathbf{w}^*$.



(a) Gaussian

(b) Cubic Gaussian

(c) Haar

(d) Whitened cubic Gaussian

Fig. 6: The probability of successful recovery of the planted linear neuron by solving the group ℓ_1-minimization problem (12) over 5 independent trials. The black lines represent the boundaries of successful recovery with probability 1. Here the planted neuron $\mathbf{w}^*$ is randomly generated from $\mathcal{N}(0, \mathbf{I}_d)$.

- Test distance: generate a test set $\tilde{\mathbf{X}}$ with the same distribution as $\mathbf{X}$, then the prediction of the learned model is

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{w}_0 + \sum_{j=1}^{P}(\tilde{\mathbf{X}}\mathbf{w}_j)_+.$$

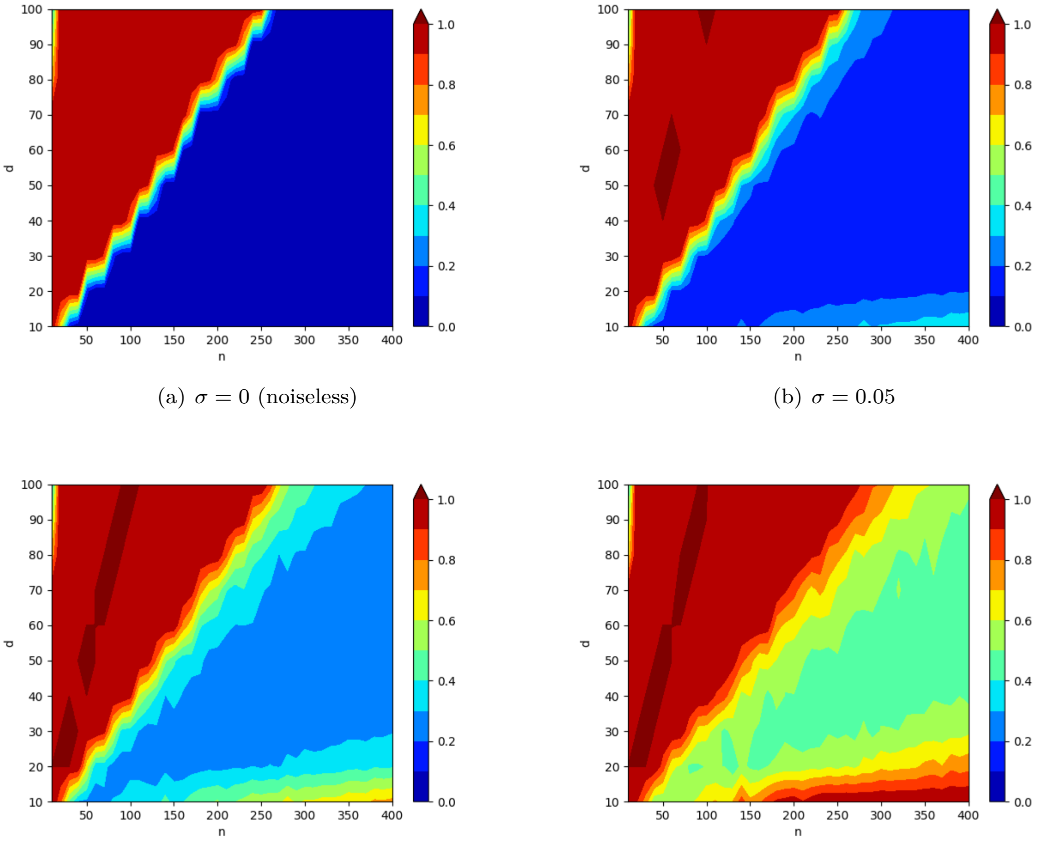
Then the test distance is defined as the ℓ_2 distance between the prediction $\tilde{\mathbf{y}}$ and the ground truth $\mathbf{y}^* = \tilde{\mathbf{X}}\mathbf{w}^*$.

The boundaries of red regions in Figure 7, which represents highly unsuccessful recovery, remain around $n = 2d$ for various noise levels σ. When σ increases, the area of dark blue regions of small absolute distance/test error gradually vanishes. This implies that the linear part of the neural network no longer approximates the planted linear neuron and the gated ReLU neurons fit the noise. In Appendix M-A, we will observe the same pattern for absolute distance in Figure 22.

In the third part, we study the generalization property of ReLU networks with skip connections using convex/non-convex training methods. Results for the convex training methods are provided in Appendix M-A.

For the nonconvex training method, we solve the regularized non-convex training problem (6) with $\beta = 10^{-6}$ as an approximation of the minimum norm problem (7). We set the number of neurons to be $m = n + 1$ and train the ReLU neural network with skip connection for 400 epochs. We use the AdamW optimizer and set the weight decay to be $\beta = 10^{-6}$. We note that the nonconvex training may still reach local minimizers. Thus, the absolute distance to the planted linear neuron does not show a clear phase transition as the convex training. However, the transitions of test error generally follow the patterns of the group ℓ_1-minimization problem. In Figure 8, we show that the test error increases as n/d increases, and the rate of increase becomes sharper around $n = 2d$ (the boundary of the orange and yellow region).

As a remark to the number of neurons in the NN, by the theorems in [2], for all $m \geq n + 1$, the convex program (6) and the non-convex problem (8) have identical optimal values. Thus it is sufficient to set $m = n + 1$ neurons in the



(a) $\sigma = 0$ (noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 7: Averaged absolute distance to the planted linear neuron by solving the group ℓ_1-minimization problem (12) derived from training ReLU networks with skip connection over 5 independent trials.

NN. Readers can refer to Appendix M-D for experiments of training NNs with more neurons. The numerical results show that the test error still follow the same pattern in Figure 8.

B. Multi-neuron recovery and irrepresentability condition

In this subsection, we analyze the recovery for ReLU networks with a normalization layer. Results for single-neuron recovery can be found in Appendix M-B. Here we focus on the case where the label vector is the combination of several normalized ReLU neurons. We will test for three types of planted neurons.

- $k = 2, \mathbf{w}_1^* = \mathbf{w}^*,\ \mathbf{w}_2^* = -\mathbf{w}^*$, where $\mathbf{w}^* \sim \mathcal{U}(\mathbb{S}^{n-1})$. In this case, the hyperplane arrangements of two neurons do not intersect.
- $k = 2, \mathbf{w}_1^*, \mathbf{w}_2^* \sim \mathcal{U}(\mathbb{S}^{n-1})$. It is a general case where the hyperplane arrangements of two neurons can intersect.
- $k \geq 2, \mathbf{w}_i^* = \mathbf{e}_i$, where $\mathbf{e}_i$ is the i-th standard basis in $\mathbb{R}^n$.

We consider the noisy observation model, i.e., the observation is the combination of several normalized ReLU neurons and Gaussian noise, i.e.,

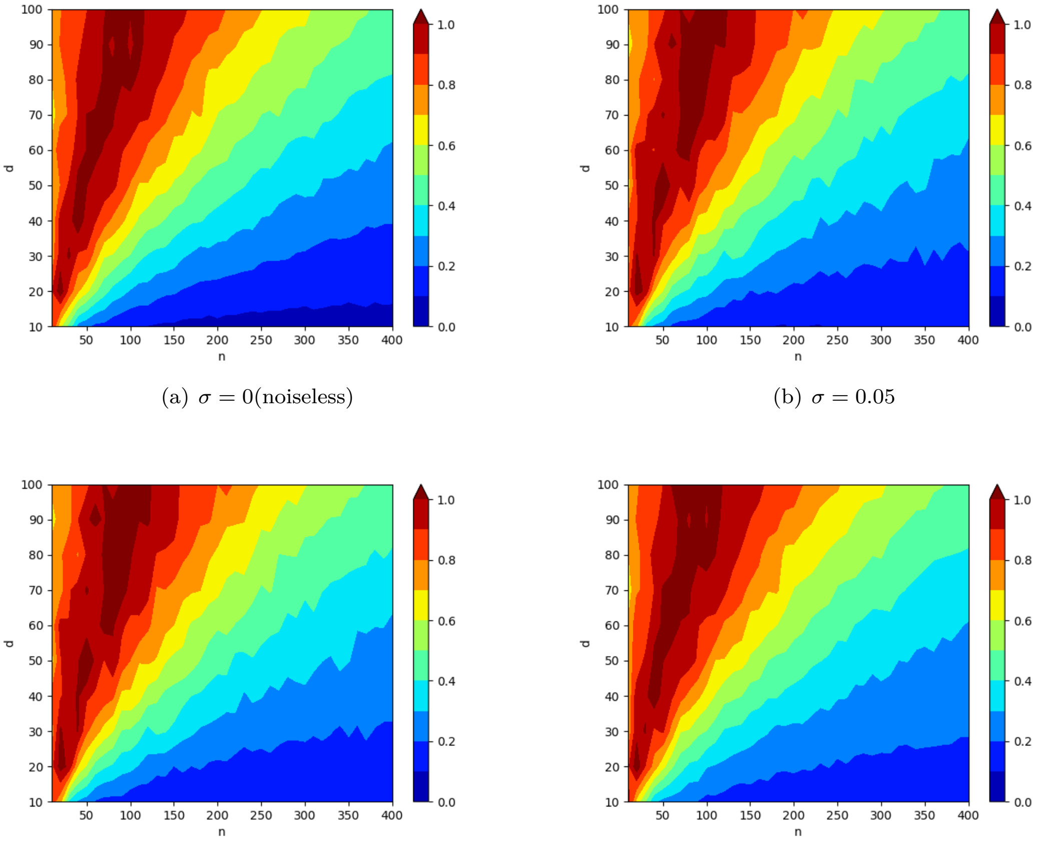
$$\mathbf{y} = \sum_{i=1}^{k} \frac{(\mathbf{X}\mathbf{w}_i^*)_+}{\|(\mathbf{X}\mathbf{w}_i^*)_+\|_2} + \mathbf{z},$$

where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2/n)$.

As a performance metric, the absolute distance is defined as $\left(\sum_{i=1}^{k} \|\mathbf{w}_{s_i} - \tilde{\mathbf{w}}_i^*\|_2^2\right)^{1/2}$.

In Figure 9, we show the phase transition graph when the planted neurons satisfy $\mathbf{w}_1^* = \mathbf{w}^*,\ \mathbf{w}_2^* = -\mathbf{w}^*,\ \mathbf{w}^* \sim \mathcal{U}(\mathbb{S}^{n-1})$. Results for the other two cases can be found in Appendix M-C.

It is also worth noting that we check the NIC-k given in (NIC-k), which guarantees recovery for the convex problem (13) and (14). We compute the probability that the NIC holds for d ranging from 10 to 80 and n ranging from 10 to



(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 8: Averaged test error by training ReLU networks with skip connection on the regularized non-convex problem (6) over 10 independent trials.

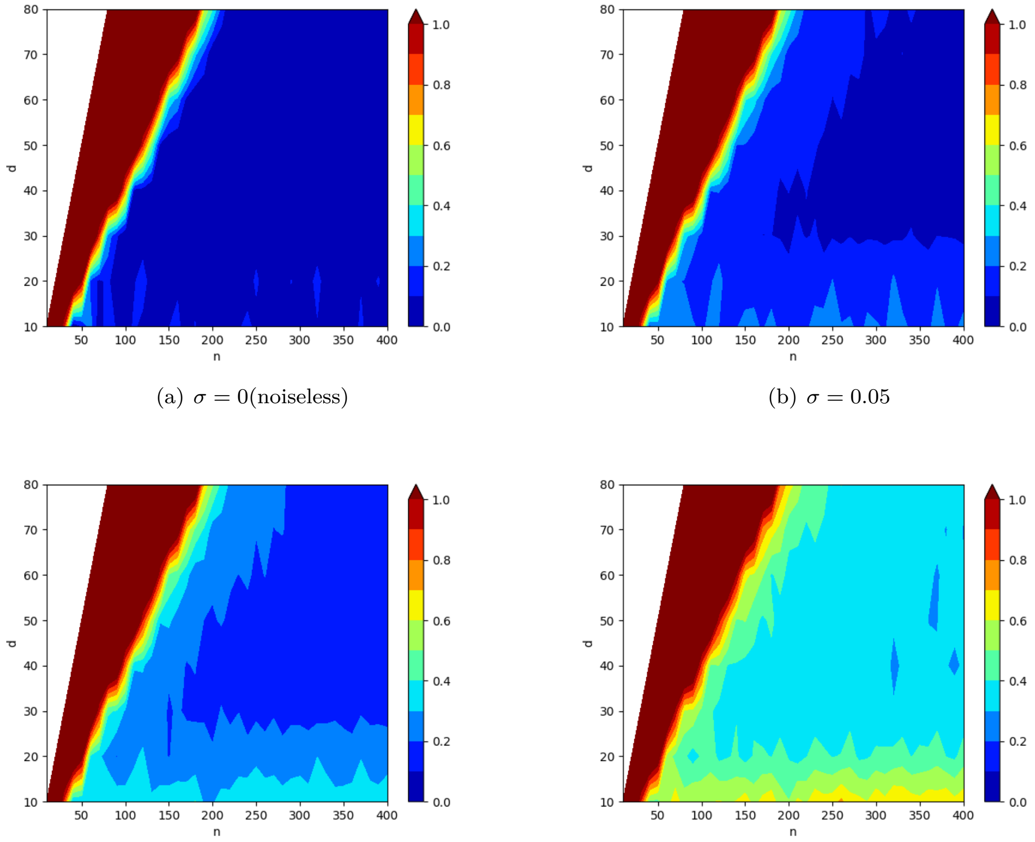
400 with 50 independent trials. For each pair of (n, d). For each data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, we use a random subset H' of the set $H\backslash H_S$ to validate the inequality (NIC-k), where H is the set of all possible hyperplane arrangements and H_S is the set of hyperplane arrangements generated by the planted neurons, i.e.

$$H_S = \{\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}_i^*))|i \in [k]\}.$$

In numerical experiments, we generate random hyperplane arrangements. If (NIC-k) holds for all $\mathbf{D}_j \in H'$, we say $\mathbf{X}$ satisfies the NIC-k numerically.

VIII. Conclusion

We presented a framework to analyze the recovery properties of ReLU neural networks through convex reparameterizations. We introduced Neural Isometry Conditions, which are deterministic conditions on the training data that ensure the recovery of planted neurons via training two-layer ReLU networks and the variants with skip connection and normalization layers. Viewing the non-convex neural network training problem from a convex optimization perspective, we establish theorems analogous to sparse recovery and compressed sensing by using probabilistic methods. For randomly generated training data matrices, we showed the existence of a sharp phase transition in the recovery of simple planted models. Interestingly, ReLU neural networks with an arbitrary number of neurons exactly recover simple planted models, such as a combination of a few ReLU neurons, when the number of samples exceeds a critical threshold. Therefore, these models can perfectly generalize even with an extremely large number of parameters when the labels are generated by simple models. This phenomenon not only aligns with the results developed in sparse recovery theory but is also validated by our numerical experiments. Namely, when the training data is i.i.d. Gaussian and the number of data points is smaller than a critical threshold, the convex program cannot recover the planted model. On the other hand, when the number of data points is above a critical threshold, the solution of the convex optimization problem uniquely recovers the planted solution. Our main contribution is that we explicitly characterize



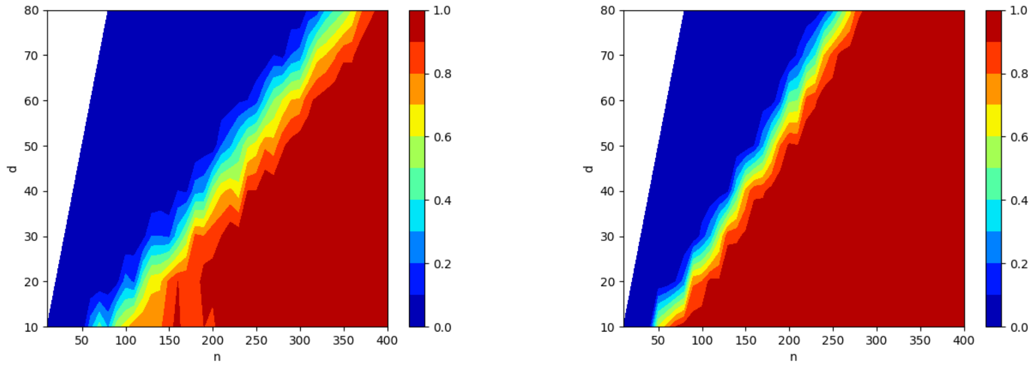
(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 9: Averaged absolute distance to the planted normalized ReLU neurons by solving the convex problem (14) from training ReLU networks with normalization layer over 5 independent trials. Here we set $k = 2$ planted neurons which satisfy $\mathbf{w}_1^* = \mathbf{w}^*,\ \mathbf{w}_2^* = -\mathbf{w}^*,\ \mathbf{w}^* \sim \mathcal{U}(\mathbb{S}^{n-1})$.

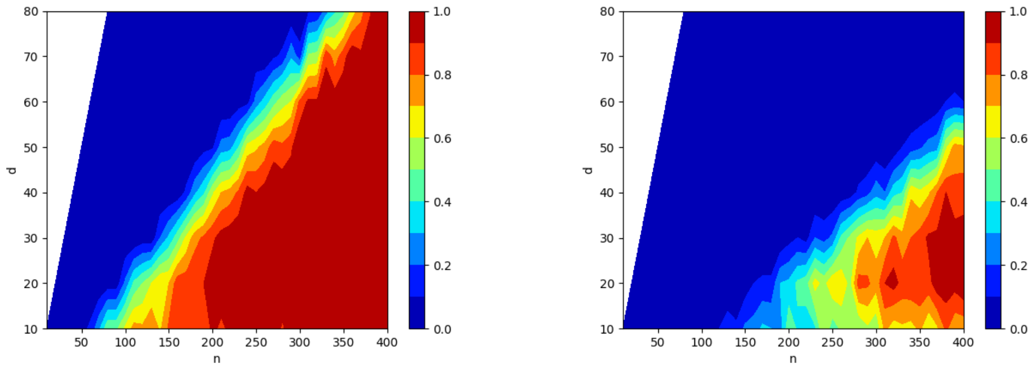


(a) $k = 2,\ \mathbf{w}_1^* = \mathbf{w}^*,\ \mathbf{w}_2^* = -\mathbf{w}^*,\ \mathbf{w}^* \sim \mathcal{U}(\mathbb{S}^{n-1})$

(b) $k = 2,\ \mathbf{w}_1^*, \mathbf{w}_2^* \sim \mathcal{U}(\mathbb{S}^{n-1})$

the data isometry conditions that imply exact recovery of planted ReLU and linear neurons, and the specific relation between the number of data points and problem dimensions for random data matrices for successful recovery with high probability. We also extend our results to the case where the observation is noisy and show that the neural network can still learn a simple model, even with an arbitrary number of neurons, when the noise component is not too large and the regularization parameter lies in an appropriate interval.

An immediate open problem is extending our results to neural networks of depth greater than two, and investigating



(c) $k=2,\ \mathbf{w}_1^*=\mathbf{e}_1,\ \mathbf{w}_2^*=\mathbf{e}_2,$

(d) $k=3,\ \mathbf{w}_i^*=\mathbf{e}_i(i=1,2,3)$

Fig. 10: The probability that the irrepresentability condition (NIC-k) holds for the group ℓ_1-minimization problem (14) over 50 independent trials. The label vector $\mathbf{y}$ is the combination of several normalized ReLU neurons.

modern DNN structures such as convolutional and transformer layers. Moreover, the analysis of the exact recovery threshold for an arbitrary number of ReLU neurons is an important open problem. Our numerical experiments suggest the trend $n \geq 2k$ for recovering k ReLU neurons when the training matrix is composed of i.i.d. Gaussian random variables. Finally, we note that non-convex training methods might get stuck at local minimizers or stationary points that are avoided by the convex formulation. The relationship between the convex and non-convex training process should be further revealed to explain this phenomenon. We leave the analysis of the non-convex training process when the observation is derived from a simple model as an open research problem for future work.

Appendix A
Review of linear sparse recovery via ℓ_1 minimization

We briefly review conditions required to ensure recovery via ℓ_1-norm minimization in the linear observation case. Suppose that $\mathbf{y} = \mathbf{X}\mathbf{w}^*$ denotes linear observations, where the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ represents measurements and $\mathbf{w}^* \in \mathbb{R}^d$ is a k-sparse vector of dimension d. Consider the following ℓ_1-minimization problem to recover $\mathbf{w}^*$ from measurements $\mathbf{y}$:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\|_1 \text{ s.t. } \mathbf{X}\mathbf{w} = \mathbf{y}. \tag{29}$$

It can be shown that $\mathbf{w}^*$ can be exactly recovered from $\mathbf{y}$ when the data matrix $\mathbf{X}$ satisfies certain isometry conditions, which are analogous to the ones developed in this work. The KKT optimality conditions of the above convex optimization problem are

$$\exists \boldsymbol{\lambda} \in \mathbb{R}^n \qquad \text{s.t.} \qquad \begin{aligned} &|\boldsymbol{\lambda}^T \mathbf{x}_j^{\text{col}}| \le 1, && \text{for } w_j = 0,\\ &\boldsymbol{\lambda}^T \mathbf{x}_j^{\text{col}} = \text{sign}(w_j), && \text{for } w_j \neq 0, \end{aligned} \tag{30}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ is a dual variable and $\mathbf{x}_j^{\text{col}} \in \mathbb{R}^d$ denotes the j-th column of $\mathbf{X}$.

Let $S = \{i \in [d] \,|\, w_i^* \neq 0\}$ denote the support set of w^* and $S^c = [n]/S$ its complement. For the support set $S \subseteq [d]$ of size $|S|$, denote the subvector of $\mathbf{w}^*$ that corresponds to entries restricted to S as $\mathbf{w}_S = (w_i)_{i\in S} \in \mathbb{R}^{|S|}$, and the submatrix $\mathbf{X}_S \in \mathbb{R}^{n\times|S|}$ of $\mathbf{X}$ formed with the columns S that correspond to the support of $\mathbf{w}$. The irrepresentability condition is a simpler sufficient condition that implies that the KKT conditions in (30) hold for $\mathbf{w} = \mathbf{w}^*$. This is ensured by the choice $\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}:\, \mathbf{X}_S^T\boldsymbol{\lambda} = \text{sign}(\mathbf{w}_S^*)} \|\boldsymbol{\lambda}\|_2 = \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\text{sign}(\mathbf{w}_S^*)$ assuming that $\mathbf{X}_S^T\mathbf{X}_S$ is invertible, and leads to the condition

$$\|\mathbf{X}_{S^c}^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\text{sign}(\mathbf{w}_S^*)\|_\infty < 1. \tag{31}$$

Intuitively, the above condition is expected to hold under three conditions: (i) the matrix $(\mathbf{X}_S^T\mathbf{X}_S)^{-1}$ is well-conditioned, (ii) the columns of $\mathbf{X}$ have small inner-products with each other, i.e., $\|\mathbf{X}_{S^c}^T\mathbf{X}_S\|_{\infty,\infty}$ is small, and (iii) the size of the subset S is not too large. Moreover,the irrepresentability condition in (31) also ensures that $\mathbf{w}^*$ is the unique optimal solution to (29) [7].

The Restricted Isometry Property (RIP) is a stronger condition that imposes well-conditioning of submatrices uniformly over all size-k subsets. RIP is stated as follows

$$(1-\delta_k)\|\mathbf{w}_S\|_2^2 \le \|\mathbf{X}_S\mathbf{w}_S\|_2^2 \le (1+\delta_k)\|\mathbf{w}_S\|_2^2, \qquad \forall \mathbf{w}_S \in \mathbb{R}^{|S|}, \quad \forall S \subseteq [d] \,:\, |S| \le k$$

where $\delta_k \in (0,1)$ is the Restricted Isometry Constant for some positive integer k. RIP implies that all $n\times|S|$ submatrices of $\mathbf{X} \in \mathbb{R}^{n\times d}$ are well-conditioned for all subsets S of size at most k. Examples of matrices that satisfy the RIP property include i.i.d. sub-Gaussian random matrices, random Haar matrices, as well subsampled orthonormal systems, e.g., Fourier and Hadamard matrices under conditions on the dimensions n and d, see details in [9, 47]. In [48], it has been shown that RIP with a sufficiently small constant δ_k implies the irrepresentability condition, and hence the recovery of a linear k-sparse vector from the observations $\mathbf{y} = \mathbf{X}\mathbf{w}^*$.

Appendix B
Permutation and splitting of neural networks

In this section, we present the definition of permutation and splitting of two-layer ReLU neural networks and their variants with skip connections or normalization layers.

For a ReLU network $\Theta = (\mathbf{W}_1, \mathbf{w}_2)$ where $\mathbf{W}_1 \in \mathbb{R}^{d\times m}, \mathbf{w}_2 \in \mathbb{R}^m$, a permutation of Θ is any neural network $\Theta' = (\mathbf{W}_1', \mathbf{w}_2')$ with $\mathbf{W}_1' \in \mathbb{R}^{d\times m}, \mathbf{w}_2' \in \mathbb{R}^m$ such that $\mathbf{w}_{1,j}' = \mathbf{w}_{1,\pi(j)}$ and $w_{2,j}' = w_{2,\pi(j)}$ for $j \in [m]$. Here $\pi : [m] \to [m]$ is a permutation of $[m]$.

Given a neuron-pair $(\mathbf{w}_1, w_2)$, we say that a collection of neuron-pairs $\{(\mathbf{w}_{1,j}, w_{2,j})\}_{j=1}^k$ is a splitting of $(\mathbf{w}_1, w_2)$ if $(\mathbf{w}_{1,j}, w_{2,j}) = (\sqrt{\gamma_j}\mathbf{w}_1, \sqrt{\gamma_j}w_2)$ for some $\gamma_j \ge 0$ and $\sum_{j=1}^k \gamma_j = 1$. Given a ReLU neural network $\Theta = (\mathbf{W}_1, \mathbf{w}_2)$ with $\mathbf{W}_1 \in \mathbb{R}^{d\times m}, \mathbf{w}_2 \in \mathbb{R}^m$, a *splitting* of Θ is any neural network $\Theta' = (\mathbf{W}_1', \mathbf{w}_2')$ with $\mathbf{W}_1' \in \mathbb{R}^{d\times m'}, \mathbf{w}_2' \in \mathbb{R}^{m'}$ such that the non-zero neurons of Θ' can be partitioned into splittings of the neurons of Θ.

For a ReLU network with skip connection $\Theta = (\mathbf{W}_1, \mathbf{w}_2)$ where $\mathbf{W}_1 \in \mathbb{R}^{d\times m}, \mathbf{w}_2 \in \mathbb{R}^m$, the permutation and splitting of this network refers to the permutation and splitting of the ReLU neurons $(\mathbf{w}_{1,i}, w_{2,i})_{i=2}^m$ in this network.

For a ReLU network with normalization layer $\Theta = (\mathbf{W}_1, \mathbf{w}_2, \boldsymbol{\alpha})$ where $\mathbf{W}_1 \in \mathbb{R}^{d\times m}, \mathbf{w}_2, \boldsymbol{\alpha} \in \mathbb{R}^m$, a permutation of Θ is any neural network $\Theta' = (\mathbf{W}_1', \mathbf{w}_2', \boldsymbol{\alpha}')$ such that $\mathbf{w}_{1,j}' = \mathbf{w}_{1,\pi(j)}$, $w_{2,j}' = w_{2,\pi(j)}$ and $\alpha_j' = \alpha_{\pi(j)}$ for $j \in [m]$. Here $\pi : [m] \to [m]$ is a permutation of $[m]$.

Given a neuron-pair $(\mathbf{w}_1, w_2, \alpha)$, we say that a collection of neuron-pairs $\{(\mathbf{w}_{1,j}, w_{2,j})\}_{j=1}^k$ is a splitting of $(\mathbf{w}_1, w_2)$ if $(w_{2,j}, \alpha_j) = (\sqrt{\gamma_j}w_2, \sqrt{\gamma_j}\alpha)$ for some $\gamma_j \ge 0$ and $\sum_{j=1}^k \gamma_j = 1$ and $\mathbf{w}_{1,j}$ is positively colinear with $\mathbf{w}_1$ for $j \in [m]$. Namely, there exists $\zeta_j > 0$ such that $\mathbf{w}_{1,j} = \zeta_j\mathbf{w}_{1,j}$. Given a ReLU neural network with normalization layer

$\Theta = (\mathbf{W}_1, \mathbf{w}_2, \boldsymbol{\alpha})$ with $\mathbf{W}_1 \in \mathbb{R}^{d\times m}, \mathbf{w}_2, \boldsymbol{\alpha} \in \mathbb{R}^m$, a *splitting* of Θ is any neural network $\Theta' = (\mathbf{W}_1', \mathbf{w}_2', \boldsymbol{\alpha}')$ with $\mathbf{W}_1' \in \mathbb{R}^{d\times m'}, \mathbf{w}_2', \boldsymbol{\alpha}' \in \mathbb{R}^{m'}$ such that the non-zero neurons of Θ' can be partitioned into splittings of the neurons of Θ.

Appendix C
Recovering a linear model using ReLU NN

For a linear model $y = \mathbf{X}\mathbf{w}$, it can be decomposed into a ReLU NN with two neurons

$$y = \max(0, \mathbf{X}\mathbf{w}) - \max(0, -\mathbf{X}\mathbf{w}).$$

This can be addressed theoretically by Proposition 8 in the asymptotic setting (by adding an additional normalization layer). In practice, we note that recovering a linear model using only ReLUs needs a significantly larger number of observations n compared to the case of using ReLU NNs with skip connections. In the revision, we added new experiments and figures to illustrate this. For instance, in Figure 11 below, with $n = 5$ and $d = 2$, the ReLU NN cannot recover the linear model, while the ReLU NN with skip connection recovers, leading to zero test error with high probability as in Figure 12. Moreover, we present the average test error with respect to the number of samples n in Figure 13. This figure shows that the ReLU network without skip connection can not reliably recover the linear model with high probability. On the other hand, we observe a clear phase transition for the model with skip connection, indicating that it recovers the linear model exactly in a certain regime. We believe that this observation justifies our focus on the ReLU NNs with skip connections.

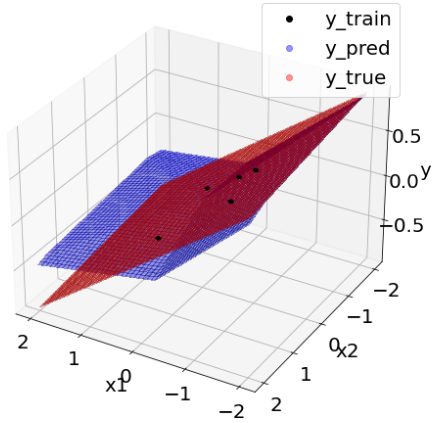


Fig. 11: ReLU NN.

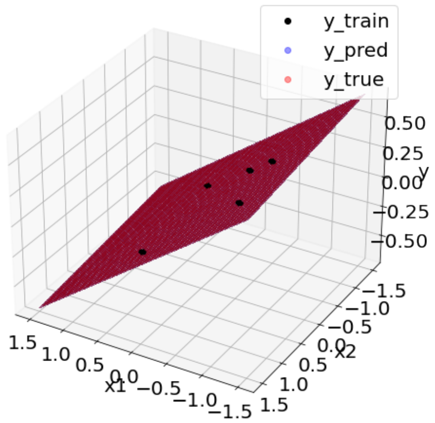


Fig. 12: ReLU NN with skip connection.

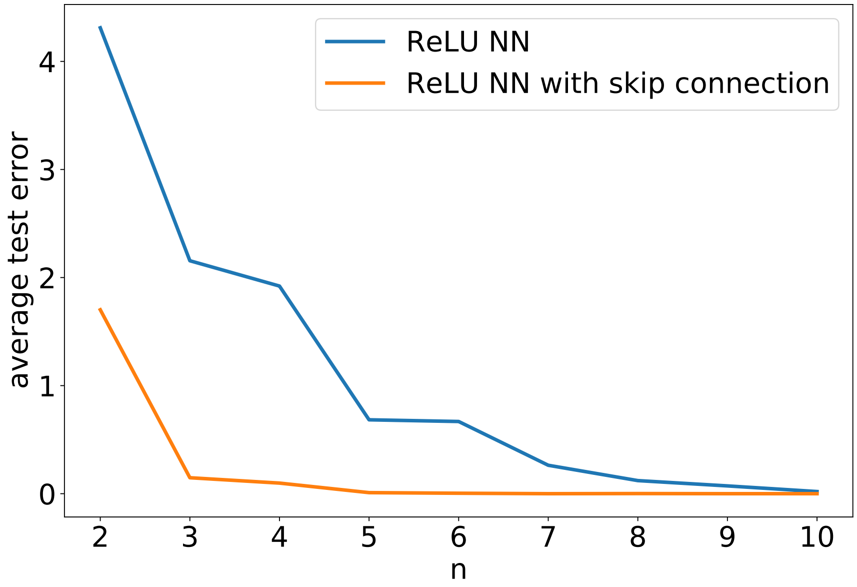


Fig. 13: Comparison between ReLU NN with/without skip connection. We take $\mathbf{X} \sim \mathcal{N}(0, 1/n)$, $\mathbf{w} \sim \mathcal{N}(0, 1)$, $\mathbf{y} = \mathbf{X}\mathbf{w}$ and $\mathbf{X}_{\text{test}}$ as the uniform grid over $[-3, 3]^2$. We plot the average test error over 100 independent trials.

Appendix D
Justification for excluding the hyperplane arrangement induced by the zero vector

Consider $\tilde{H} = \{\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0)) | \mathbf{h} \in \mathbb{R}^d\}$. Firstly, we note that $\mathbf{I}_n = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))$. If $\mathbf{I}_n \in H$, then we have $\tilde{H} = H$. If $\mathbf{I}_n \notin H$, then this implies that $\{\mathbf{w} \in \mathbb{R}^d : (2\mathbf{I}_n - \mathbf{I}_n)\mathbf{X}\mathbf{w} \geq 0\} = \{0\}$. Therefore, excluding the hyperplane arrangement induced by $\mathbf{h} = 0$ does not change the convex program (8).

Appendix E
The convex program for gated ReLU networks

Consider the minimum norm interpolation problem

$$\min_{\Theta} \frac{1}{2} \left(\|\mathbf{W}_1\|_F^2 + \|\mathbf{w}_2\|_2^2 \right), \text{ s.t. } f^{\text{gReLU}}(\mathbf{X}; \Theta) = \mathbf{y}, \tag{32}$$

where $f^{\text{gReLU}}(\mathbf{X}; \Theta)$ defined in (10) is the output of a gated ReLU network. We first reformulate (32) in the following way.

Proposition 10 *The problem* (32) *can be reformulated as*

$$\min_{\Theta} \|\mathbf{w}_2\|_1, \ \textit{s.t. } f^{\text{gReLU}}(\mathbf{X}; \Theta) = \mathbf{y}, \|\mathbf{w}_{1,i}\|_2 \leq 1, i \in [m]. \tag{33}$$

For the reformulated problem (33), we can derive the dual problem.

Proposition 11 *The dual problem of* (33) *is given by*

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y}, \ \textit{s.t. } \max_{\|\mathbf{w}\|_2 \leq 1, \mathbf{h} \neq 0} \left| \boldsymbol{\lambda}^T \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0)) \mathbf{X}\mathbf{w} \right| \leq 1. \tag{34}$$

Based on the hyperplane arrangement described in (2), the dual problem is also equivalent to

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y}, \text{ s.t. } \max_{j \in [p], \|\mathbf{w}\|_2 \leq 1} \left| \boldsymbol{\lambda}^T \mathbf{D}_j \mathbf{X}\mathbf{w} \right| \leq 1 = \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T \mathbf{y}, \text{ s.t. } \left\| \mathbf{X}^T \mathbf{D}_j \boldsymbol{\lambda} \right\|_2 \leq 1, j \in [p]. \tag{35}$$

Indeed, the bi-dual problem (dual of the dual problem (34)) is the group Lasso problem (9).

Proposition 12 *The dual of the problem* (34) *is the group Lasso problem* (9)*. For a sufficiently large m, the minimum norm interpolation problem* (32) *is equivalent to* (9)*.*

A. Proof of Proposition 10

Proof For $i \in [m]$, consider $\hat{\mathbf{w}}_{1,i} = \alpha_i \mathbf{w}_{1,i}$ and $\hat{w}_{2,i} = \alpha_i^{-1} w_{2,i}$, where $\alpha_i > 0$. Let $\hat{\Theta} = (\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2, \mathbf{H})$. Then, we note that $f^{\mathrm{gReLU}}(\mathbf{X};\Theta) = f^{\mathrm{gReLU}}(\mathbf{X};\hat{\Theta})$. This implies that $\hat{\Theta}$ is feasible for (32). From the inequality of arithmetic and geometric mean, we note that

$$\frac{1}{2}\sum_{i=1}^{m}\left(\alpha_i^2\|\mathbf{w}_{1,i}\|_2^2 + \alpha_i^{-2} w_{2,i}^2\right) \geq \sum_{i=1}^{m}\|\mathbf{w}_{1,i}\|_2 |w_{2,i}|. \tag{36}$$

The equality is achieved when $\alpha_i = \sqrt{\frac{|w_{2,i}|}{\|\mathbf{w}_{1,i}\|_2}}$. As the scaling operation does not change $\sum_{i=1}^m \|\mathbf{w}_{1,i}\|_2 |w_{2,i}|$, we can set $\|\mathbf{w}_{1,i}\|_2 = 1$ and then the lower bound of the objective value becomes $\sum_{i=1}^m |w_{2,i}| = \|\mathbf{w}_2\|_1$. This completes the proof.

B. Proof of Proposition 11

Consider the Lagrangian function

$$L(\mathbf{W}_1, \mathbf{w}_2, \mathbf{H}, \boldsymbol{\lambda}) = \|\mathbf{w}_2\|_1 + \boldsymbol{\lambda}^T\left(\mathbf{y} - \sum_{i=1}^{m} \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_i \geq 0))\mathbf{X}\mathbf{w}_{1,i} w_{2,i}\right). \tag{37}$$

The problem (32) is equivalent to

$$\begin{aligned}
&\min_{\mathbf{W}_1,\mathbf{w}_2,\mathbf{H}} \max_{\boldsymbol{\lambda}} L(\mathbf{W}_1, \mathbf{w}_2, \mathbf{H}, \boldsymbol{\lambda}), \text{ s.t. } \|\mathbf{w}_{1,i}\|_2 \leq 1, \mathbf{h}_i \neq 0, i \in [m]\\
=&\min_{\mathbf{W}_1,\mathbf{H}} \max_{\boldsymbol{\lambda}} \min_{\mathbf{w}_2} L(\mathbf{W}_1, \mathbf{w}_2, \mathbf{H}, \boldsymbol{\lambda}), \text{ s.t. } \|\mathbf{w}_{1,i}\|_2 \leq 1, \mathbf{h}_i \neq 0, i \in [m]\\
=&\min_{\mathbf{W}_1,\mathbf{H}} \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\mathbf{y} - \sum_{i=1}^{m} \tilde{\mathbb{I}}\left(\left|\boldsymbol{\lambda}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_i \geq 0))\mathbf{X}\mathbf{w}_{1,i}\right| \leq 1\right),\\
&\text{s.t. } \|\mathbf{w}_{1,i}\|_2 \leq 1, \mathbf{h}_i \neq 0, i \in [m].
\end{aligned} \tag{38}$$

Here $\tilde{\mathbb{I}}(S) = 0$ if the statement S is correct and $\tilde{\mathbb{I}}(S) = +\infty$ otherwise. By exchanging the order of min and max, we obtain the dual problem

$$\begin{aligned}
&\max_{\boldsymbol{\lambda}} \min_{\mathbf{W}_1,\mathbf{H}} \boldsymbol{\lambda}^T\mathbf{y} - \sum_{i=1}^{m} \tilde{\mathbb{I}}\left(\left|\boldsymbol{\lambda}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_i \geq 0))\mathbf{X}\mathbf{w}_{1,i}\right| \leq 1\right), \text{ s.t. } \|\mathbf{w}_{1,i}\|_2 \leq 1, \mathbf{h}_i \neq 0, i \in [m]\\
=&\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\mathbf{y}, \text{ s.t. } \max_{\|\mathbf{w}_{1,i}\|_2 \leq 1, \mathbf{h}_i \neq 0} \left|\boldsymbol{\lambda}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}_i \geq 0))\mathbf{X}\mathbf{w}_{1,i}\right| \leq 1, i \in [m]\\
=&\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\mathbf{y}, \text{ s.t. } \max_{\|\mathbf{w}\|_2 \leq 1, \mathbf{h} \neq 0} \left|\boldsymbol{\lambda}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\mathbf{X}\mathbf{w}\right| \leq 1.
\end{aligned} \tag{39}$$

C. Proof of Proposition 12

Proof As the group lasso problem (9) is a convex problem, it is sufficient to show that the dual problem of (9) is exactly (34). Consider the Lagrangian function

$$L(\mathbf{w}_1, \ldots, \mathbf{w}_p, \boldsymbol{\lambda}) = \sum_{j=1}^{p} \|\mathbf{w}_j\|_2 + \boldsymbol{\lambda}^T(\mathbf{y} - \sum_{j=1}^{p} \mathbf{D}_j\mathbf{X}\mathbf{w}_j). \tag{40}$$

The dual problem follows

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_p} L(\mathbf{w}_1, \ldots, \mathbf{w}_p, \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\mathbf{y}, \text{ s.t. } \left\|\mathbf{X}^T\mathbf{D}_j\boldsymbol{\lambda}\right\|_2 \leq 1, j \in [p], \tag{41}$$

which is equivalent to (34). This implies that the optimal value of (9) serves as a lower bound for (32). For a sufficiently large m, any feasible point $(\mathbf{w}_1, \ldots, \mathbf{w}_p)$ to (9) also corresponds to a feasible neural network for (32). In this case, the minimum norm interpolation problem (32) is equivalent to (9).

Appendix F
Proofs in Section II

A. Proof of Lemma 1

Proof Let m^* be the minimal number of neurons of the optimal solution to the convex program (3). Similar to the proof of Theorem 1 in [6], assuming that $m \geq m^*$, for any globally optimal solution to the non-convex problem (7), we can merge its ReLU neurons into a minimal neural network defined in [6]. This minimal neural network combining with the linear part corresponds to an optimal solution to the convex program (3). Therefore, we can view this globally optimal neural network as a possibly split and permuted version of an optimal solution to the convex program (3). For the case where $\mathbf{y} = \mathbf{X}\mathbf{w}^*$, the ReLU network with skip connection can represent $\mathbf{y}$ using the linear part. In this case, $m^* = 1$.

B. Proof of Proposition 1

Proof We first show that the linear neural isometry condition implies the recovery of the planted linear model by solving (12).

Proposition 13 *Suppose that the linear neural isometry condition* (NIC-L) *holds. Consider* $\mathbf{W}^* = (\mathbf{w}_0^*, \dots, \mathbf{w}_p^*)$ *such that* $\mathbf{w}_0^* = \mathbf{w}^*$ *while* $\mathbf{w}_j^* = 0$ *for* $j \neq 0$*. Then,* $\mathbf{W}^*$ *is the unique optimal solution to* (12)*.*

Proof The KKT conditions for (12) consist of

$$\begin{aligned} &\left\|\mathbf{X}^T\mathbf{D}_j\boldsymbol{\lambda}^T\right\| \leq 1, && \text{if } \mathbf{w}_j = 0,\\ &\mathbf{X}^T\mathbf{D}_j\boldsymbol{\lambda}^T = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}, && \text{if } \mathbf{w}_j \neq 0,\\ &\sum_{j=0}^p \mathbf{D}_j\mathbf{X}\mathbf{w}_j = \mathbf{y}. \end{aligned} \tag{42}$$

As a direct corollary of Proposition 14, $\mathbf{W}^*$ is the unique optimal solution to (12).

We then present the proof of Proposition 1. As (12) is derived by dropping all inequality constraints in (3), the optimal value of (8) is lower bounded by (12). Note that $\mathbf{W}^*$ is the unique optimal solution to (12). Hence, the optimal value of (12) is $\|\mathbf{w}^*\|_2$. On the other hand, $\hat{\mathbf{W}}$ is feasible for (3) and it leads to an objective value of $\|\mathbf{w}^*\|_2$. This implies that $\hat{\mathbf{W}}$ is an optimal solution to (3) and the optimal values of (3) and (12) are the same. Suppose that we have another solution $\bar{\mathbf{W}}$ which is optimal to (3). Consider $\breve{\mathbf{W}} = \{\breve{\mathbf{w}}_j | j = 0, 1, \dots, p\}$ where $\breve{\mathbf{w}}_0 = \tilde{\mathbf{w}}_0$ and $\breve{\mathbf{w}}_j = \bar{\mathbf{w}}_j - \bar{\mathbf{w}}_j'$ for $j = 1, \dots, p$. Then, $\bar{\mathbf{W}}$ is also optimal to (12). As $\mathbf{W}^*$ is the unique optimal solution to (9), this implies that $\bar{\mathbf{W}} = \mathbf{W}^*$. Hence, we have $\bar{\mathbf{w}}_0 = \bar{\mathbf{w}}_0 = \mathbf{w}^*$. For $j = 1, 2, \dots, p$, we have

$$\|\bar{\mathbf{w}}_j\|_2 + \|\bar{\mathbf{w}}_j'\|_2 \geq 0 = \|\mathbf{w}_j^*\|_2. \tag{43}$$

The equality holds when $\bar{\mathbf{w}}_j = \bar{\mathbf{w}}_j' = 0$. As $\bar{\mathbf{W}}$ is optimal to (3), we have $\bar{\mathbf{w}}_j = \bar{\mathbf{w}}_j' = 0$ for $j = 1, \dots, p$. This implies that $\bar{\mathbf{W}} = \hat{\mathbf{W}}$. Thus, $\hat{\mathbf{W}}$ is the unique optimal solution.

Appendix G
Proofs in Section III

A. Proof of Theorem 3

Proof Let m^* be the minimal number of neurons of the optimal solution to the convex program (8). In the case there are multiple optimal solutions, we may take the one with minimal cardinality. We can view the minimal norm problem (7) as

$$\min_{\Theta} \ell(f(\mathbf{X}; \Theta); \mathbf{y}) + R(\Theta), \tag{44}$$

where the loss function is defined by

$$\ell(\mathbf{z}; \mathbf{y}) = \begin{cases} 0, & \mathbf{z} = \mathbf{y},\\ +\infty, & \mathbf{z} \neq \mathbf{y}. \end{cases}$$

Hence, by applying Theorem 1 in [6], for $m \geq m^*$, any globally optimal solutions of the non-convex problem (7) of ReLU networks can be computed via the optimal solutions of the convex program (8) up to splitting and permutation. For the case $\mathbf{y} = (\mathbf{X}\mathbf{w}^*)_+$, we have $m^* = 1$.

B. Proof of Theorem 4

Proof Because $\mathbf{y} = \frac{\mathbf{X}\mathbf{w}^*}{\|\mathbf{X}\mathbf{w}^*\|_2}$, the ReLU network with skip connection can represent $\mathbf{y}$ using $m^* = 1$ neurons. Let $\mathbf{T}_1, \dots, \mathbf{T}_q$ be the enumeration of all possible diagonal arrangement patterns

$$\{\mathbf{diag}(\text{sign}(\mathbf{X}\mathbf{w})) | \mathbf{w} \in \mathbb{R}^d, \mathbf{w} \neq 0\}.$$

We denote Q_i be closed convex cone of solution vectors for $\text{sign}(\mathbf{X}\mathbf{w}) = \mathbf{T}_i$ for $i \in [q]$. Let $B_i = Q_i \times \mathbb{R}_{>0}$ and $B_{i+q} = Q_i \times \mathbb{R}_{<0}$ for $i \in [q]$. Similar to the definition of minimal neural networks in [6], we can define the minimal neural networks with normalization layer as follows:

- We say that a ReLU neural network with normalization layer $\Theta = (\mathbf{W}_1, \mathbf{w}_2, \alpha)$ is minimal if (i) it is scaled, i.e., $|w_{2,i}| = |\alpha_i|$ (ii) for each cone B_i where $i \in [2q]$, the minimal neural network has at most a single non-zero neuron $(\mathbf{w}_{1,j}, w_{2,j}, \alpha_j)$ such that $(\mathbf{w}_{1,j}, w_{2,j}\alpha_j) \in B_i$.

Similar to the proof of Theorem 1 in [6], for any globally optimal solution to the non-convex problem (7), we can merge it into a minimal neural network with normalization layer. This minimal neural network corresponds to an optimal solution to the convex program (13). Therefore, we can view this globally optimal neural network as the split and permuted version of an optimal solution to the convex program (13).

Appendix H
Proofs in Section IV

To begin with, consider a general group ℓ_1-minimization problem

$$\min_{\{\mathbf{w}_j\}_{j=1}^k} \quad \sum_{j=1}^k \left\|\mathbf{w}_j\right\|_2, \text{s.t.} \quad \sum_{j=1}^k \mathbf{A}_j\mathbf{w}_j = \mathbf{y}, \tag{45}$$

where $\mathbf{A}_j \in \mathbb{R}^{n\times r_j}$ and $\mathbf{w}_j \in \mathbb{R}^{r_j}$ for $j \in [k]$. The KKT condition follows

$$\begin{aligned} &\left\|\mathbf{A}_j^T\boldsymbol{\lambda}\right\|_2 \leq 1, && \text{if } \mathbf{w}_j = 0,\\ &\mathbf{A}_j^T\boldsymbol{\lambda} = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|_2}, && \text{if } \mathbf{w}_j \neq 0,\\ &\sum_{j=1}^k \mathbf{A}_j\mathbf{w}_j = \mathbf{y}, \end{aligned} \tag{46}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the dual variable. Suppose that $\mathbf{y} = \mathbf{A}_{i^*}\mathbf{w}^*$ is the label vector, where $\mathbf{w}^* \in \mathbb{R}^{s_{i^*}}$. Assume that $\mathbf{A}_{i^*}^T\mathbf{A}_{i^*}$ is invertible. Then, the irrepresentability condition follows

$$\left\|\mathbf{A}_j^T\mathbf{A}_{i^*}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*})^{-1}\frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}\right\|_2 < 1, \forall j \neq i^*. \tag{47}$$

Proposition 14 *Let* $\mathbf{y} = \mathbf{A}_{i^*}\mathbf{w}^*$*. Suppose that the irrepresentability condition* (47) *holds. Consider* $\mathbf{W}^* = (\mathbf{w}_1^*, \dots, \mathbf{w}_k^*)$ *such that* $\mathbf{w}_{i^*}^* = \mathbf{w}^*$ *while* $\mathbf{w}_j^* = 0$ *for* $j \neq i^*$*. Then,* $\mathbf{W}^*$ *is the unique optimal solution to* (45)*.*

Proof We first show that $\mathbf{W}^*$ is the optimal solution to (45). Let

$$\boldsymbol{\lambda} = \mathbf{A}_{i^*}\left(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*}\right)^{-1}\frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}.$$

We can examine that $\mathbf{A}_{i^*}^T\boldsymbol{\lambda} = \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$. For $j \neq i^*$, from the irrepresentability condition (47), we have

$$\left\|\mathbf{A}_j^T\boldsymbol{\lambda}\right\|_2 = \left\|\mathbf{A}_j^T\mathbf{A}_{i^*}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*})^{-1}\frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}\right\|_2 < 1.$$

Therefore, $(\mathbf{W}^*, \boldsymbol{\lambda})$ satisfies the KKT condition (46). This implies that $\mathbf{W}^*$ is optimal to (45).

Then, we prove the uniqueness. Suppose that $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ is another optimal solution to (9). Denote $\Delta_j = \mathbf{w}_j$ for $j \neq i^*$ and $\Delta_{i^*} = \mathbf{w}_{i^*} - \mathbf{w}_{i^*}^*$. Then, we have

$$\sum_{j\neq i^*}\mathbf{A}_j\Delta_j + \mathbf{A}_{i^*}\Delta_{i^*} = 0. \tag{48}$$

We note that

$$\begin{aligned} &\|\mathbf{w}_{i^*}\|_2 + \sum_{j\neq i^*}\|\mathbf{w}_j\|_2\\ \geq&\|\mathbf{w}_{i^*}^*\|_2 + \left(\frac{\mathbf{w}_{i^*}^*}{\|\mathbf{w}_{i^*}^*\|_2}\right)^T\Delta_{i^*} + \sum_{j\neq i^*}\|\mathbf{w}_j\|_2\\ =&\|\mathbf{w}_{i^*}^*\|_2 + \boldsymbol{\lambda}^T\mathbf{A}_{i^*}\Delta_{i^*} + \sum_{j\neq i^*}\|\Delta_j\|_2\\ =&\|\mathbf{w}_{i^*}^*\|_2 - \boldsymbol{\lambda}^T\sum_{j\neq i^*}\mathbf{A}_j\Delta_j + \sum_{j\neq i^*}\|\Delta_j\|_2\\ \geq&\|\mathbf{w}_{i^*}^*\|_2 + \sum_{j\neq i^*}\|\Delta_j\|_2\left(1 - \left\|\mathbf{A}_j^T\boldsymbol{\lambda}\right\|_2\right) \geq \|\mathbf{w}_{i^*}^*\|_2 \end{aligned} \tag{49}$$

The equality holds when $\Delta_j = 0$ for all $j \neq i^*$ and $\Delta_{i^*} = \gamma\frac{\mathbf{w}_{i^*}^*}{\|\mathbf{w}_{i^*}^*\|_2}$ for certain $\gamma \geq 0$. This implies that $\mathbf{W}^*$ is the unique optimal solution to (9).

Consider the following weak irrepresentability condition

$$\left\|\mathbf{A}_j^T\mathbf{A}_{i^*}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*})^{-1}\mathbf{w}^*\right\|_2 < 1, j \in S_1, \left\|\mathbf{A}_j^T\mathbf{A}_{i^*}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*})^{-1}\mathbf{w}^*\right\|_2 = 1, j \in S_2, \tag{50}$$

where $S_1 \cup S_2 = [k]/\{i^*\}$. We have the following results when the weak irrepresentability condition holds.

Proposition 15 *Suppose that the weak irrepresentability condition holds. Consider* $\mathbf{W}^* = (\mathbf{w}_1^*, \dots, \mathbf{w}_k^*)$ *such that* $\mathbf{w}_{i^*}^* = \mathbf{w}^*$ *while* $\mathbf{w}_j^* = 0$ *for* $j \neq i^*$*. Then,* $\mathbf{W}^*$ *is an optimal solution to* (45)*. All optimal solutions* $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ *shall satisfy* $\mathbf{w}_j = 0$ *for* $j \in S_1$*.*

Proof Similar to the proof of Proposition 14, we can show that $\mathbf{W}^*$ is the optimal solution to (45). Suppose that $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ is another optimal solution to (9). Denote $\Delta_j = \mathbf{w}_j$ for $j \neq i^*$ and $\Delta_{i^*} = \mathbf{w}_{i^*} - \mathbf{w}_{i^*}^*$. Then, we have

$$\sum_{j \neq i^*} \mathbf{A}_j \Delta_j + \mathbf{A}_{i^*} \Delta_{i^*} = 0. \tag{51}$$

We note that

$$\begin{aligned}
&\|\mathbf{w}_{i^*}\|_2 + \sum_{j \neq i^*} \|\mathbf{w}_j\|_2 \\
\geq & \|\mathbf{w}_{i^*}^*\|_2 + \left(\frac{\mathbf{w}_{i^*}^*}{\|\mathbf{w}_{i^*}^*\|_2} \right)^T \Delta_{i^*} + \sum_{j \neq i^*} \|\mathbf{w}_j\|_2 \\
= & \|\mathbf{w}_{i^*}^*\|_2 + \boldsymbol{\lambda}^T \mathbf{A}_{i^*} \Delta_{i^*} + \sum_{j \neq i^*} \|\Delta_j\|_2 \\
= & \|\mathbf{w}_{i^*}^*\|_2 - \boldsymbol{\lambda}^T \sum_{j \neq i^*} \mathbf{A}_j \Delta_j + \sum_{j \neq i^*} \|\Delta_j\|_2 \\
\geq & \|\mathbf{w}_{i^*}^*\|_2 + \sum_{j \neq i^*} \|\Delta_j\|_2 \left(1 - \left\|\mathbf{A}_j^T \boldsymbol{\lambda}\right\|_2 \right) \\
\geq & \|\mathbf{w}_{i^*}^*\|_2.
\end{aligned} \tag{52}$$

The equality holds when $\Delta_j = 0$ for all $j \in S_1$ and $\Delta_{i^*} = \gamma \frac{\mathbf{w}_{i^*}^*}{\|\mathbf{w}_{i^*}^*\|_2}$ for certain $\gamma \geq 0$. This completes the proof.

We then consider the case where $\mathbf{y} = \sum_{i=1}^{l} \mathbf{A}_{s_i} \mathbf{w}_{s_i}^*$, where $S = \{s_1, \dots, s_l\} \subseteq [k]$. Denote $\mathbf{A}_S^{\text{aug}} = \begin{bmatrix} \mathbf{A}_{s_1}^T \\ \dots \\ \mathbf{A}_{s_k}^T \end{bmatrix}$. Suppose that $\mathbf{A}_S^{\text{aug}} (\mathbf{A}_S^{\text{aug}})^T =: \begin{bmatrix} \mathbf{A}_{s_1}^T \mathbf{A}_{s_1} & \dots & \mathbf{A}_{s_1}^T \mathbf{A}_{s_k} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{s_k}^T \mathbf{A}_{s_1} & \dots & \mathbf{A}_{s_k}^T \mathbf{A}_{s_k} \end{bmatrix}$ is invertible. Then, the irrepresentability condition follows

$$\left\| \mathbf{A}_j^T (\mathbf{A}_S^{\text{aug}})^T (\mathbf{A}_S^{\text{aug}} (\mathbf{A}_S^{\text{aug}})^T)^{-1} \begin{bmatrix} \mathbf{w}_{s_1}^* / \left\|\mathbf{w}_{s_1}^*\right\|_2 \\ \vdots \\ \mathbf{w}_{s_1}^* / \left\|\mathbf{w}_{s_k}^*\right\|_2 \end{bmatrix} \right\|_2 < 1, \forall j \notin S. \tag{53}$$

or equivalently,

$$\left\| \mathbf{A}_j^T (\mathbf{A}_S^{\text{aug}})^\dagger \begin{bmatrix} \mathbf{w}_{s_1}^* / \left\|\mathbf{w}_{s_1}^*\right\|_2 \\ \vdots \\ \mathbf{w}_{s_1}^* / \left\|\mathbf{w}_{s_k}^*\right\|_2 \end{bmatrix} \right\|_2 < 1, \forall j \notin S. \tag{54}$$

Proposition 16 *Suppose that the irrepresentability condition* (53) *holds. Consider* $\hat{\mathbf{W}}^* = (\hat{\mathbf{w}}_1^*, \dots, \hat{\mathbf{w}}_k^*)$ *such that* $\hat{\mathbf{w}}_i^* = \mathbf{w}_i^*$ *for* $i \in S$ *while* $\mathbf{w}_j^* = 0$ *for* $j \notin S$*. Then,* $\hat{\mathbf{W}}^*$ *is the unique optimal solution to* (45)*.*

Proof We first show that $\hat{\mathbf{W}}^*$ is the optimal solution to (45). Let

$$\boldsymbol{\lambda} = (\mathbf{A}_S^{\text{aug}})^T (\mathbf{A}_S^{\text{aug}} (\mathbf{A}_S^{\text{aug}})^T)^{-1} \begin{bmatrix} \mathbf{w}_{s_1}^* / \left\|\mathbf{w}_{s_1}^*\right\|_2 \\ \vdots \\ \mathbf{w}_{s_1}^* / \left\|\mathbf{w}_{s_k}^*\right\|_2 \end{bmatrix}.$$

We can examine that $\mathbf{A}_i^T \boldsymbol{\lambda} = \frac{\mathbf{w}_i^*}{\|\mathbf{w}_i^*\|_2}$ for $i \in S$. For $j \notin S$, from the irrepresentability condition (53), we have $\left\|\mathbf{A}_j^T \boldsymbol{\lambda}\right\|_2 < 1$. Therefore, $(\mathbf{W}^*, \boldsymbol{\lambda})$ satisfies the KKT condition (46). This implies that $\mathbf{W}^*$ is optimal to (45).

Then, we prove the uniqueness. Suppose that $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ is another optimal solution to (9). Denote $\Delta_j = \mathbf{w}_j - \hat{\mathbf{w}}_j^*$ for $j \in [k]$. Then, we have

$$\sum_{j \notin S} \mathbf{A}_j \Delta_j + \sum_{i \in S} \mathbf{A}_i \Delta_i = 0. \tag{55}$$

We note that

$$\begin{aligned}&\sum_{i\in S}\|\mathbf{w}_i\|_2+\sum_{j\notin S}\|\mathbf{w}_j\|_2\\ \geq&\sum_{i\in S}\|\hat{\mathbf{w}}_i^*\|_2+\sum_{i\in S}\left(\frac{\hat{\mathbf{w}}_i^*}{\|\hat{\mathbf{w}}_i^*\|_2}\right)^T\Delta_i+\sum_{j\notin S}\|\mathbf{w}_j\|_2\\ =&\sum_{i\in S}\|\hat{\mathbf{w}}_i^*\|_2+\boldsymbol{\lambda}^T\sum_{i\in S}\mathbf{A}_i\Delta_i+\sum_{j\notin S}\|\Delta_j\|_2\\ =&\sum_{i\in S}\|\hat{\mathbf{w}}_i^*\|_2-\boldsymbol{\lambda}^T\sum_{j\notin S}\mathbf{A}_j\Delta_j+\sum_{j\notin S}\|\Delta_j\|_2\\ \geq&\sum_{i\in S}\|\hat{\mathbf{w}}_i^*\|_2+\sum_{j\notin S}\|\Delta_j\|_2\left(1-\left\|\mathbf{A}_j^T\boldsymbol{\lambda}\right\|_2\right)\\ \geq&\sum_{i\in S}\|\hat{\mathbf{w}}_i^*\|_2.\end{aligned}\tag{56}$$

The equality holds when $\Delta_j=0$ for all $j\notin S$ and $\Delta_i=\gamma_i\frac{\mathbf{w}_{i^*}^*}{\|\mathbf{w}_{i^*}^*\|_2}$ for $i\in S$. Here $\gamma_i\geq 0$ for $i\in S$. As $\Delta_j=0$ for all $j\notin S$, we have

$$\sum_{i\in S}\mathbf{A}_i\Delta_i=0.\tag{57}$$

Note that $\begin{bmatrix}\mathbf{A}_{s_1}^T\mathbf{A}_{s_1} & \dots & \mathbf{A}_{s_1}^T\mathbf{A}_{s_k}\\ \vdots & \ddots & \vdots\\ \mathbf{A}_{s_k}^T\mathbf{A}_{s_1} & \dots & \mathbf{A}_{s_k}^T\mathbf{A}_{s_k}\end{bmatrix}$ is invertible. This implies that $\Delta_i=0$ for $i\in S$. Hence, $\hat{\mathbf{W}}^*$ is the unique optimal solution to (9).

A. Proof of Proposition 2

Proof We first show that the neural isometry condition (NIC-1) implies the recovery of the planted model via (9).

Proposition 17 *Suppose that the neural isometry condition* (NIC-1) *holds. Consider* $\mathbf{W}^*=(\mathbf{w}_1^*,\dots,\mathbf{w}_p^*)$ *such that* $\mathbf{w}_{i^*}^*=\mathbf{w}^*$ *while* $\mathbf{w}_j^*=0$ *for* $j\neq i^*$. *Then,* $\mathbf{W}^*$ *is the unique optimal solution of* (9).

Proof The KKT conditions for (9) include

$$\begin{aligned}&\left\|\mathbf{X}^T\mathbf{D}_j\boldsymbol{\lambda}^T\right\|\leq 1, && \text{if } \mathbf{w}_j=0,\\ &\mathbf{X}^T\mathbf{D}_j\boldsymbol{\lambda}^T=\frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}, && \text{if } \mathbf{w}_j\neq 0,\\ &\sum_{j=1}^p\mathbf{D}_j\mathbf{X}\mathbf{w}_j=\mathbf{y},\end{aligned}\tag{58}$$

where $\boldsymbol{\lambda}\in\mathbb{R}^n$ is the dual variable. As a direct corollary of Proposition 14, $\mathbf{W}^*$ is the unique optimal solution to (9).

Then, we present the proof of Proposition 2. As (9) is derived by dropping all inequality constraints in (8), the optimal value of (8) is lower bounded by (9). As $\mathbf{W}^*$ is the unique optimal solution to (9), the optimal value of (9) is $\|\mathbf{w}^*\|_2$. On the other hand, $\hat{\mathbf{W}}$ is feasible for (8) and it leads to an objective value of $\|\mathbf{w}^*\|_2$. This implies that $\hat{\mathbf{W}}$ is an optimal solution to (8) and the optimal values of (8) and (9) are the same. Suppose that we have another solution $\bar{\mathbf{W}}$ which is optimal to (8). Consider $\breve{\mathbf{W}}=\{\breve{\mathbf{w}}_j|j\in[p]\}$ where $\breve{\mathbf{w}}_j=\bar{\mathbf{w}}_j-\bar{\mathbf{w}}_j'$. Then, $\breve{\mathbf{W}}$ is also optimal to (9). As $\mathbf{W}^*$ is the unique optimal solution to (9), this implies that $\breve{\mathbf{W}}=\mathbf{W}^*$. Then, we note that for $j\neq i^*$,

$$\|\bar{\mathbf{w}}_j\|_2+\|\bar{\mathbf{w}}_j'\|_2\geq 0=\|\mathbf{w}_j^*\|_2.\tag{59}$$

The equality holds when $\bar{\mathbf{w}}_j=\bar{\mathbf{w}}_j'=0$. We also note that $\bar{\mathbf{w}}_{i^*}-\bar{\mathbf{w}}_{i^*}'=\mathbf{w}^*$. This implies that

$$\|\bar{\mathbf{w}}_{i^*}\|_2+\|\bar{\mathbf{w}}_{i^*}'\|_2\geq\|\bar{\mathbf{w}}_{i^*}-\bar{\mathbf{w}}_{i^*}'\|_2=\|\mathbf{w}^*\|_2.\tag{60}$$

The equality holds if and only if there exists $\alpha\in[0,1]$ such that $\bar{\mathbf{w}}_{i^*}=\alpha\mathbf{w}^*$ and $\bar{\mathbf{w}}_{i^*}'=-(1-\alpha)\mathbf{w}^*$. If $\alpha=1$, then $\bar{\mathbf{W}}=\mathbf{W}^*$. If $\alpha<1$, as $(2\mathbf{D}_{i^*}-\mathbf{I}_n)\mathbf{X}\bar{\mathbf{w}}_{i^*}'\geq 0$ and $(2\mathbf{D}_{i^*}-\mathbf{I}_n)\mathbf{X}\mathbf{w}^*\geq 0$, this implies that $(2\mathbf{D}_{i^*}-\mathbf{I}_n)\mathbf{X}\mathbf{w}^*=0$. Therefore, we have $\mathbf{X}\mathbf{w}^*=0$ and $\mathbf{X}^T\mathbf{X}\mathbf{w}^*=0$. This leads to a contradiction because $\mathbf{w}^*\neq 0$ and $\mathbf{X}^T\mathbf{X}$ is invertible with probability 1 for $n>d$.

B. Proof of Proposition 3

Proof We first show that the normalized neural isometry condition implies the the recovery of the planted model via (13).

Proposition 18 *Suppose that the neural isometry condition* (NIC-L) *holds. Consider* $\mathbf{W}^* = (\mathbf{w}_0^*, \dots, \mathbf{w}_p^*)$ *such that* $\mathbf{w}_{i^*}^* = \frac{\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*}{\|\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*\|_2}$ *while* $\mathbf{w}_j^* = 0$ *for* $j \neq i^*$. *Then,* $\mathbf{W}^*$ *is the unique optimal solution to* (14).

Proof We note that

$$\mathbf{y} = \frac{(\mathbf{X}\mathbf{w}^*)_+}{\|(\mathbf{X}\mathbf{w}^*)_+\|_2} = \frac{\mathbf{D}_{i^*}\mathbf{X}\mathbf{w}^*}{\|\mathbf{D}_{i^*}\mathbf{X}\mathbf{w}^*\|_2} = \frac{\mathbf{U}_{i^*}\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*}{\|\mathbf{U}_{i^*}\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^*\|_2} = \mathbf{U}_{i^*}\mathbf{w}_{i^*}^*.$$

The KKT conditions for (14) include

$$\begin{aligned} &\left\|\mathbf{U}_j\boldsymbol{\lambda}^T\right\| \leq 1, &&\text{if } \mathbf{w}_j = 0,\\ &\mathbf{U}_j^T\boldsymbol{\lambda}^T = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}, &&\text{if } \mathbf{w}_j \neq 0,\\ &\sum_{j=1}^p \mathbf{U}_j\mathbf{w}_j = \mathbf{y}, \end{aligned} \tag{61}$$

As a direct corollary of Proposition 14, $\mathbf{W}^*$ is the unique optimal solution to (14).

We then present the proof of Proposition 3. As (14) is derived by dropping all inequality constraints in (13), the optimal value of (13) is lower bounded by (14). As $\mathbf{W}^*$ is the unique optimal solution to (14), the optimal value of (14) is 1. On the other hand, $\hat{\mathbf{W}}$ is feasible for (8) and it leads to an objective value of 1. This implies that $\hat{\mathbf{W}}$ is an optimal solution to (13) and the optimal values of (13) and (14) are the same. Suppose that we have another solution $\bar{\mathbf{W}}$ which is optimal to (13). Consider $\breve{\mathbf{W}} = \{\breve{\mathbf{w}}_j | j \in [p]\}$ where $\breve{\mathbf{w}}_j = \bar{\mathbf{w}}_j - \bar{\mathbf{w}}_j'$. Then, $\breve{\mathbf{W}}$ is also optimal to (14). As $\mathbf{W}^*$ is the unique optimal solution to (14), this implies that $\breve{\mathbf{W}} = \mathbf{W}^*$. Then, we note that for $j \neq i^*$,

$$\|\bar{\mathbf{w}}_j\|_2 + \|\bar{\mathbf{w}}_j'\|_2 \geq 0 = \|\mathbf{w}_j^*\|_2. \tag{62}$$

The equality holds when $\bar{\mathbf{w}}_j = \bar{\mathbf{w}}_j' = 0$. We also note that

$$\|\bar{\mathbf{w}}_{i^*}\|_2 + \|\bar{\mathbf{w}}_{i^*}'\|_2 \geq \|\bar{\mathbf{w}}_{i^*} - \bar{\mathbf{w}}_{i^*}'\|_2 = \|\mathbf{w}_{i^*}^*\|_2. \tag{63}$$

The equality holds if and only if there exists $\alpha \in [0,1]$ such that $\bar{\mathbf{w}}_{i^*} = \alpha\mathbf{w}^*$ and $\bar{\mathbf{w}}_{i^*}' = -(1-\alpha)\mathbf{w}_{i^*}^*$. If $\alpha = 1$, then $\bar{\mathbf{W}} = \bar{\mathbf{W}}$. If $\alpha < 1$, as $(2\mathbf{D}_{i^*} - \mathbf{I}_n)\mathbf{X}\mathbf{V}_i^T\mathbf{\Sigma}_i^{-1}\bar{\mathbf{w}}_{i^*}' \geq 0$ and $(2\mathbf{D}_{i^*} - \mathbf{I}_n)\mathbf{X}\mathbf{V}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{w}_{i^*}^* \geq 0$, this implies that $(2\mathbf{D}_{i^*} - \mathbf{I}_n)\mathbf{X}\mathbf{V}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{w}_{i^*}^* = 0$. Therefore, we have $\mathbf{X}\mathbf{V}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{w}^* = 0$ and $\mathbf{X}^T\mathbf{X}\mathbf{V}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{w}_{i^*}^* = 0$. This leads to a contradiction because $\mathbf{V}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{w}^* \neq 0$ and $\mathbf{X}^T\mathbf{X}$ is invertible with probability 1 for $n > d$.

C. Proof of Proposition 4

Proof We first show that the neural isometry condition (NIC-k) implies the recovery of the planted model by solving (9).

Proposition 19 *Consider* $\tilde{\mathbf{W}}^* = (\tilde{\mathbf{w}}_1^*, \dots, \tilde{\mathbf{w}}_p^*)$ *satisfies that* $\tilde{\mathbf{w}}_{s_i}^* = r_i^*\mathbf{w}_i^*$ *for* $i \in \{1, \dots, k\}$ *and* $\tilde{\mathbf{w}}_j^* = 0$ *for* $j \notin S$. *Suppose that the neural isometry condition* (NIC-k) *is satisfied. Then,* $\tilde{\mathbf{W}}^*$ *is the unique optimal solution to* (14).

Proof The KKT conditions for (9) consist of

$$\begin{aligned} &\left\|\mathbf{D}_j\mathbf{X}\boldsymbol{\lambda}^T\right\| \leq 1, &&\text{if } \tilde{\mathbf{w}}_j = 0,\\ &\mathbf{D}_j\mathbf{X}^T\boldsymbol{\lambda}^T = \frac{\tilde{\mathbf{w}}_j}{\|\tilde{\mathbf{w}}_j\|}, &&\text{if } \tilde{\mathbf{w}}_j \neq 0,\\ &\sum_{j=1}^p \mathbf{D}_j\mathbf{X}\tilde{\mathbf{w}}_j = \mathbf{y}, \end{aligned} \tag{64}$$

As a direct corollary of Proposition 16, $\tilde{\mathbf{W}}^*$ is the unique optimal solution to (9).

We then present the proof of Proposition 4. From Proposition 19, we note that $\tilde{\mathbf{W}}^*$ is the unique optimal solution to (14). As (9) is derived by dropping all inequality constraints in (8), the optimal value of (8) is lower bounded by (9). As $\tilde{\mathbf{W}}^*$ is the unique optimal solution to (9), the optimal value of (9) is $\sum_{i=1}^k \|\mathbf{w}_i^*\|_2$. On the other hand, $\hat{\mathbf{W}}$ is feasible for (8) and it leads to an objective value of $\sum_{i=1}^k \|\mathbf{w}_i^*\|_2$. This implies that $\hat{\mathbf{W}}$ is an optimal solution to (8) and the optimal values of (8) and (9) are the same.

Suppose that we have another solution $\bar{\mathbf{W}}$ which is optimal to (8). Consider $\breve{\mathbf{W}} = \{\breve{\mathbf{w}}_j | j \in [p]\}$ where $\breve{\mathbf{w}}_j = \bar{\mathbf{w}}_j - \bar{\mathbf{w}}'_j$. Then, $\bar{\mathbf{W}}$ is also optimal to (9). As $\mathbf{W}^*$ is the unique optimal solution to (9), this implies that $\breve{\mathbf{W}} = \mathbf{W}^*$, or equivalently, $\bar{\mathbf{w}}_j - \bar{\mathbf{w}}'_j = \tilde{\mathbf{w}}_j$ for $j \in [p]$. Then, we note that for $j \notin S$

$$\|\bar{\mathbf{w}}_j\|_2 + \|\bar{\mathbf{w}}'_j\|_2 \geq 0 = \|\mathbf{w}^*_j\|_2. \tag{65}$$

The equality holds when $\bar{\mathbf{w}}_j = \bar{\mathbf{w}}'_j = 0$. For $i \in [k]$, we have $\bar{\mathbf{w}}_{s_i} - \bar{\mathbf{w}}'_{s_i} = \tilde{\mathbf{w}}_{s_i} = r^*_i \mathbf{w}^*_i$. This implies that

$$\|\bar{\mathbf{w}}_{s_i}\|_2 + \|\bar{\mathbf{w}}'_i\|_2 \geq \|\bar{\mathbf{w}}_i - \bar{\mathbf{w}}'_i\|_2 = \|\mathbf{w}^*_i\|_2. \tag{66}$$

The equality holds if and only if there exists $\alpha \in [0,1]$ such that $\bar{\mathbf{w}}_{s_i} = \alpha r^*_i \mathbf{w}^*_i$ and $\bar{\mathbf{w}}'_{s_i} = -(1-\alpha) r^*_i \mathbf{w}^*_i$. If $0 < \alpha < 1$, as $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\bar{\mathbf{w}}'_{s_i} \geq 0$ and $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\bar{\mathbf{w}}_{s_i} \geq 0$, this implies that $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\mathbf{w}^*_i = 0$. Therefore, we have $\mathbf{X}\mathbf{w}^*_i = 0$ and $\mathbf{X}^T\mathbf{X}\mathbf{w}^*_i = 0$. This leads to a contradiction because $\mathbf{w}^*_i \neq 0$ and $\mathbf{X}^T\mathbf{X}$ is invertible with probability 1 for $n > d$. If $r^*_i = 1$, then $\alpha = 1$. Otherwise, we have $\alpha = 0$ and this leads to $\bar{\mathbf{w}}'_{s_i} = -\mathbf{w}^*_i$, which is contradictory to $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\bar{\mathbf{w}}'_{s_i} \geq 0$. If $r^*_i = -1$, similarly, we have $\alpha = 0$. Otherwise, it follows that $\alpha = 1$ and this leads to $\bar{\mathbf{w}}_{s_i} = -\mathbf{w}^*_i$, which is contradictory to $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\bar{\mathbf{w}}_{s_i} \geq 0$.

D. Proof of Proposition 5

Proof We first show that the neural isometry condition (NNIC-k) implies the recovery of the planted model by solving (14).

Proposition 20 *Consider $\tilde{\mathbf{W}}^* = (\tilde{\mathbf{w}}^*_1, \dots, \tilde{\mathbf{w}}^*_p)$ satisfies that $\tilde{\mathbf{w}}^*_i = r^*_i \frac{\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i}{\|\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i\|_2}$ for $i \in S$ and $\tilde{\mathbf{w}}^*_j = 0$ for $j \notin S$. Suppose that the neural isometry condition* (NNIC-k) *is satisfied. Then, $\tilde{\mathbf{W}}^*$ is the unique optimal solution to* (14).

Proof We note that for $i \in [k]$, we have

$$\frac{(\mathbf{X}\mathbf{w}^*_i)_+}{\|(\mathbf{X}\mathbf{w}^*_i)_+\|_2} = \frac{\mathbf{D}_{s_i}\mathbf{X}\mathbf{w}^*_i}{\left\|\mathbf{D}_{s_i}\mathbf{X}\mathbf{w}^*_i\right\|_2} = \frac{\mathbf{U}_{s_i}\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i}{\left\|\mathbf{U}_{s_i}\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i\right\|_2} = \mathbf{U}_{s_i} r^*_i \tilde{\mathbf{w}}^*_i.$$

This implies that

$$\mathbf{y} = \sum_{i=1}^k r^*_i \frac{(\mathbf{X}\mathbf{w}^*_i)_+}{\|(\mathbf{X}\mathbf{w}^*_i)_+\|_2} = \sum_{i=1}^k \mathbf{U}_{s_i}\mathbf{w}^*_i \tilde{\mathbf{w}}^*_i.$$

The KKT conditions for (14) consist of

$$\begin{aligned} &\left\|\mathbf{U}_j \boldsymbol{\lambda}^T\right\| \leq 1, && \text{if } \tilde{\mathbf{w}}_j = 0,\\ &\mathbf{U}_j^T \boldsymbol{\lambda}^T = \frac{\tilde{\mathbf{w}}_j}{\|\tilde{\mathbf{w}}_j\|}, && \text{if } \tilde{\mathbf{w}}_j \neq 0,\\ &\sum_{j=1}^p \mathbf{U}_j \tilde{\mathbf{w}}_j = \mathbf{y}, && \end{aligned} \tag{67}$$

As a direct corollary of Proposition 16, $\tilde{\mathbf{W}}^*$ is the unique optimal solution to (14).

We then present the proof of Proposition 4. From Proposition 16, $\tilde{\mathbf{W}}^*$ is the unique optimal solution to (14). Because (14) is derived by dropping all inequality constraints in (13), the optimal value of (13) is lower bounded by (14). As $\tilde{\mathbf{W}}^*$ is the unique optimal solution to (14), the optimal value of (14) is k. On the other hand, $\hat{\mathbf{W}}$ is feasible for (13) and it leads to an objective value of k. This implies that $\hat{\mathbf{W}}$ is an optimal solution to (13) and the optimal values of (13) and (14) are the same.

Suppose that we have another solution $\bar{\mathbf{W}}$ which is optimal to (8). Consider $\breve{\mathbf{W}} = \{\breve{\mathbf{w}}_j | j \in [p]\}$ where $\breve{\mathbf{w}}_j = \bar{\mathbf{w}}_j - \bar{\mathbf{w}}'_j$. Then, $\bar{\mathbf{W}}$ is also optimal to (9). As $\mathbf{W}^*$ is the unique optimal solution to (9), this implies that $\breve{\mathbf{W}} = \mathbf{W}^*$, or equivalently, $\bar{\mathbf{w}}_j - \bar{\mathbf{w}}'_j = \tilde{\mathbf{w}}_j$ for $j \in [p]$. Then, we note that for $j \notin S$

$$\|\bar{\mathbf{w}}_j\|_2 + \|\bar{\mathbf{w}}'_j\|_2 \geq 0 = \|\mathbf{w}^*_j\|_2. \tag{68}$$

The equality holds when $\bar{\mathbf{w}}_j = \bar{\mathbf{w}}'_j = 0$. For $i \in [k]$, we have $\bar{\mathbf{w}}_{s_i} - \bar{\mathbf{w}}'_{s_i} = \tilde{\mathbf{w}}_{s_i} = r^*_i \frac{\mathbf{\Sigma}_i\mathbf{V}_i\mathbf{w}^*_i}{\|\mathbf{\Sigma}_i\mathbf{V}_i\mathbf{w}^*_i\|_2}$. This implies that

$$\|\bar{\mathbf{w}}_{s_i}\|_2 + \|\bar{\mathbf{w}}'_i\|_2 \geq \|\bar{\mathbf{w}}_i - \bar{\mathbf{w}}'_i\|_2 = 1. \tag{69}$$

The equality holds if and only if there exists $\alpha \in [0,1]$ such that $\bar{\mathbf{w}}_{s_i} = \alpha r^*_i \frac{\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i}{\|\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i\|_2}$ and $\bar{\mathbf{w}}'_{s_i} = -(1-\alpha) r^*_i \frac{\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i}{\|\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}^*_i\|_2}$. If $0 < \alpha < 1$, as $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\bar{\mathbf{w}}'_{s_i} \geq 0$ and $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\mathbf{w}_{s_i} \geq 0$, this implies that $(2\mathbf{D}_{s_i} - \mathbf{I}_n)\mathbf{X}\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\mathbf{w}^*_i = 0$. Therefore, we have $\mathbf{X}\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\mathbf{w}^*_i = 0$ and $\mathbf{X}^T\mathbf{X}\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\mathbf{w}^*_i = 0$. This leads to a contradiction because $\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\mathbf{w}^*_i \neq 0$ and $\mathbf{X}^T\mathbf{X}$ is invertible with probability 1 for $n > d$. If $r^*_i = 1$, then we have $\alpha = 1$. Otherwise, we have $\alpha = 0$ and this leads to $\bar{\mathbf{w}}'_{s_i} = -\frac{\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}_i^*}{\|\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}_i^*\|_2}$, which is contradictory to $(2\mathbf{D}_{s_i}-\mathbf{I}_n)\mathbf{X}\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\bar{\mathbf{w}}'_{s_i}\geq 0$. If $r_i^* = -1$, similarly, we have $\alpha = 0$. Otherwise, we have $\alpha = 1$ and this leads to $\bar{\mathbf{w}}_{s_i} = -\frac{\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}_i^*}{\|\mathbf{\Sigma}_{s_i}\mathbf{V}_{s_i}\mathbf{w}_i^*\|_2}$, which is contradictory to $(2\mathbf{D}_{s_i}-\mathbf{I}_n)\mathbf{X}\mathbf{V}_{s_i}^T\mathbf{\Sigma}_{s_i}^{-1}\bar{\mathbf{w}}_{s_i}\geq 0$.

Appendix I
Proofs in Section V

A. Proof of Lemma 2

Proof We first illustrate Theorem 1 in [1] as follows:

Theorem 12 *Fix a tolerance $\eta \in (0,1)$. Let C and K be convex cones in $\mathbb{R}^n$. Draw a random orthogonal basis $\mathbf{W} \in \mathbb{R}^{n\times n}$. Then,*

$$P(K \cap \mathbf{W}C \neq \{0\}) \begin{cases} \leq \eta & \delta(C) + \delta(K) \leq n - a_\eta\sqrt{n}, \\ \geq 1-\eta & \delta(C) + \delta(K) \geq n + a_\eta\sqrt{n}. \end{cases} \tag{70}$$

where $a_\eta = 8\sqrt{\log(4/\eta)}$.

We can take $C = \left\{ t \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix} | t \geq 0 \right\}$, where $\mathbf{0} \in \mathbb{R}^{n-d}$ is a vector of 0s. Then, $P(\mathbf{X}\mathbf{w} \in K) = P(\mathbf{W}C \cap K \neq \{0\})$. As $w(C) = d$, for $w(K) + d < n$, we have $a_\eta = \frac{n-\delta(K)-d}{\sqrt{n}}$. This implies that $\eta = 4e^{-\frac{(n-\delta(K)-d)^2}{64n}}$. Hence, by taking $\alpha = \frac{(n-\delta(K)-d)^2}{64n^2}$, (18) holds. Similarly, for $\delta(K)+d > n$, we have $a_\eta = \frac{w(K)+d-n}{\sqrt{n}}$, which implies that $\eta = 4e^{-\frac{(n-\delta(K)-d)^2}{64n}}$. By taking $\alpha = \frac{(n-\delta(K)-d)^2}{64n^2}$, (18) holds

B. Proof of Theorem 8

Proof We first note that the matrix in the irrepresentability condition (NIC-L) has the following upper bound

$$\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0} \left\| \mathbf{X}^T \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \right\|_2 \leq \frac{1}{\sigma_{\min}^2} \max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0} \|\mathbf{X}^T \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\mathbf{X}\|_2,$$

where $\sigma_{\min}$ is the smallest singular value of $\mathbf{X}$. We introduce a lemma to bound the norm of $\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\mathbf{X}$.

Lemma 3 *Let $h \in \mathbb{R}^d$ and $t > 0$ be fixed. Suppose that each element $x_{i,j}$ of $\mathbf{X} \in \mathbb{R}^{n\times d}$ are i.i.d. random variables following a mean-zero sub-Gaussian distribution with variance proxy σ^2 such that*

- $\mathbb{E}[x_{i,j}^2] = \frac{1}{n}$.
- *$-x_{i,j}$ has the same distribution as $x_{i,j}$.*

Then, for $d \geq 2$, with probability at least

$$1 - 4\exp\left(-\frac{nt^2}{162\sqrt{2}\sigma^2} + \frac{d\log(54n)}{2}\right)$$

we have

$$\max_{\mathbf{h}\in\mathbb{R}^d,\mathbf{h}\neq 0} \|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\mathbf{X}\|_2 \leq \frac{1}{2} + t. \tag{71}$$

From Lemma 3, by taking $t = 1/4$, for $n \geq 4000\sigma^2 d\log(54n)$, we have

$$P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0} \|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\mathbf{X}\|_2 \leq \frac{3}{4}\right) \geq 1 - 4\exp\left(-\frac{n}{8000\sigma^2}\right). \tag{72}$$

On the other hand, from Theorem 4.6.1 in [47], we can lower bound the smallest eigenvalue of $\mathbf{X}$,

$$P\left(\sigma_{\min} \geq 1 - 2\sqrt{\frac{d}{n}}\right) \geq 1 - 2\exp(-d). \tag{73}$$

For $n \geq 1024d$. we have

$$P\left(\sigma_{\min}^2 \geq 7/8\right) \geq P\left(\sigma_{\min} \geq 1 - 1/16\right) \geq P\left(\sigma_{\min} \geq 1 - 2\sqrt{\frac{d}{n}}\right) \geq 1 - 2\exp(-d). \quad \text{‘} \tag{74}$$

This implies that for n satisfying $n \geq \max\{1024d, 4000\sigma^2 d\log(54n)\}$, with probability at least $1 - 2\exp(-d) - 4\exp\left(-\frac{n}{8000\sigma^2}\right)$, we have

$$\max_{j\in[p]} \left\| \mathbf{X}^T\mathbf{D}_j\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \right\|_2 \leq \frac{3}{4\sigma_{\min}^2} \leq \frac{3}{4}\cdot\frac{8}{7} < 1. \tag{75}$$

Conditioned on the above event, we note that for any $\mathbf{w}^* \in \mathbb{R}^d, \mathbf{w}^* \neq 0$, we have

$$\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0} \left\|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{Xh}\geq 0))\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}\right\|_2 \leq \max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0} \left\|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{Xh}\geq 0))\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right\|_2 < 1, \tag{76}$$

i.e., the irrepresentability condition (NIC-L) holds. This completes the proof.

C. Proof of Lemma 3

PROOF For a positive semi-definite matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$, we have $\|\mathbf{A}\|_2 = \max_{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1}\mathbf{z}^T\mathbf{A}\mathbf{z}$. Note that

$$\begin{aligned}
&\left|\max_{\mathbf{h}\in\mathbb{R}^d,\mathbf{h}\neq 0}\|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{Xh}\geq 0))\mathbf{X}\|_2 - \frac{1}{2}\right| \\
=&\left|\max_{\mathbf{h}\in\mathbb{R}^d,\mathbf{h}\neq 0}\left\|\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0))\right\|_2 - \frac{1}{2}\right| \\
=&\left|\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\sum_{i=1}^n \mathbf{z}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{z}\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)) - \frac{1}{2}\right| \\
\leq&\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)) - \frac{1}{2}\right|.
\end{aligned} \tag{77}$$

By rescaling $\mathbf{x}_i = \sqrt{n}\mathbf{x}_i$, we can rewrite the above quantity as

$$\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)) - \frac{1}{2}\right|. \tag{78}$$

As $-x_{i,j}$ has the same distribution as $x_{i,j}$, we have

$$\begin{aligned}
&P(\mathbf{h}^T\mathbf{x}_i\geq 0) = P(\mathbf{h}^T\mathbf{x}_i\leq 0), \\
&\mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2|\mathbf{h}^T\mathbf{x}_i\geq 0] = \mathbb{E}[(-\mathbf{z}^T\mathbf{x}_i)^2|\mathbf{h}^T(-\mathbf{x}_i)\geq 0] = \mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2|\mathbf{h}^T\mathbf{x}_i\leq 0].
\end{aligned} \tag{79}$$

We note that $\mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2] = \|\mathbf{z}\|_2^2 = 1$ and

$$\mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2] = \mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2|\mathbf{h}^T\mathbf{x}_i\geq 0]P(\mathbf{h}^T\mathbf{x}_i\geq 0) + \mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2|\mathbf{h}^T\mathbf{x}_i\leq 0]P(\mathbf{h}^T\mathbf{x}_i\leq 0). \tag{80}$$

This implies that $\mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2|\mathbf{h}^T\mathbf{x}_i\geq 0]P(\mathbf{h}^T\mathbf{x}_i\geq 0) = \frac{1}{2}$, or equivalently,

$$\mathbb{E}[(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)] = \frac{1}{2}. \tag{81}$$

Therefore, it is sufficient to upper-bound the following probability

$$P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0) - \frac{1}{2}\right|\geq t\right) \tag{82}$$

We note that

$$\begin{aligned}
&P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i')^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right]\right|\geq t\right) \\
&\cdot P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i')^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right]\right|\geq t\right) \\
\geq& P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0) - \frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i')^2\mathbb{I}(\mathbf{v}^T\mathbf{x}_i'\geq 0)\right|\geq 2t\right).
\end{aligned} \tag{83}$$

where $\mathbf{x}_i'$ are i.i.d. random vectors following the same distribution of $\mathbf{x}_i$ and they are independent with $\mathbf{x}_1,\dots,\mathbf{x}_n$. This implies that

$$\begin{aligned}
&P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i')^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right]\right|\geq t\right) \\
\leq& P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0) - \frac{1}{n}\sum_{i=1}^n (\mathbf{z}^T\mathbf{x}_i')^2\mathbb{I}(\mathbf{v}^T\mathbf{x}_i'\geq 0)\right|\geq 2t\right)^{\frac{1}{2}},
\end{aligned} \tag{84}$$

By introducing i.i.d. random variables ϵ_i uniformly distributed in $\{-1,1\}$, we have the following bound

$$\begin{aligned}
&P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)-\frac{1}{n}\sum_{i=1}^n(\mathbf{z}^T\mathbf{x}_i')^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right|\geq 2t\right)\\
=&P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\left((\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)-(\mathbf{z}^T\mathbf{x}_i')^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right)\right|\geq 2t\right)\\
=&P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)\right|\geq t\right),
\end{aligned}\tag{85}$$

where $\epsilon_{n+1},\dots,\epsilon_{2n}$ are i.i.d. copies of $\epsilon_1,\dots,\epsilon_n$.

Decoupling: Next, we apply a decoupling result in [49, Theorem 3.4.1] to obtain the following upper bound:

$$\begin{aligned}
&P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)\right|\geq t\right)\\
\leq&8P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right|\geq t\right),
\end{aligned}\tag{86}$$

were $\{\mathbf{x}_i'\}_{i=1}^{2n}$ is an independent and identically distributed copy of the sequence $\{\mathbf{x}_i\}_{i=1}^{2n}$.

ϵ-net bound: For each realization of $x_1',\dots,\mathbf{x}_{2n}'$, let $\mathcal{P}'=\{\mathbf{diag}(\mathbb{I}(\mathbf{X}'\mathbf{h}\geq 0))|\mathbf{h}\in\mathbb{R}^d,\mathbf{h}\neq 0\}$, $p'=|\mathcal{P}'|$ and we write $\mathcal{P}'=\{\mathbf{D}_1,\dots,\mathbf{D}_{p'}\}$. According to [42], we have the upper bound $p'\leq 2d(2en/d)^d$. We note that

$$\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right|=\max_{j\in[p']}\max_{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|\tag{87}$$

Consider an ϵ-net of $\{\mathbf{z}\in\mathbb{R}^d:\|z\|_2=1\}$, $\{\mathbf{z}_1,\dots,\mathbf{z}_N\}$, where $N\leq(1+2/\epsilon)^d$. Namely, for any $\mathbf{z}\in\mathbb{R}^d$ satisfying $\|\mathbf{z}\|_2=1$, there exists $k\in[N]$ such that $\mathbf{z}=\mathbf{z}_k+\Delta$, where $\|\Delta\|_2\leq\epsilon$. Then, we have

$$\begin{aligned}
&\max_{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|\\
\leq&\max_{k\in[N],\|\Delta\|_2\leq\epsilon}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i((\mathbf{z}_k+\Delta)^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|\\
\leq&\max_{k\in[N]}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}_k^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|+(2\epsilon+\epsilon^2)\max_{\|\mathbf{z}\|_2\leq 1}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|
\end{aligned}\tag{88}$$

Here we utilize that for an arbitrary symmetric matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$ and arbitrary vector $\mathbf{z}\in\mathbb{R}^d$ with $\|\mathbf{z}\|_2=1$,

$$\max_{\|\Delta\|_2\leq\epsilon}|(\mathbf{z}+\Delta)^T\mathbf{A}(\mathbf{z}+\Delta)|\leq|\mathbf{z}^T\mathbf{A}\mathbf{z}|+(2\epsilon+\epsilon^2)\|\mathbf{A}\|_2.\tag{89}$$

This implies that

$$\max_{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|\leq\frac{1}{1-2\epsilon-\epsilon^2}\max_{k\in[N]}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}_k^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|\tag{90}$$

For fixed k, we note that $\mathbf{z}_k^T\mathbf{x}_i$ is also sub-Gaussian with variance proxy σ^2. Let $g_i=(\mathbf{z}_k^T\mathbf{x}_i)^2$. Therefore, $\epsilon_i g_i^2(\mathbf{D}_j)_{i,i}$ is sub-exponential with parameters $(4\sqrt{2}\sigma^2(\mathbf{D}_j)_{i,i},4\sigma^2)$. This implies that $\sum_{i=1}^{2n}\epsilon_i g_i^2(\mathbf{D}_j)_{i,i}$ is sub-exponential with parameter $(\nu^*,4\sigma^2)$, where

$$\nu^*=\sqrt{\sum_{i=1}^{2n}4\sqrt{2}\sigma^2(\mathbf{D}_j)_{i,i}}=\sqrt{4\sqrt{2}\sigma^2\operatorname{tr}(\mathbf{D}_j)}.\tag{91}$$

Therefore, for $t\leq\frac{(\nu^*)^2}{4\sigma^2}=\sqrt{2}\operatorname{tr}(\mathbf{D}_j)$, we have

$$P\left(\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}_k^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|>t\right)\leq 2\exp\left(-2nt^2/(2(\nu^*)^2)\right)=2\exp\left(-\frac{n^2t^2}{4\sqrt{2}\sigma^2\operatorname{tr}(\mathbf{D}_j)}\right)\leq 2\exp\left(-\frac{nt^2}{4\sqrt{2}\sigma^2}\right).$$

This implies that

$$P\left(\max_{k\in[N]}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}_k^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|>t\right)\leq 2N\exp\left(-\frac{nt^2}{4\sqrt{2}\sigma^2}\right)\leq 2\exp\left(-\frac{nt^2}{4\sqrt{2}\sigma^2}+d\log(3/\epsilon)\right).\tag{92}$$

Again, by applying the union bound, we have

$$\begin{aligned}
&P\left(\max_{j\in[p']}\max_{k\in[N]}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}_k^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|>t\right)\\
\leq& p'\cdot P\left(\max_{k\in[N]}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}_k^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|>t\right)\\
\leq& 2\exp\left(-\frac{nt^2}{4\sqrt{2}\sigma^2}+d\log(3/\epsilon)+d(\log(2n)-\log d+1)+\log(2d)\right)\\
\leq& 2\exp\left(-\frac{nt^2}{4\sqrt{2}\sigma^2}+d\log(3/\epsilon)+d\log(6n)\right),
\end{aligned}\tag{93}$$

where we assume that $d\geq 2$. As a result, by taking $\epsilon=1/3$, we have

$$\begin{aligned}
&P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i\geq 0)-\frac{1}{2}\right|\geq t\right)^2\\
\leq& 8\,P\left(\max_{\mathbf{z},\mathbf{h}\in\mathbb{R}^d:\|\mathbf{z}\|_2=1,\mathbf{h}\neq 0}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}^T\mathbf{x}_i)^2\mathbb{I}(\mathbf{h}^T\mathbf{x}_i'\geq 0)\right|\geq t\right)\\
\leq& 8P\left(\max_{j\in[p']}\max_{k\in[N]}\left|\frac{1}{2n}\sum_{i=1}^{2n}\epsilon_i(\mathbf{z}_k^T\mathbf{x}_i)^2(\mathbf{D}_j)_{i,i}\right|>(1-2\epsilon-\epsilon^2)t\right)\\
\leq& 16\exp\left(-\frac{nt^2(1-2\epsilon-\epsilon^2)^2}{4\sqrt{2}\sigma^2}+d\log(3/\epsilon)+d\log(6n)\right)\\
=& 16\exp\left(-\frac{nt^2}{81\sqrt{2}\sigma^2}+d\log(54n)\right).
\end{aligned}\tag{94}$$

This completes the proof.

D. Proof of Theorem 5

Proof Without the loss of generality, we can let $\mathbf{w}^*$ satisfies that $\|\mathbf{w}^*\|_2=1$. As $n>2d$, $\mathbf{X}^T\mathbf{X}$ is invertible with probability 1. Consider the event

$$E=\left\{\max_{\mathbf{h}\in\mathbb{R}^d:\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)\neq\mathbf{1}}\|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0))\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w}^*\|_2<1\right\}.\tag{95}$$

Firstly, we show that

$$P(E)=1.\tag{96}$$

Denote $A(\mathbf{h})=\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0))\mathbf{X}$ and $\tilde{A}(\mathbf{h})=(\mathbf{X}^T\mathbf{X})^{-1}A(\mathbf{h})^TA(\mathbf{h})(\mathbf{X}^T\mathbf{X})^{-1}$. We note that $A(\mathbf{h})$ and $\mathbf{X}^T\mathbf{X}$ are positive semi-definite and symmetric. As $A(\mathbf{h})\succeq\mathbf{X}^T\mathbf{X}$, we have $A(\mathbf{h})^TA(\mathbf{h})\preceq(\mathbf{X}^T\mathbf{X})^2$ and

$$\tilde{A}(\mathbf{h})\preceq I.$$

The equality holds if and only if $A(\mathbf{h})=\mathbf{X}^T\mathbf{X}$. This is equivalent to $\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)-\mathbf{1})\mathbf{X}=0$, which contradicts with $\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)\neq\mathbf{1}$. Hence, we also have

$$\tilde{A}(\mathbf{h})\neq I.$$

Recall our notation $\operatorname{eigmax}\left(\tilde{A}(\mathbf{h})\right)$ used for the subspace of maximal eigenvectors of the symmetric matrix $\tilde{A}(\mathbf{h})$. We note that $\|\tilde{A}(\mathbf{h})\mathbf{w}^*\|_2=1$ if and only if $\|\tilde{A}(\mathbf{h})\|_2=1$ and $\mathbf{w}^*\in\operatorname{eigmax}\left(\tilde{A}(\mathbf{h})\right)$. As $\tilde{A}(\mathbf{h})\neq\mathbf{I}_n$, conditioned on $\|\tilde{A}(\mathbf{h})\|_2=1$, $\operatorname{eigmax}\left(\tilde{A}(\mathbf{h})\right)$ is a random subspace with dimension at most $d-1$. This implies that

$$P(\|\tilde{A}(\mathbf{h})\mathbf{w}^*\|_2=1)=P\left(\|\tilde{A}(\mathbf{h})\|_2=1,\mathbf{w}^*\in\operatorname{eigmax}\left(\tilde{A}(\mathbf{h})\right)\right)=0.\tag{97}$$

For $\boldsymbol{\sigma}\in\{0,1\}^n$, define $B(\boldsymbol{\sigma})=\mathbf{X}^T\mathbf{diag}(\boldsymbol{\sigma})\mathbf{X}$ and $\tilde{B}(\boldsymbol{\sigma})=(\mathbf{X}^T\mathbf{X})^{-1}B(\boldsymbol{\sigma})^TB(\boldsymbol{\sigma})(\mathbf{X}^T\mathbf{X})^{-1}$. Then, we have

$$\begin{aligned}
&P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)\neq\mathbf{1}}\|\tilde{B}(\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0))\mathbf{w}^*\|_2=1\right)\\
=&P\left(\max_{\boldsymbol{\sigma}\in\{0,1\}^n:\exists\mathbf{h}\neq 0,\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)\neq\mathbf{1},\boldsymbol{\sigma}=\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)}\|\tilde{B}(\boldsymbol{\sigma})\mathbf{w}^*\|_2=1\right)\\
\leq&\sum_{\boldsymbol{\sigma}\in\{0,1\}^n}P(\|\tilde{B}(\boldsymbol{\sigma})\mathbf{w}^*\|_2=1,\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)\neq\mathbf{1},\boldsymbol{\sigma}=\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0))=0.
\end{aligned}\tag{98}$$

From the kinematic formula, for $n > 2d$, we have $P(\mathbf{I}_n \notin H) \geq 1 - \exp(-n\alpha)$. In this case, the event E implies that the neural isometry condition (NIC-L) holds. This completes the proof.

E. Proof of Proposition 7

Proof Without the loss of generality, we can let $\mathbf{w}^*$ satisfies that $\|\mathbf{w}^*\|_2 = 1$. As $n > d$, $\mathbf{X}^T\mathbf{X}$ is invertible with probability 1. Consider the event

$$E = \left\{ \max_{\mathbf{h}\in\mathbb{R}^d:\mathbb{I}(\mathbf{Xh}\geq 0)\neq \mathbf{1}} \|\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{Xh} \geq 0))\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w}^*\|_2 < 1 \right\}. \tag{99}$$

First, we show that

$$P(E) = 1. \tag{100}$$

Denote $A(\mathbf{h}) = \mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{Xh} \geq 0))\mathbf{X}$ and $\tilde{A}(\mathbf{h}) = (\mathbf{X}^T\mathbf{X})^{-1}A(\mathbf{h})^TA(\mathbf{h})(\mathbf{X}^T\mathbf{X})^{-1}$. We note that $A(\mathbf{h})$ and $\mathbf{X}^T\mathbf{X}$ are positive semi-definite and symmetric. As $A(\mathbf{h}) \preceq \mathbf{X}^T\mathbf{X}$, we have $A(\mathbf{h})^TA(\mathbf{h}) \preceq (\mathbf{X}^T\mathbf{X})^2$ and

$$\tilde{A}(\mathbf{h}) \preceq I.$$

The equality holds if and only if $A(\mathbf{h}) = \mathbf{X}^T\mathbf{X}$. This is equivalent to $\mathbf{X}^T\mathbf{diag}(\mathbb{I}(\mathbf{Xh} \geq 0) - \mathbf{1})\mathbf{X} = 0$, which contradicts with $\mathbb{I}(\mathbf{Xh} \geq 0) \neq \mathbf{1}$. Hence, we also have

$$\tilde{A}(\mathbf{h}) \neq I.$$

We note that $\|\tilde{A}(\mathbf{h})\mathbf{w}^*\|_2 = 1$ if and only if $\|\tilde{A}(\mathbf{h})\|_2 = 1$ and $\mathbf{w}^* \in \text{eigmax}\left(\tilde{A}(\mathbf{h})\right)$. As $\tilde{A}(\mathbf{h}) \neq \mathbf{I}_n$, conditioned on $\|\tilde{A}(\mathbf{h})\|_2 = 1$, $\text{eigmax}\left(\tilde{A}(\mathbf{h})\right)$ is a random subspace with dimension at most $d-1$. This implies that

$$P(\|\tilde{A}(\mathbf{h})\mathbf{w}^*\|_2 = 1) = P\left(\|\tilde{A}(\mathbf{h})\|_2 = 1, \mathbf{w}^* \in \text{eigmax}\left(\tilde{A}(\mathbf{h})\right)\right) = 0. \tag{101}$$

For $\boldsymbol{\sigma} \in \{0,1\}^n$, define $B(\boldsymbol{\sigma}) = \mathbf{X}^T\mathbf{diag}(\boldsymbol{\sigma})\mathbf{X}$ and $\tilde{B}(\boldsymbol{\sigma}) = (\mathbf{X}^T\mathbf{X})^{-1}B(\boldsymbol{\sigma})^TB(\boldsymbol{\sigma})(\mathbf{X}^T\mathbf{X})^{-1}$. Then, we have

$$\begin{aligned}
&P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbf{1}} \|\tilde{B}(\mathbb{I}(\mathbf{Xh} \geq 0))\mathbf{w}^*\|_2 = 1\right)\\
=&P\left(\max_{\boldsymbol{\sigma}\in\{0,1\}^n:\exists\mathbf{h}\neq 0,\mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbf{1},\boldsymbol{\sigma}=\mathbb{I}(\mathbf{Xh}\geq 0)} \|\tilde{B}(\boldsymbol{\sigma})\mathbf{w}^*\|_2 = 1\right)\\
\leq& \sum_{\boldsymbol{\sigma}\in\{0,1\}^n} P(\|\tilde{B}(\boldsymbol{\sigma})\mathbf{w}^*\|_2 = 1, \mathbb{I}(\mathbf{Xh} \geq 0) \neq \mathbf{1}, \boldsymbol{\sigma} = \mathbb{I}(\mathbf{Xh} \geq 0)) = 0.
\end{aligned} \tag{102}$$

Then, according to Proposition 15, conditioned on the event E, the optimal solution $\mathbf{W} = (\mathbf{w}_0, \dots, \mathbf{w}_p)$ to (12) shall satisfy that $\mathbf{w}_j = 0$ for $\mathbf{D}_j \neq \mathbf{I}_n$. If there does not exist $j \in [p]$ such that $\mathbf{D}_j = \mathbf{I}_n$, then, $\mathbf{W}^* = (\mathbf{w}^*, 0, \dots, 0)$ is the unique optimal solution to (12). Thus, it is also the unique optimal solution to (3).

If there exists $i^* \in [p]$ such that $\mathbf{D}_{i^*} = \mathbf{I}_n$. Let $\mathbf{W} = (\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_1', \dots, \mathbf{w}_p, \mathbf{w}_p')$ be an optimal solution to (3). Let $\hat{\mathbf{w}}_j = \mathbf{w}_j - \mathbf{w}_j'$ for $j \in [p]$. Then, $\hat{\mathbf{W}} = (\mathbf{w}_0, \hat{\mathbf{w}}_1, \dots, \hat{\mathbf{w}}_p)$ is also optimal to (12). This implies that $\hat{\mathbf{w}}_j = 0$ for $j \in [p]$ such that $j \neq i^*$. For $j \in [p]$ such that $j \neq i^*$, we have

$$\|\hat{\mathbf{w}}_j\|_2 + \|\hat{\mathbf{w}}_j'\|_2 \geq 0. \tag{103}$$

We also note that

$$\mathbf{X}(\mathbf{w}_0 + \mathbf{w}_{i^*} - \mathbf{w}_{i^*}') = \mathbf{X}\mathbf{w}^*. \tag{104}$$

As $\mathbf{X}^T\mathbf{X}$ is invertible, we have $\mathbf{w}_0 + \mathbf{w}_{i^*} - \mathbf{w}_{i^*}' = \mathbf{w}^*$. Thus, we have

$$\|\mathbf{w}_0\|_2 + \|\mathbf{w}_{i^*}\|_2 + \|\mathbf{w}_{i^*}'\|_2 \geq \|\mathbf{w}^*\|_2. \tag{105}$$

The equality holds when there exists $\alpha_1, \alpha_2, \alpha_3 \geq 0$ such that $\alpha_1 + \alpha_2 + \alpha_3 = 1$, $\mathbf{w}_0 = \alpha_1\mathbf{w}^*$, $\mathbf{w}_{i^*} = \alpha_2\mathbf{w}^*$ and $\mathbf{w}_{i^*}' = -\alpha_3\mathbf{w}^*$. However, as $\mathbf{X}\mathbf{w}^* \geq 0$ does not hold and $\mathbf{X}\mathbf{w}_{i^*} \geq 0$, $\mathbf{X}\mathbf{w}_{i^*}' \geq 0$, we have $\alpha_2 = \alpha_3 = 0$. This implies that $\mathbf{W}^* = (\mathbf{w}^*, 0, \dots, 0)$ is the unique optimal solution to (12).

F. Proof of Theorem 9

PROOF Let $K_j = \{u : D_j u \geq 0\}$ and denote $\mathcal{C} = \{j : \mathrm{tr}(D_j) > n - d\}$. We note that

$$\begin{aligned} &P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0} \mathrm{tr}(\mathbb{I}(\mathbf{Xh} \geq 0)) \leq n - d\right) \\ =&P(\exists \mathbf{w} \in \mathbb{R}^d : \mathbf{Xw} \in \cup_{j\in\mathcal{C}} K_j, \mathbf{w} \neq 0). \end{aligned} \tag{106}$$

As $\cup_{j\in\mathcal{C}} K_j$ is not a convex set, we cannot directly apply the kinematic formula. Let $K = (\cup_{j\in\mathcal{C}} K_j) \cap S^{n-1}$, where $S^{n-1} = \{\mathbf{z} \in \mathbb{R}^n | \|\mathbf{z}\|_2 = 1\}$. It is a closed subset of S^{n-1}. We give the lower bound of the success probability based on the Gordon's escape through a mesh theorem.

Lemma 4 *Let K be a closed subset of S^{n-1}. Define the Gaussian width of K by:*

$$w(K) = \mathbb{E}_{\mathbf{g}\sim\mathcal{N}(0,I_n)}\left[\max_{\mathbf{z}\in K} \mathbf{g}^T\mathbf{z}\right]. \tag{107}$$

Define $a_k = \mathbb{E}_{\mathbf{g}\sim\mathcal{N}(0,I_k)}[\|\mathbf{g}\|_2]$ for $k \in \mathbb{N}$. Then, for a $n - k$ dimensional subspace $L \subseteq \mathbb{R}^n$ drawn at random, we have

$$P(L \cap K \neq \varnothing) \geq 1 - \frac{7}{2} e^{-\frac{1}{18}(a_k - w(K))}. \tag{108}$$

Note that $\mathbf{Xw}$ is a random d-dimensional subspace of $\mathbb{R}^n$. According to the Gordon's escape through a mesh theorem, we have

$$P(\exists \mathbf{w} \in \mathbb{R}^d : \mathbf{Xw} \in \cup_{j\in\mathcal{C}} K_j, \mathbf{w} \neq 0) \geq 1 - \frac{7}{2} e^{-\frac{1}{18}(a_{n-d} - w(K))}. \tag{109}$$

To ensure that there exists $\mathbf{w} \neq 0$ such that $\mathbf{Xw} \in \cup_{j\in\mathcal{C}} K_j$ with high probability, we require that $a_{n-d}^2 > w(K)^2$. As $\frac{k}{k+1}k \leq a_k^2 \leq k$, it is sufficient to have $w(K)^2 < n - d$. Therefore, it suffices to calculate the squared Gaussian width of K. We can compute that

$$\begin{aligned} w(K)^2 =& \left(\mathbb{E}\left[\max_{j\in\mathcal{C}} \max_{\mathbf{z}\in\mathbb{R}^n : D_j\mathbf{z}\geq 0, \|\mathbf{z}\|_2 = 1} \mathbf{g}^T\mathbf{z}\right]\right)^2 \\ =& \left(\mathbb{E}\left[\max_{j\in\mathcal{C}} \|(\mathbf{D}_j\mathbf{g})_+ + ((\mathbf{D}_j - \mathbf{I}_n)\mathbf{g})_+\|_2\right]\right)^2 \\ \leq& \mathbb{E}\left[\left(\max_{j\in\mathcal{C}} \|(\mathbf{D}_j\mathbf{g})_+ + ((\mathbf{D}_j - \mathbf{I}_n)\mathbf{g})_+\|_2\right)^2\right] \\ =& \mathbb{E}\left[\max_{j\in\mathcal{C}} \|(\mathbf{D}_j\mathbf{g})_+\|_2^2 + \|((\mathbf{D}_j - \mathbf{I}_n)\mathbf{g})_+\|_2^2\right]. \end{aligned} \tag{110}$$

By noting that

$$\begin{aligned} &\|(\mathbf{D}_j\mathbf{g})_+\|_2^2 + \|((\mathbf{D}_j - \mathbf{I}_n)\mathbf{g})_+\|_2^2 \\ =&\|(\mathbf{g})_+\|_2^2 - \|((\mathbf{I}_n - \mathbf{D}_j)\mathbf{g})_+\|_2^2 + \|((\mathbf{D}_j - \mathbf{I}_n)\mathbf{g})_+\|_2^2, \end{aligned}$$

we have

$$\begin{aligned} &\mathbb{E}\left[\max_{j\in\mathcal{C}} \|(\mathbf{D}_j\mathbf{g})_+\|_2^2 + \|((\mathbf{D}_j - \mathbf{I}_n)\mathbf{g})_+\|_2^2\right] \\ =&\mathbb{E}\left[\|(\mathbf{g})_+\|_2^2 + \max_{j\in\mathcal{C}} \left(\|((\mathbf{D}_j - \mathbf{I}_n)\mathbf{g})_+\|_2^2 - \|((\mathbf{I}_n - \mathbf{D}_j)\mathbf{g})_+\|_2^2\right)\right] \\ =&\frac{n}{2} + \mathbb{E}\left[\max_{j\in\mathcal{C}} \sum_{i:(\mathbf{D}_j)_{i,i}=0} -\mathrm{sgn}(\mathbf{g}_i)\mathbf{g}_i^2\right] = \frac{n}{2} + \mathbb{E}\left[\max_{S\subseteq[n]:|S|\leq d-1} \sum_{i\in S} \mathrm{sgn}(\mathbf{g}_i)\mathbf{g}_i^2\right] \end{aligned} \tag{111}$$

Let $R = \mathrm{sgn}(G)G^2$, where $G \sim \mathcal{N}(0, 1)$. Denote F_R be the CDF of the random variable R. Suppose that $d, n \to \infty$ with a fixed ratio $d = \theta n$, $\frac{1}{n}\mathbb{E}\left[\max_{S\subseteq[n]:|S|\leq d-1} \sum_{i\in S} \mathrm{sgn}(\mathbf{g}_i)\mathbf{g}_i^2\right]$ will converge to

$$\int_{F_r^{-1}(1-\theta)}^{\infty} r dF_R(r). \tag{112}$$

We note that

$$1 - F_R(r) = P(R \geq r) = \frac{1}{2}P(G^2 \geq r) = \frac{1}{2}(1 - F_{\chi^2}(r)).$$

Therefore, we can rewrite (112) as $\frac{1}{2}\int_{F_{\chi^2}^{-1}(1-2\theta)}^{\infty} r dF_{\chi^2}(r)$. Denote

$$g(\theta) = \frac{1}{2} + \frac{1}{2}\int_{F_{\chi^2}^{-1}(1-2\theta)}^{\infty} r dF_{\chi^2}(r) + \theta. \tag{113}$$

We note that $g(\theta)$ monotonically increases for $\theta \in [0, 1/2)$. By noting that $g(0) = 1/2$ and $g(0.5) = 1.5$, there uniquely exists $\theta^* \in (0, 1/2)$ such that $g(\theta) = 1$. We also note that

$$\lim_{n\to\infty, d=\theta n} \frac{1}{n}\left(w^2(K) + d\right) = g(\theta). \tag{114}$$

Therefore, for sufficiently large n, d with $d < n\theta$, we have $w^2(K) < n - d$, which implies that (23) holds w.h.p..

We present a numerical way to compute $g(\theta)$. Denote $q = F_{\chi^2}^{-1}(1-2\theta)$.By integration by parts, we can compute that

$$\begin{aligned}
&\frac{1}{2}\int_q^\infty r dF_{\chi^2}(r) = -\frac{1}{2}\int_q^\infty r d(1 - F_{\chi^2}(r))\\
=&-\frac{1}{2}r(1-F_{\chi^2}(r))|_q^\infty + \int_q^\infty (1 - F_{\chi^2}(r))dr\\
=&q\theta + \frac{1}{2}\int_q^\infty (1 - F_{\chi^2}(r))dr.
\end{aligned} \tag{115}$$

By the numerical quadrature of the survival function of the χ^2 random variable with 1 degree of freedom, we plot $g(\theta)$ in Figure 14. Note that when $\theta \approx 0.1314$, we have $g(\theta) \approx 1$.

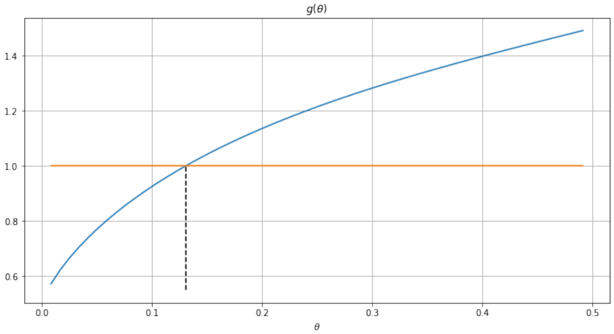


Fig. 14: $g(\theta)$ as a function of θ in $[0, 0.5]$.

G. Proof of Theorem 7

Proof For $\mathbf{h} \in \mathbb{R}^d$ with $\mathbf{h} \neq 0$, denote $A(\mathbf{h}) = \mathbf{X}^T \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\mathbf{X}$. We first prove the case where $n < 2d$. According to the kinematic formula, $P(\exists \mathbf{h} \in \mathbb{R}^d : \mathbf{h} \neq 0, A(\mathbf{h}) = \mathbf{I}_n) \geq 1 - \exp^{-\alpha n}$, where $\alpha = \frac{(n/2-d)^2}{64n^2}$. In other words, there exists an all-ones hyperplane arrangement with probability at least $1 - \exp(-n\alpha)$. In this case, let $\mathbf{D}_j = \mathbf{I}_n$. Construct $\tilde{\mathbf{w}}_i = \mathbf{w}'_i = 0$ if $i \neq j$, $\tilde{\mathbf{w}}_j = \mathbf{w}^*$ and $\tilde{\mathbf{w}}'_j = 0$. Then, $\mathbf{W}' = (0, \tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}'_1, \dots, \tilde{\mathbf{w}}_p, \tilde{\mathbf{w}}'_p)$ is also a feasible solution to (12).

Then, we consider the case where $n > 2d$. We show that with probability at least $1 - \exp(-\alpha n)$, the event (orth-NIC-L) holds. For $\boldsymbol{\sigma} \in \{0,1\}^n$, we let $B(\boldsymbol{\sigma}) = \mathbf{X}^T \mathbf{diag}(\boldsymbol{\sigma})\mathbf{X}$. As $\mathbf{X}^T\mathbf{X} = \mathbf{I}_d$, we note that $\|B(\boldsymbol{\sigma})\mathbf{w}^*\|_2 = 1$ if and only if $\|B(\boldsymbol{\sigma})\|_2 = 1$ and $\mathbf{w}^* \in \text{eigmax}(B(\boldsymbol{\sigma}))$. For $\boldsymbol{\sigma} \in \{0,1\}^n$ with $\boldsymbol{\sigma} \neq \mathbf{1}$ and $\|B(\boldsymbol{\sigma})\|_2 = 1$, as $B(\boldsymbol{\sigma}) \neq I$, we have $\dim(\text{eigmax}(B(\boldsymbol{\sigma}))) \leq d - 1$. This implies that $P(\mathbf{w}^* \in \text{eigmax}(B(\boldsymbol{\sigma})|\|B(\boldsymbol{\sigma})\|_2 = 1) = 0$. Therefore, we have the bound:

$$\begin{aligned}
&P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0} \|A(\mathbf{h})\mathbf{w}^*\|_2 = 1\right)\\
=&P\left(\max_{\boldsymbol{\sigma}:\exists \mathbf{h}\neq 0, \boldsymbol{\sigma}=\mathbb{I}(\mathbf{X}\mathbf{h}\geq 0)} \|B(\boldsymbol{\sigma})\mathbf{w}^*\|_2 = 1\right)\\
\leq&\sum_{\boldsymbol{\sigma}\in\{0,1\}^n} P(\|B(\boldsymbol{\sigma})\mathbf{w}^*\|_2 = 1, \exists \mathbf{h} \neq 0, \boldsymbol{\sigma} = \mathbb{I}(\mathbf{X}\mathbf{h} \geq 0))\\
\leq&P(\exists \mathbf{h} \neq 0, \mathbf{1} = \mathbb{I}(\mathbf{X}\mathbf{h} \geq 0)) + \sum_{\boldsymbol{\sigma}\in\{0,1\}^n, \boldsymbol{\sigma}\neq\mathbf{1}} P(\|B(\boldsymbol{\sigma})\mathbf{w}^*\|_2 < 1)\\
=&P(\exists \mathbf{h} \neq 0, \mathbf{1} = \mathbb{I}(\mathbf{X}\mathbf{h} \geq 0)).
\end{aligned} \tag{116}$$

According to the kinematic formula, for $n > 2d$, $P(\exists \mathbf{h} \neq 0, \mathbf{1} = \mathbb{I}(\mathbf{X}\mathbf{h} \geq 0)) \leq \exp^{-\alpha n}$. This implies that

$$P\left(\max_{\mathbf{h}\in\mathbb{R}^d, \mathbf{h}\neq 0} \|A(\mathbf{h})\mathbf{w}^*\|_2 < 1\right) = 1 - P\left(\max_{\mathbf{h}\in\mathbb{R}^d, \mathbf{h}\neq 0} \|A(\mathbf{h})\mathbf{w}^*\|_2 = 1\right) \geq 1 - \exp^{-\alpha n}.$$

Therefore for $n > 2d$, the neural isometry condition (orth-NIC-L) holds with probability at least $1 - \exp(-n\alpha)$. From Proposition 13, the neural isometry condition (orth-NIC-L) implies that $\mathbf{W}$ is the unique optimal solution to (12).

H. Proof of Proposition 21

PROOF To ensure that $\mathbf{W}$ is an optimal solution to (125), we only require that the KKT conditions (117) at $\mathbf{W}$ are satisfied, i.e.,

$$\begin{aligned}
&\left\|\mathbf{A}_j^T\boldsymbol{\lambda}\right\|_2 \leq \beta, \qquad\qquad \text{if } j \neq i^*,\\
&\mathbf{A}_{i^*}^T\boldsymbol{\lambda} = -\beta\frac{\mathbf{w}_{i^*}}{\|\mathbf{w}_{i^*}\|_2},\\
&\mathbf{A}_{i^*}(\mathbf{w}_{i^*} - \mathbf{w}^*) - \mathbf{z} = \boldsymbol{\lambda}.
\end{aligned} \tag{117}$$

The last two equations give

$$\mathbf{A}_{i^*}^T\mathbf{A}_{i^*}\mathbf{w}_{i^*} + \beta\frac{\mathbf{w}_{i^*}}{\|\mathbf{w}_{i^*}\|_2} = \mathbf{A}_{i^*}^T\mathbf{A}_{i^*}\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}. \tag{118}$$

Let us write $\mathbf{w}_{i^*} = r\mathbf{w}$ where $r = \|\mathbf{w}_{i^*}\|_2$ and $\mathbf{w} = \frac{\mathbf{w}_{i^*}}{\|\mathbf{w}_{i^*}\|_2}$. Then, the above expression gives

$$(r\mathbf{A}_{i^*}^T\mathbf{A}_{i^*} + \beta\mathbf{I})\mathbf{w} = \mathbf{A}_{i^*}^T\mathbf{A}_{i^*}\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}. \tag{119}$$

As $r\mathbf{A}_{i^*}^T\mathbf{A}_{i^*} + \beta\mathbf{I}$ is invertible, we obtain an explicit solution for $\mathbf{w}$, i.e.,

$$\mathbf{w} = (r\mathbf{A}_{i^*}^T\mathbf{A}_{i^*} + \beta\mathbf{I})^{-1}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*}\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}).$$

Because $\mathbf{w}$ satisfies $\|\mathbf{w}\|_2 = 1$, the scalar r shall satisfy

$$1 = \left\|(r\mathbf{A}_{i^*}^T\mathbf{A}_{i^*} + \beta\mathbf{I})^{-1}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*}\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z})\right\|_2. \tag{120}$$

Because $\mathbf{A}_{i^*}^T\mathbf{A}_{i^*} = \mathbf{I}$, we have $r + \beta = \|\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}\|_2$. This implies that $r = \|\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}\|_2 - \beta$. As $r \geq 0$, we require that $\beta \leq \|\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}\|_2$. A sufficient condition is that $\beta \leq \|\mathbf{w}^*\|_2 - \|\mathbf{z}\|_2$. We can write the expression of $\boldsymbol{\lambda}$ as

$$\begin{aligned}
\boldsymbol{\lambda} =&\mathbf{A}_{i^*}(\mathbf{w}_{i^*} - \mathbf{w}^*) - \mathbf{z}\\
=&\mathbf{A}_{i^*}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*} + \beta/r\mathbf{I})^{-1}(\mathbf{A}_{i^*}^T\mathbf{A}_{i^*}\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}) - \mathbf{A}_{i^*}\mathbf{w}^* - \mathbf{z}\\
=&\frac{r}{r+\beta}\mathbf{A}_{i^*}\mathbf{w}^* - \mathbf{A}_{i^*}\mathbf{w} + \frac{r}{r+\beta}\mathbf{A}_{i^*}\mathbf{A}_{i^*}^T\mathbf{z} - \mathbf{z}\\
=&-\frac{\beta}{r+\beta}\mathbf{A}_{i^*}\mathbf{w}^* + \left(\mathbf{I} - \frac{r}{r+\beta}\mathbf{A}_{i^*}\mathbf{A}_{i^*}^T\right)\mathbf{z}.
\end{aligned} \tag{121}$$

Because $\mathbf{A}_{i^*}\mathbf{A}_{i^*}^T$ is a projection matrix whose eigenvalues are 0 and 1, we have $\|\mathbf{I} - r/(r+\beta)\mathbf{A}_{i^*}\mathbf{A}_{i^*}^T\|_2 \leq 1$. Therefore, for $j \neq i^*$, we have the upper bound

$$\begin{aligned}
\|\mathbf{A}_j^T\boldsymbol{\lambda}\|_2 &= \left\|-\frac{\beta}{r+\beta}\mathbf{A}_j^T\mathbf{A}_{i^*}\mathbf{w}^* + \mathbf{A}_j^T\left(\mathbf{I} - \frac{r}{r+\beta}\mathbf{A}_{i^*}\mathbf{A}_{i^*}^T\right)\mathbf{z}\right\|_2\\
&\leq \frac{\beta}{r+\beta}\|\mathbf{A}_j^T\mathbf{A}_{i^*}\mathbf{w}^*\|_2 + \|\mathbf{A}_j\|_2\|\mathbf{z}\|_2\\
&= \frac{\beta}{\left\|\mathbf{w}^* + \mathbf{A}_{i^*}^T\mathbf{z}\right\|_2}\|\mathbf{A}_j^T\mathbf{A}_{i^*}\mathbf{w}^*\|_2 + \|\mathbf{A}_j\|_2\|\mathbf{z}\|_2\\
&\leq \frac{\beta}{\|\mathbf{w}^*\|_2 - \|\mathbf{A}_{i^*}^T\mathbf{z}\|_2}\|\mathbf{A}_j^T\mathbf{A}_{i^*}\mathbf{w}^*\|_2 + \|\mathbf{z}\|_2\\
&\leq \frac{\beta}{\|\mathbf{w}^*\|_2 - \|\mathbf{z}\|_2}\|\mathbf{A}_j^T\mathbf{A}_{i^*}\mathbf{w}^*\|_2 + \|\mathbf{z}\|_2\\
&\leq \beta\frac{1-\gamma}{1 - \|\mathbf{z}\|_2/\|\mathbf{w}^*\|_2} + \|\mathbf{z}\|_2 \leq \beta.
\end{aligned} \tag{122}$$

Here we utilize that $\|\mathbf{A}_j\|_2 \leq 1$ and

$$\beta \geq \|\mathbf{z}\|_2\frac{1 - \|\mathbf{z}\|_2/\|\mathbf{w}^*\|_2}{\gamma - \|\mathbf{z}\|_2/\|\mathbf{w}^*\|_2}. \tag{123}$$

Therefore, it implies that there exists a solution $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ such that $\mathbf{w}_j^* = 0$ for $j \neq i^*$. In addition, we provide the upper bound for the ℓ_2 norm $\|\mathbf{w}_{i^*} - \mathbf{w}^*\|_2$.

$$\begin{aligned}
\|\mathbf{w}_{i^*} - \mathbf{w}^*\|_2 &= \|r\mathbf{w} - \mathbf{w}^*\|_2 \\
&= \left\| \frac{r}{r+\beta} \left(\mathbf{w}^* + \mathbf{A}_{i^*}^T \mathbf{z} \right) - \mathbf{w}^* \right\|_2 \\
&\leq \frac{\beta}{r+\beta} \|\mathbf{w}^*\|_2 + \frac{r}{r+\beta} \left\| \mathbf{A}_{i^*}^T \mathbf{z} \right\|_2 \\
&\leq \frac{\beta \|\mathbf{w}^*\|_2}{\left\| \mathbf{w}^* + \mathbf{A}_{i^*}^T \mathbf{z} \right\|_2} + \|\mathbf{z}\|_2 \\
&\leq \frac{\beta \|\mathbf{w}^*\|_2}{\|\mathbf{w}^*\|_2 - \|\mathbf{z}\|_2} + \|\mathbf{z}\|_2 .
\end{aligned} \tag{124}$$

I. Proof of Theorem 10

In order to prove these theorems, we consider a generic group Lasso problem

$$\min_{\{\mathbf{w}_j\}_{j=1}^k} \quad \left\| \sum_{j=1}^k \mathbf{A}_j \mathbf{w}_j - \mathbf{y} \right\|_2^2 + \beta \sum_{j=1}^k \|\mathbf{w}_j\|_2 . \tag{125}$$

The next result provides a sufficient condition on the regularization parameter and the norm of the noise component to ensure successful support recovery, as well as an estimation of the upper bound on the ℓ_2 distance between the optimal solution and the embedded neuron.

Proposition 21 *Let $\mathbf{y} = \mathbf{A}_{i^*} \mathbf{w}^* + \mathbf{z}$, where $\mathbf{A}_{i^*}$ satisfies $(\mathbf{A}_{i^*})^T \mathbf{A}_{i^*} = \mathbf{I}_d$. Assume that $\|\mathbf{A}_j\|_2 \leq 1$ for $j \neq i^*$. Suppose that the following condition holds.*

$$\max_{j \neq i^*} \|\mathbf{A}_j^T \mathbf{A}_{i^*} \mathbf{w}^*\|_2 \leq (1-\gamma) \|\mathbf{w}^*\|_2$$

for a certain scalar constant $\gamma > 0$. Further, suppose that $\|\mathbf{z}\|_2 \leq \frac{\gamma}{2} \|\mathbf{w}^*\|_2$. Then, for $\beta \in \left[\|\mathbf{z}\|_2 \frac{\|\mathbf{w}^*\|_2 - \|\mathbf{z}\|_2}{\gamma \|\mathbf{w}^*\|_2 - \|\mathbf{z}\|_2}, \|\mathbf{w}^*\|_2 - \|\mathbf{z}\|_2 \right]$, there exists a solution $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$ such that $\mathbf{w}_j^* = 0$ for $j \neq i^*$. Moreover, the ℓ_2 norm $\|\mathbf{w}_{i^*} - \mathbf{w}^*\|_2$ is bounded as follows:

$$\|\mathbf{w}_{i^*} - \mathbf{w}^*\|_2 \leq \frac{\beta \|\mathbf{w}^*\|_2}{\|\mathbf{w}^*\|_2 - \|\mathbf{z}\|_2} + \|\mathbf{z}\|_2 .$$

In order to apply Proposition 21, we need to estimate the upper bound of

$$\max_{\mathbf{h} \in \mathbb{R}^d : \mathbf{h} \neq 0} \left\| \mathbf{U}^T \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0)) \mathbf{U} \right\|_2 .$$

From Lemma 3, by taking $t = 1/4$, for $n \geq 4000\sigma^2 d \log(54n)$, we have

$$P\left(\max_{\mathbf{h} \in \mathbb{R}^d : \mathbf{h} \neq 0} \|\mathbf{X}^T \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{h} \geq 0)) \mathbf{X}\|_2 \leq \frac{3}{4} \right) \geq 1 - 4 \exp\left(-\frac{n}{8000\sigma^2} \right) . \tag{126}$$

Note that $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, therefore

$$\left\| \mathbf{X}^T \mathbf{D}_j \mathbf{X} \right\|_2 = \left\| \mathbf{\Sigma} \mathbf{U}^T \mathbf{D}_j \mathbf{U} \mathbf{\Sigma} \right\|_2 \geq \sigma_{\min}^2 \left\| \mathbf{U}^T \mathbf{D}_j \mathbf{U} \right\|_2 , \tag{127}$$

where $\sigma_{\min}$ is the smallest singular value of $\mathbf{X}$.

On the other hand, from Theorem 4.6.1 in [47], we have the following lower bound.

$$P\left(\sigma_{\min} \geq 1 - 2\sqrt{\frac{d}{n}} \right) \geq 1 - 2\exp(-d). \tag{128}$$

For $n \geq 1024d$, we have

$$P\left(\sigma_{\min}^2 \geq 7/8 \right) \geq P\left(\sigma_{\min} \geq 1 - 1/16 \right) \geq P\left(\sigma_{\min} \geq 1 - 2\sqrt{\frac{d}{n}} \right) \geq 1 - 2\exp(-d). \tag{129}$$

This implies that for n satisfying $n \geq \max\{1024d, 4000\sigma^2 d \log(54n)\}$, we have

$$\begin{aligned}
&P\left(\max_{j\in[p]} \left\|\mathbf{X}^T\mathbf{D}_j\mathbf{X}\right\|_2 \leq \frac{3}{4}\right) \leq P\left(\sigma_{\min}^2 \max_{j\in[p]} \left\|\mathbf{U}^T\mathbf{D}_j\mathbf{U}\right\|_2 \leq \frac{3}{4}\right)\\
\leq &P\left(\sigma_{\min}^2 \max_{j\in[p]} \left\|\mathbf{U}^T\mathbf{D}_j\mathbf{U}\right\|_2 \leq \frac{3}{4} \mid \sigma_{\min}^2 \geq \frac{7}{8}\right) P\left(\sigma_{\min}^2 \geq \frac{7}{8}\right)\\
&+ P\left(\sigma_{\min}^2 \max_{j\in[p]} \left\|\mathbf{U}^T\mathbf{D}_j\mathbf{U}\right\|_2 \leq \frac{3}{4} \mid \sigma_{\min}^2 \leq \frac{7}{8}\right) P\left(\sigma_{\min}^2 \leq \frac{7}{8}\right)\\
\leq &P\left(\max_{j\in[p]} \left\|\mathbf{U}^T\mathbf{D}_j\mathbf{U}\right\|_2 \leq \frac{6}{7}\right) P\left(\sigma_{\min}^2 \geq \frac{7}{8}\right) + 1 - P\left(\sigma_{\min}^2 \geq \frac{7}{8}\right).
\end{aligned} \tag{130}$$

Therefore, we calculate that

$$\begin{aligned}
&P\left(\max_{j\in[p]} \left\|\mathbf{U}^T\mathbf{D}_j\mathbf{U}\right\|_2 \leq \frac{6}{7}\right)\\
\geq &1 - 4\exp\left(-\frac{n}{8000\sigma^2}\right) / P\left(\sigma_{\min}^2 \geq \frac{7}{8}\right)\\
\geq &1 - 4\exp\left(-\frac{n}{8000\sigma^2}\right) / (1 - 2\exp(-d))\\
\geq &1 - 2\exp(-d) - 4\exp\left(-\frac{n}{8000\sigma^2}\right)
\end{aligned} \tag{131}$$

Applying Proposition 21 with $\gamma = \frac{1}{7}$ and $\mathbf{A}_{i^*} = \mathbf{U}, \mathbf{A}_j = \mathbf{D}_j\mathbf{U}, \mathbf{w}^* = \mathbf{\Sigma}\mathbf{V}^T\mathbf{w}^*$, we conclude that if the assumptions in Theorem 10 is satisfied, then with probability at least $1 - 4\exp(-n/8000\sigma^2) - 2\exp(-d)$ there exists $\mathbf{W} = (\mathbf{w}, 0, \dots, 0)$ such that $\mathbf{W}$ is the optimal solution to both (24) and (25) whenever $n \geq \max\{4000\sigma^2 d \log(54n), 1024d\}$.

Moreover, we obtain the desired upper bound

$$\left\|\mathbf{w} - \mathbf{\Sigma}\mathbf{V}^T\mathbf{w}^*\right\|_2 \leq \frac{\beta\eta}{\eta - \|\mathbf{z}\|_2} + \|\mathbf{z}\|_2.$$

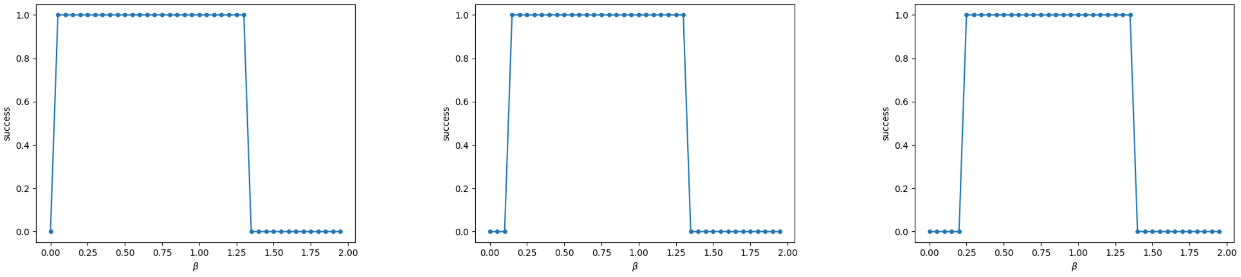
Finally, we provide high probability upper and lower bounds for $\eta = \|\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}^*\|_2$ as follows. Again from Theorem 4.6.1 in [47], we know that for $n \geq 1024d$, with probability at least $1 - 2\exp(-d)$ it holds that

$$1 - 1/16 \leq \sigma_{\min} \leq \sigma_{\max} \leq 1 + 1/16, \tag{132}$$

which immediately implies that $(1 - 1/16)\|\mathbf{w}^*\|_2 \leq \eta \leq (1 + 1/16)\|\mathbf{w}^*\|_2$.

J. Numerical Verification

In this subsection, we numerically verify Theorem 10. We take $n = 40, d = 10$ and test for $\sigma = 0, 1/8, 1/4$. For each σ, we solve the regularized group Lasso problem (25) for $\beta \in [0, 2]$. Then we analyze the solution and record the number of active neurons. The recovery is regarded as success if there is exactly one active neuron. In Figure 15, the recovery displays a failure-success-failure pattern when β increases. Besides, the lower bound of β that ensures successful recovery shifts right as σ increases, while the upper bound generally remains the same.



(a) $\sigma = 0$ (b) $\sigma = 1/8$ (c) $\sigma = 1/4$

Fig. 15: The pattern of successful recovery of one planted linear neuron by solving regularized group Lasso problem (125) derived from training ReLU networks with skip connections.

Appendix J
Proofs in Section III-C

A. Proof of Theorem 11

Proof We first introduce an auxillary lemma:

Lemma 5 *Suppose that $\mathbf{U}_i, \mathbf{U}_j \in \mathbb{R}^{n\times d}$ are column orthonormal, i.e., $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}_d$ and $\mathbf{U}_j^T\mathbf{U}_j = \mathbf{I}_d$. If $\mathbf{U}_i^T\mathbf{U}_j = \mathbf{I}_d$, then, we have $\mathbf{U}_i = \mathbf{U}_j$.*

Proof For $k \in [d]$, we let $\mathbf{w}_{i,k}$ and $\mathbf{w}_{j,k}$ be the k-th column of $\mathbf{U}_i$ and $\mathbf{U}_j$. Note that $1 = (\mathbf{U}_i^T\mathbf{U}_j)_{k,k} = \mathbf{w}_{i,k}^T\mathbf{w}_{j,k}$. As $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}_d$ and $\mathbf{U}_j^T\mathbf{U}_j = \mathbf{I}_d$, we have $\|\mathbf{w}_{i,k}\|_2 = \|\mathbf{w}_{j,k}\|_2 = 1$. Therefore, $\mathbf{w}_{i,k}^T\mathbf{w}_{j,k} = 1$ implies that $\mathbf{w}_{i,k} = \mathbf{w}_{j,k}$. Hence, we have $\mathbf{U}_i = \mathbf{U}_j$.

Consider the event

$$E_1 = \left\{\sum_{i=1}^n \mathbb{I}(\mathbf{x}_i^T\tilde{\mathbf{w}}^* \geq 0) \geq d\right\}. \tag{133}$$

We first show that for $n > 2d$, E_1 holds with high probability. As $\mathbf{x}_i \sim \mathcal{N}(0, I_d/n)$, $\mathbb{I}(\mathbf{x}_i^T\mathbf{w}^* \geq 0)$ are i.i.d. random variables following $\mathbf{Bern}(1/2)$. Denote $B_i = \mathbb{I}(\mathbf{x}_i^T\mathbf{w}^* \geq 0)$ and let $B = \frac{1}{n}\sum_{i=1}^n B_i$. We note that $\mathbb{E}[B] = \frac{1}{2}$. According to the Chernoff bound, we have

$$P(E_1^c) = P\left(B < \frac{d}{n}\right) = P\left(B < \left(1 - \frac{n-2d}{n}\right)\mathbb{E}[B]\right) \leq \exp\left(-\frac{1}{6}\left(\frac{n-2d}{n}\right)^2 n\right). \tag{134}$$

This implies that $P(E_1) \geq 1 - \exp\left(-\frac{1}{6}\left(\frac{n-2d}{n}\right)^2 n\right)$.

Denote $E_2 = \{\sigma_{\min}\left(\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X}\right) > 0\}$, where $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbb{I}(\mathbf{X}\mathbf{w}^* \geq 0))$. We note that

$$\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\mathbb{I}(\mathbf{x}_i^T\mathbf{w}^* \geq 0).$$

As $\mathbf{x}_i \sim \mathcal{N}(0, I_d/n)$, we have $P(E_2^c|E_1) = \frac{P(E_2^c, E_1)}{P(E_1)} = 0$, which implies that $P(E_2|E_1) = 1$.

Conditioned on the event E_2, we have $\mathbf{U}_{i^*} \in \mathbb{R}^{n\times d}$. For $j \neq i^*$, we show that $\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*} \neq \mathbf{I}_d$. Let $\mathbf{U}_j \in \mathbb{R}^{n\times r_j}$. If $r_j < d$, then $\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}$ is with rank at most $r_j < d$. Hence, we have $\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*} \neq \mathbf{I}_d$. If $r_j = d$, suppose that $\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*} = \mathbf{I}_d$. Denote $\mathbf{P} = \mathbf{U}_j^T\mathbf{U}_{i^*} \in \mathbb{R}^{d\times d}$. Then, $\mathbf{P}^T\mathbf{P} = \mathbf{I}_d$. This implies that $\mathbf{P}$ is orthogonal. Hence, $\mathbf{U}_j^T(\mathbf{U}_{i^*}\mathbf{P}^T) = \mathbf{P}\mathbf{P}^T = \mathbf{I}_d$. We note that $\mathbf{U}_{i^*}\mathbf{P}^T$ is also column orthonormal. This is because

$$(\mathbf{U}_{i^*}\mathbf{P}^T)^T\mathbf{U}_{i^*}\mathbf{P}^T = \mathbf{P}\mathbf{U}_{i^*}^T\mathbf{U}_{i^*}\mathbf{P}^T = \mathbf{P}\mathbf{P}^T = \mathbf{I}_d.$$

As the matrices $\mathbf{U}_j$ and $\mathbf{U}_{i^*}\mathbf{P}^T$ are column orthonormal and $\mathbf{U}_j^T(\mathbf{U}_{i^*}\mathbf{P}^T) = \mathbf{I}_d$, from Lemma 5, we have $\mathbf{U}_j = \mathbf{U}_{i^*}\mathbf{P}^T$. As $\mathbf{D}_{i^*} \neq \mathbf{D}_j$, there exists $k \in [n]$ such that either of following statements will hold.

- $(\mathbf{D}_{i^*})_{k,k} = 1$ and $(\mathbf{D}_j)_{k,k} = 0$.
- $(\mathbf{D}_{i^*})_{k,k} = 0$ and $(\mathbf{D}_j)_{k,k} = 1$.

For the first case, we note that

$$\mathbf{w}_{i^*,k} = (\mathbf{D}_i)_{k,k}\mathbf{x}_k^T\mathbf{V}_{i^*}\mathbf{\Sigma}_{i^*}, \mathbf{w}_{j,k} = (\mathbf{D}_j)_{k,k}\mathbf{x}_k^T\mathbf{V}_j\mathbf{\Sigma}_j. \tag{135}$$

Because $\mathbf{x}_k \neq 0$ and $\mathbf{V}_{i^*}\mathbf{\Sigma}_{i^*}$ is invertible, we have $\|\mathbf{w}_{i^*,k}\|_2 > 0$. As $\mathbf{U}_{i^*} = \mathbf{U}_j\mathbf{P}^T$, we have $\mathbf{w}_{i^*,k}^T = \mathbf{w}_{j,k}^T\mathbf{P}^T$. As $\mathbf{w}_{j,k} = 0$, this implies that $\mathbf{w}_{i^*,k} = 0$, which leads to a contradiction. For the second case, we note that $\mathbf{U}_{i^*} = \mathbf{U}_j\mathbf{P}$. Similarly, this will lead to a contradiction.

Therefore, conditioned on E_2, for $j \neq i^*$, we have $\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*} \neq \mathbf{I}_d$. We note that $\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*} \succeq \mathbf{U}_{i^*}^T\mathbf{U}_{i^*} = \mathbf{I}_d$. This implies that $\|\mathbf{U}_j^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 = 1$ if and only if we have $\|\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\|_2 = 1$ and $\tilde{\mathbf{w}}^* \in \text{eigmax}\{\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\}$. As $\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*} \neq \mathbf{I}_d$, if it follows that $\|\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\|_2 = 1$, then $\text{eigmax}\{\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\}$ is with dimension at most $d-1$. In other words, conditioned on $\|\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\|_2 = 1$, $\text{eigmax}\{\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\}$ is a random subspace with dimension at most $d-1$. This implies that

$$\begin{aligned}&P(\tilde{\mathbf{w}}^* \in \text{eigmax}\{\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\}, \|\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\|_2 = 1|E_2)\\ =&P(\mathbf{\Sigma}_{i^*}\mathbf{V}_{i^*}\mathbf{w}^* \in \text{eigmax}\{\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\}, \|\mathbf{U}_{i^*}^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_{i^*}\|_2 = 1|E_2) = 0.\end{aligned} \tag{136}$$

Immediately, for $j \neq i^*$, we have

$$P(\|\mathbf{U}_j^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 = 1) = 0 \tag{137}$$

For $\boldsymbol{\sigma} \in \{0,1\}^n$, define $\mathbf{U}(\boldsymbol{\sigma})$ as the left singular vector of $\mathbf{diag}(\boldsymbol{\sigma})\mathbf{X}$. Then, we have

$$\begin{aligned}
&P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0,\mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbb{I}(\mathbf{X}\tilde{\mathbf{w}}^*\geq 0)} \|\mathbf{U}(\mathbb{I}(\mathbf{Xh}\geq 0))^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 = 1 \middle| E_2\right)\\
=&P\left(\max_{\boldsymbol{\sigma}\in\{0,1\}^n:\exists\mathbf{h}\neq 0,\mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbb{I}(\mathbf{X}\tilde{\mathbf{w}}^*\geq 0),\boldsymbol{\sigma}=\mathbb{I}(\mathbf{Xh}\geq 0)} \|\mathbf{U}(\boldsymbol{\sigma})^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 = 1 \middle| E_2\right)\\
\leq&\sum_{\boldsymbol{\sigma}\in\{0,1\}^n} P(\|\mathbf{U}(\boldsymbol{\sigma})^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 = 1, \mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbb{I}(\mathbf{X}\tilde{\mathbf{w}}^*\geq 0), \boldsymbol{\sigma}=\mathbb{I}(\mathbf{Xh}\geq 0)|E_2) = 0.
\end{aligned}\tag{138}$$

We note that $P(E_2) \geq P(E_1, E_2) = P(E_1) \geq 1 - \exp\left(-\frac{1}{6}\left(\frac{n-2d}{n}\right)^2 n\right)$. This implies that

$$\begin{aligned}
&P\left(\max_{\substack{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0,\\ \mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbb{I}(\mathbf{X}\tilde{\mathbf{w}}^*\geq 0)}} \|\mathbf{U}(\mathbb{I}(\mathbf{Xh}\geq 0))^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 < 1\right)\\
=&P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0,\mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbb{I}(\mathbf{X}\tilde{\mathbf{w}}^*\geq 0)} \|\mathbf{U}(\mathbb{I}(\mathbf{Xh}\geq 0))^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 < 1, E_2\right)\\
=&P\left(\max_{\mathbf{h}\in\mathbb{R}^d:\mathbf{h}\neq 0,\mathbb{I}(\mathbf{Xh}\geq 0)\neq\mathbb{I}(\mathbf{X}\tilde{\mathbf{w}}^*\geq 0)} \|\mathbf{U}(\mathbb{I}(\mathbf{Xh}\geq 0))^T\mathbf{U}_{i^*}\tilde{\mathbf{w}}^*\|_2 < 1 \middle| E_2\right) P(E_2)\\
=&P(E_2) \geq 1 - \exp\left(-\frac{1}{6}\left(\frac{n-2d}{n}\right)^2 n\right).
\end{aligned}\tag{139}$$

This completes the proof.

B. Proof of Proposition 8

PROOF For simplicity, we assume that $\|\mathbf{w}_1\|_2 = \|\mathbf{w}_2\|_2 = \|\mathbf{h}_j\|_2 = 1$. We first consider the case where $\cos\angle(\mathbf{w}_1, \mathbf{w}_2) = -1$. Then, we have $\mathbf{w}_2 = -\mathbf{w}_1$. In this case, we have $\mathbf{D}_1\mathbf{D}_2 = 0$. Hence, it follows that

$$\begin{bmatrix}\mathbf{U}_{s_1}^T\mathbf{U}_{s_1} & \mathbf{U}_{s_1}^T\mathbf{U}_{s_2}\\ \mathbf{U}_{s_2}^T\mathbf{U}_{s_1} & \mathbf{U}_{s_2}^T\mathbf{U}_{s_2}\end{bmatrix} = \begin{bmatrix}\mathbf{I}_d & 0\\ 0 & \mathbf{I}_d\end{bmatrix}.\tag{140}$$

Then, we can simply that

$$\begin{aligned}
&\mathbf{U}_j^T\begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^T\mathbf{U}_1 & \mathbf{U}_1^T\mathbf{U}_2\\ \mathbf{U}_2^T\mathbf{U}_1 & \mathbf{U}_2^T\mathbf{U}_2\end{bmatrix}^{-1}\begin{bmatrix}\tilde{\mathbf{w}}_1\\ \tilde{\mathbf{w}}_2\end{bmatrix}\\
=&\mathbf{U}_j^T\mathbf{U}_1\tilde{\mathbf{w}}_1 + \mathbf{U}_j^T\mathbf{U}_2\tilde{\mathbf{w}}_2\\
=&\boldsymbol{\Sigma}_j^{-1}\mathbf{V}_j^T\mathbf{X}^T\mathbf{D}_j\left(\mathbf{D}_1\mathbf{X}\mathbf{V}_1\boldsymbol{\Sigma}_1^{-1}\frac{\boldsymbol{\Sigma}_1\mathbf{V}_1^T\mathbf{w}_1}{\|\boldsymbol{\Sigma}_1\mathbf{V}_1^T\mathbf{w}_1\|_2}\right) + \boldsymbol{\Sigma}_j^{-1}\mathbf{V}_j^T\mathbf{X}^T\mathbf{D}_j\left(\mathbf{D}_2\mathbf{X}\mathbf{V}_2\boldsymbol{\Sigma}_2^{-1}\frac{\boldsymbol{\Sigma}_2\mathbf{V}_2^T\mathbf{w}_2}{\|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\mathbf{w}_2\|_2}\right)\\
=&\frac{1}{\|\boldsymbol{\Sigma}_1\mathbf{V}_1^T\mathbf{w}_1\|_2}\boldsymbol{\Sigma}_j^{-1}\mathbf{V}_j^T\mathbf{X}^T\mathbf{D}_j\mathbf{D}_1\mathbf{X}\mathbf{w}_1 + \frac{1}{\|\boldsymbol{\Sigma}_2\mathbf{V}_2^T\mathbf{w}_2\|_2}\boldsymbol{\Sigma}_j^{-1}\mathbf{V}_j^T\mathbf{X}^T\mathbf{D}_j\mathbf{D}_2\mathbf{X}\mathbf{w}_2
\end{aligned}\tag{141}$$

As $n \to \infty$, $\mathbf{X}^T\mathbf{D}_i\mathbf{D}_j\mathbf{X}$ converges in probability to

$$\mathbf{M}(\mathbf{h}_i, \mathbf{h}_j) = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[\mathbf{x}\mathbf{x}^T\mathbb{I}(\mathbf{x}^T\mathbf{h}_i \geq 0)\mathbb{I}(\mathbf{x}^T\mathbf{h}_j \geq 0)].\tag{142}$$

As $n \to \infty$, we have $\mathbf{X}^T\mathbf{D}_j\mathbf{X} \xrightarrow{p} \frac{1}{2}\mathbf{I}_d$ and $\boldsymbol{\Sigma}_j \xrightarrow{p} \frac{1}{\sqrt{2}}\mathbf{I}_d$. Therefore, we have

$$\left\|\mathbf{U}_j^T\mathbf{U}_1\tilde{\mathbf{w}}_1 + \mathbf{U}_j^T\mathbf{U}_2\tilde{\mathbf{w}}_2\right\|_2 \xrightarrow{p} 2\left\|\mathbf{M}(\mathbf{h}_j, \mathbf{w}_1)\mathbf{w}_1 - \mathbf{M}(\mathbf{h}_j, -\mathbf{w}_1)\mathbf{w}_1\right\|_2.\tag{143}$$

According to Lemma 7 in [50], for $\mathbf{h}_i, \mathbf{h}_j$ satisfying $\|\mathbf{h}_i\|_2 = \|\mathbf{h}_j\|_2 = 1$, the matrix $\mathbf{M}(\mathbf{h}_i, \mathbf{h}_j)$ takes the form

$$\mathbf{M}(\mathbf{h}_i, \mathbf{h}_j) = c_1(\gamma)\mathbf{I}_d + c_2(\gamma)(\mathbf{h}_i\mathbf{h}_j^T + \mathbf{h}_j\mathbf{h}_i^T) + c_3(\gamma)(\mathbf{h}_i\mathbf{h}_i^T + \mathbf{h}_j\mathbf{h}_j^T).\tag{144}$$

Here c_1, c_2, c_3 are functions of γ. Then, we have

$$\begin{aligned}
&\mathbf{M}(\mathbf{h}_j, \mathbf{w}_1)\mathbf{w}_1 - \mathbf{M}(\mathbf{h}_j, -\mathbf{w}_1)\mathbf{w}_1\\
=&(c_1(\gamma) + \gamma c_2(\gamma) + c_3(\gamma))\mathbf{w}_1 + 2(c_2(\gamma) + \gamma c_3(\gamma))\mathbf{h}_j\\
&- (c_1(-\gamma) - \gamma c_2(\gamma) - c_3(\gamma))\mathbf{w}_1 + 2(c_2(-\gamma) - \gamma c_3(-\gamma))\mathbf{h}_j.
\end{aligned}\tag{145}$$

We observe that $\|\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i-\mathbf{M}(-\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i\|_2^2$ only depends on γ. Denote $g_1(\gamma)=2\|\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i-\mathbf{M}(-\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i\|_2$. Therefore, it is sufficient to compute the case where $\mathbf{h}_i=\mathbf{e}_1$ and $\mathbf{h}_j=\gamma\mathbf{e}_1+\sqrt{1-\gamma^2}\mathbf{e}_2$. Utilizing Lemma 7, we have

$$\begin{aligned}
g_1(\gamma)^2=&4\left\|\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i-\mathbf{M}(-\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i\right\|_2^2\\
=&4\left(\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1+\sqrt{1-\gamma^2}x_2)x_1^2\right]-\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(-\gamma x_1+\sqrt{1-\gamma^2}x_2)x_1^2\right]\right)^2\\
&+4\left(\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1+\sqrt{1-\gamma^2}x_2)x_1x_2\right]-\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(-\gamma x_1+\sqrt{1-\gamma^2}x_2)x_1x_2\right]\right)^2\\
=&4\left(\int_0^\infty\left(F\left(\frac{\gamma}{\sqrt{1-\gamma^2}}x\right)-F\left(-\frac{\gamma}{\sqrt{1-\gamma^2}}x\right)\right)p(x)x^2dx\right)^2.
\end{aligned}\tag{146}$$

We plot $g_1(\gamma)^2$ in Figure 16.

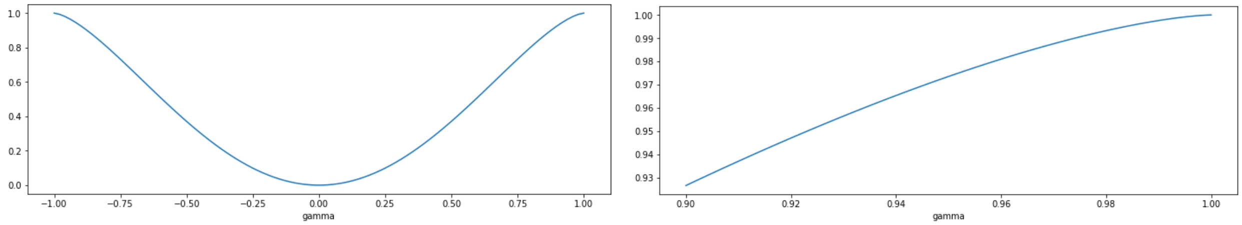


Fig. 16: The plot of $g_1(\gamma)^2$.

Then, we consider the case where $\mathbf{w}_1^T\mathbf{w}_2=0$. In this case,

$$\begin{bmatrix}\mathbf{I}_d & \mathbf{U}_1^T\mathbf{U}_2\\ \mathbf{U}_2^T\mathbf{U}_1 & \mathbf{I}_d\end{bmatrix}=\begin{bmatrix}\mathbf{I}_d & \mathbf{\Sigma}_1^{-1}\mathbf{V}_1^T\mathbf{X}^T\mathbf{D}_1\mathbf{D}_2\mathbf{X}\mathbf{V}_2\mathbf{\Sigma}_2^{-1}\\ \mathbf{\Sigma}_2^{-1}\mathbf{V}_2^T\mathbf{X}^T\mathbf{D}_2\mathbf{D}_1\mathbf{X}\mathbf{V}_1\mathbf{\Sigma}_1^{-1} & \mathbf{I}_d\end{bmatrix}\tag{147}$$

This implies that

$$\begin{bmatrix}\mathbf{V}_1 & 0\\ 0 & \mathbf{V}_2\end{bmatrix}\begin{bmatrix}\mathbf{I}_d & \mathbf{U}_1^T\mathbf{U}_2\\ \mathbf{U}_2^T\mathbf{U}_1 & \mathbf{I}_d\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T & 0\\ 0 & \mathbf{V}_2^T\end{bmatrix}\xrightarrow{p}\begin{bmatrix}\mathbf{I}_d & 2\mathbf{M}(\mathbf{w}_1,\mathbf{w}_2)\\ 2\mathbf{M}(\mathbf{w}_1,\mathbf{w}_2) & \mathbf{I}_d\end{bmatrix}\tag{148}$$

On the other hand, we note that

$$\begin{aligned}
&\mathbf{V}_j^T\mathbf{U}_j^T\begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T & 0\\ 0 & \mathbf{V}_2^T\end{bmatrix}\\
=&\mathbf{V}_j^T\mathbf{\Sigma}_j^{-1}\mathbf{V}_j\mathbf{X}^T\mathbf{D}_j\begin{bmatrix}\mathbf{D}_1\mathbf{X}\mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{V}_1^T & \mathbf{D}_2\mathbf{X}\mathbf{V}_2\mathbf{\Sigma}_2^{-1}\mathbf{V}_2^T\end{bmatrix}\xrightarrow{p}2\begin{bmatrix}\mathbf{M}(\mathbf{h}_j,\mathbf{w}_1) & \mathbf{M}(\mathbf{h}_j,\mathbf{w}_2)\end{bmatrix}.
\end{aligned}\tag{149}$$

We also have

$$\begin{bmatrix}\mathbf{V}_1 & 0\\ 0 & \mathbf{V}_2\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{w}}_1\\ \tilde{\mathbf{w}}_2\end{bmatrix}=\begin{bmatrix}\frac{\mathbf{V}_1^T\mathbf{\Sigma}_1\mathbf{V}_1\mathbf{w}_1}{\|\mathbf{\Sigma}_1\mathbf{V}_1\mathbf{w}_1\|_2} & \frac{\mathbf{V}_2^T\mathbf{\Sigma}_2\mathbf{V}_2\mathbf{w}_2}{\|\mathbf{\Sigma}_2\mathbf{V}_2\mathbf{w}_2\|_2}\end{bmatrix}\xrightarrow{p}\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\end{bmatrix}.\tag{150}$$

Therefore, we have the limit

$$\left\|\mathbf{U}_j\begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{I}_d & \mathbf{U}_1^T\mathbf{U}_2\\ \mathbf{U}_2^T\mathbf{U}_1 & \mathbf{I}_d\end{bmatrix}^{-1}\begin{bmatrix}\tilde{\mathbf{w}}_1\\ \tilde{\mathbf{w}}_2\end{bmatrix}\right\|_2\xrightarrow{p}2\left\|\begin{bmatrix}\mathbf{M}(\mathbf{h}_j,\mathbf{w}_1) & \mathbf{M}(\mathbf{h}_j,\mathbf{w}_2)\end{bmatrix}\begin{bmatrix}\mathbf{I}_d & 2\mathbf{M}(\mathbf{w}_1,\mathbf{w}_2)\\ 2\mathbf{M}(\mathbf{w}_1,\mathbf{w}_2) & \mathbf{I}_d\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\end{bmatrix}\right\|_2$$

Lemma 6 *Suppose that $\mathbf{w}_1^T\mathbf{w}_2=0$. Then, we have $c_1(0)=\frac{1}{4}, c_2(0)=\frac{1}{2\pi}$ and $c_3(0)=0$. In other words, it follows that*

$$\mathbf{M}(\mathbf{w}_1,\mathbf{w}_2)=\frac{1}{4}\mathbf{I}_d+\frac{1}{2\pi}(\mathbf{w}_1\mathbf{w}_2^T+\mathbf{w}_2^T\mathbf{w}_1).\tag{151}$$

From Lemma 6, we can compute that

$$\begin{bmatrix}\mathbf{I}_d & 2\mathbf{M}(\mathbf{w}_1,\mathbf{w}_2)\\ 2\mathbf{M}(\mathbf{w}_1,\mathbf{w}_2) & \mathbf{I}_d\end{bmatrix}=\begin{bmatrix}\mathbf{I}_d & \frac{1}{2}\mathbf{I}_d+\frac{1}{\pi}(\mathbf{w}_1\mathbf{w}_2^T+\mathbf{w}_2\mathbf{w}_1^T)\\ \frac{1}{2}\mathbf{I}_d+\frac{1}{\pi}(\mathbf{w}_1\mathbf{w}_2^T+\mathbf{w}_2\mathbf{w}_1^T) & \mathbf{I}_n\end{bmatrix}.\tag{152}$$

Suppose that

$$\begin{bmatrix}\mathbf{I}_d & \frac{1}{2}\mathbf{I}_d+\frac{1}{\pi}(\mathbf{w}_1\mathbf{w}_2^T+\mathbf{w}_2^T\mathbf{w}_1)\\ \frac{1}{2}\mathbf{I}_d+\frac{1}{\pi}(\mathbf{w}_1\mathbf{w}_2^T+\mathbf{w}_2^T\mathbf{w}_1) & \mathbf{I}_n\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}=\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\end{bmatrix}.\tag{153}$$

Then, we note that

$$\beta_1+\frac{1}{2}\beta_2+\frac{\mathbf{w}_2^T\beta_2}{\pi}\mathbf{w}_1+\frac{\mathbf{w}_1^T\beta_2}{\pi}\mathbf{w}_2=\mathbf{w}_1,\quad \frac{1}{2}\beta_1+\frac{\mathbf{w}_2^T\beta_1}{\pi}\mathbf{w}_1+\frac{\mathbf{w}_1^T\beta_1}{\pi}\mathbf{w}_2+\beta_2=\mathbf{w}_2.\tag{154}$$

This implies that $\beta_1, \beta_2 \in \text{span}\{\mathbf{w}_1, \mathbf{w}_2\}$. Let $\begin{bmatrix}\beta_1^T \\ \beta_2^T\end{bmatrix} = \mathbf{A}\begin{bmatrix}\mathbf{w}_1^T \\ \mathbf{w}_2^T\end{bmatrix}$, where $\mathbf{A} = \begin{bmatrix}a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2}\end{bmatrix}$. Then, the above linear system reduces to

$$\begin{aligned} &a_{1,1} + \frac{1}{2}a_{2,1} + \frac{1}{\pi}a_{2,2} = 1, \\ &a_{1,2} + \frac{1}{2}a_{2,2} + \frac{1}{\pi}a_{1,2} = 0, \\ &\frac{1}{2}a_{1,1} + \frac{1}{\pi}a_{2,1} + a_{2,1} = 0, \\ &\frac{1}{\pi}a_{1,1} + \frac{1}{2}a_{1,2} + a_{2,2} = 1. \end{aligned} \tag{155}$$

We can solve that

$$a_{1,1} = a_{2,2} = \left(\left(1 + \frac{1}{\pi}\right)^2 - \frac{1}{4}\right)^{-1}\left(1 + \frac{1}{\pi}\right), a_{2,1} = a_{1,2} = \left(\left(1 + \frac{1}{\pi}\right)^2 - \frac{1}{4}\right)^{-1}\frac{1}{2}. \tag{156}$$

In other words, we have $\mathbf{A} = \begin{bmatrix}1 + \frac{1}{\pi} & \frac{1}{2} \\ \frac{1}{2} & 1 + \frac{1}{\pi}\end{bmatrix}^{-1}$.

Denote $\gamma_1 = \mathbf{h}_j^T\mathbf{w}_1$ and $\gamma_2 = \mathbf{h}_j^T\mathbf{w}_2$. As $\mathbf{w}_1^T\mathbf{w}_2 = 0$, we have $\gamma_1^2 + \gamma_2^2 \leq 1$. Thus, we have

$$\begin{aligned} &\begin{bmatrix}\mathbf{M}(\mathbf{h}_j, \mathbf{w}_1) & \mathbf{M}(\mathbf{h}_j, \mathbf{w}_2)\end{bmatrix}\begin{bmatrix}\mathbf{I}_d & 2\mathbf{M}(\mathbf{w}_1, \mathbf{w}_2) \\ 2\mathbf{M}(\mathbf{w}_1, \mathbf{w}_2) & \mathbf{I}_d\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{w}_1 \\ \mathbf{w}_2\end{bmatrix} \\ =&\mathbf{M}(\mathbf{w}_1, \mathbf{h}_j)(a_{1,1}\mathbf{w}_1 + a_{1,2}\mathbf{w}_2) + \mathbf{M}(\mathbf{w}_2, \mathbf{h}_j)(a_{2,1}\mathbf{w}_1 + a_{2,2}\mathbf{w}_2) \\ =&a_{1,1}(c_1(\gamma_1) + \gamma_1 c_2(\gamma_1) + c_3(\gamma_1))\mathbf{w}_1 + 2(c_2(\gamma_1) + \gamma_1 c_3(\gamma_1))\mathbf{h}_j) \\ &+ a_{1,2}(\gamma_2 c_2(\gamma_1)\mathbf{w}_1 + c_1(\gamma_1)\mathbf{w}_2 + \gamma_2 c_3(\gamma_1)\mathbf{h}_j) \\ &+ a_{2,1}(\gamma_1 c_2(\gamma_1)\mathbf{w}_2 + c_1(\gamma_2)\mathbf{w}_1 + \gamma_1 c_3(\gamma_2)\mathbf{h}_j) \\ &+ a_{2,2}(c_1(\gamma_2) + \gamma_2 c_2(\gamma_2) + c_3(\gamma_2)\mathbf{w}_1 + 2(c_2(\gamma_2) + \gamma_2 c_3(\gamma_2))\mathbf{h}_j)). \end{aligned} \tag{157}$$

Similarly, the norm of this quantity only depends on γ_1 and γ_2. Therefore, it is sufficient to consider $\mathbf{w}_1 = \mathbf{e}_1$, $\mathbf{w}_2 = \mathbf{e}_2$ and $\mathbf{h}_j = \gamma_1\mathbf{e}_1 + \gamma_2\mathbf{e}_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}\mathbf{e}_3$. In this case, we note that

$$\begin{aligned} &\mathbf{M}(\mathbf{w}_1, \mathbf{h}_j)(a_{1,1}\mathbf{w}_1 + a_{1,2}\mathbf{w}_2) + \mathbf{M}(\mathbf{w}_2, \mathbf{h}_j)(a_{2,1}\mathbf{w}_1 + a_{2,2}\mathbf{w}_2) \\ =&a_{1,1}\begin{bmatrix}\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1^2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1x_2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1x_3\right]\end{bmatrix} \\ &+ a_{1,2}\begin{bmatrix}\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1x_2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_2^2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_2x_3\right]\end{bmatrix} \\ &+ a_{2,1}\begin{bmatrix}\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_2)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1^2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_2)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1x_2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_2)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1x_3\right]\end{bmatrix} \\ &+ a_{2,2}\begin{bmatrix}\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_2)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_1x_2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_2)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_2^2\right] \\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_2)\sigma'(\gamma_1 x_1 + \gamma_2 x_2 + \sqrt{1 - \gamma_1^2 - \gamma_2^2}x_3)x_2x_3\right]\end{bmatrix} \end{aligned} \tag{158}$$

Let $F(x)$ and $p(x)$ represent the CDF and pdf of $\mathcal{N}(0, 1)$ respectively. We introduce two lemmas for computing the expectations in (158).

Lemma 7 *Suppose that $\gamma \in [-1, 1]$. We have the following computations:*

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1 + \sqrt{1 - \gamma^2}x_2)x_1^2\right] = \int_0^\infty \left(1 - F\left(-\frac{\gamma}{\sqrt{1 - \gamma^2}}x\right)\right)p(x)x^2dx, \\ &\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1 + \sqrt{1 - \gamma^2}x_2)x_1x_2\right] = \frac{1 - \gamma^2}{2\pi}. \end{aligned} \tag{159}$$

Lemma 8 *Suppose that $\gamma_1^2+\gamma_2^2+\gamma_3^2=1$ and $\gamma_3\geq 0$. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\sigma'(x_1)\sigma'(\gamma_1x_1+\gamma_2x_2+\gamma_3x_3)x_1x_2\right]&=\frac{\sqrt{1-\gamma_1^2}|\gamma_2|}{2\pi},\\ \mathbb{E}\left[\sigma'(x_1)\sigma'(\gamma_1x_1+\gamma_2x_2+\gamma_3x_3)x_2^2\right]&=\int_{-\infty}^{\infty}\left(\int_0^{\infty}\left(1-F\left(-\frac{\gamma_1x_1+\gamma_2x_2}{\gamma_3}\right)\right)p(x_1)dx_1\right)p(x_2)x_2^2dx_2,\\ \mathbb{E}\left[\sigma'(x_1)\sigma'(\gamma_1x_1+\gamma_2x_2+\gamma_3x_3)x_2x_3\right]&=\frac{|\gamma_2|\gamma_1\gamma_3}{2\pi\left(\gamma_3^2+\gamma_2^2\right)^{3/2}}.\end{aligned}\tag{160}$$

Denote $\gamma_3=\sqrt{1-\gamma_1^2-\gamma_2^2}$. Hence, we can compute that

$$\begin{aligned}&\mathbf{M}(\mathbf{w}_1,\mathbf{h}_j)(a_{1,1}\mathbf{w}_1+a_{1,2}\mathbf{w}_2)+\mathbf{M}(\mathbf{w}_2,\mathbf{h}_j)(a_{2,1}\mathbf{w}_1+a_{2,2}\mathbf{w}_2)\\ =&a_{1,1}\begin{bmatrix}\int_0^\infty\left(1-F\left(-\frac{\gamma_1}{\sqrt{1-\gamma_1^2}}x\right)\right)p(x)x^2dx\\ \frac{\sqrt{1-\gamma_2^2}|\gamma_1|}{2\pi}\\ \frac{(1-\gamma_1^2)\gamma_3}{2\pi}\end{bmatrix}+a_{1,2}\begin{bmatrix}\frac{\sqrt{1-\gamma_2^2}|\gamma_1|}{2\pi}\\ \int_{-\infty}^\infty\left(\int_0^\infty\left(1-F\left(-\frac{\gamma_1x_1+\gamma_2x_2}{\gamma_3}\right)\right)p(x_1)dx_1\right)p(x_2)x_2^2dx_2\\ \frac{|\gamma_2|\gamma_1\gamma_3}{2\pi\left(\gamma_3^2+\gamma_2^2\right)^{3/2}}\end{bmatrix}\\ &+a_{2,1}\begin{bmatrix}\int_{-\infty}^\infty\left(\int_0^\infty\left(1-F\left(-\frac{\gamma_1x_1+\gamma_2x_2}{\gamma_3}\right)\right)p(x_2)dx_2\right)p(x_1)x_1^2dx_1\\ \frac{\sqrt{1-\gamma_1^2}|\gamma_2|}{2\pi}\\ \frac{|\gamma_1|\gamma_2\gamma_3}{2\pi\left(\gamma_3^2+\gamma_2^2\right)^{3/2}}\end{bmatrix}+a_{2,2}\begin{bmatrix}\frac{\sqrt{1-\gamma_1^2}|\gamma_2|}{2\pi}\\ \int_0^\infty\left(1-F\left(-\frac{\gamma_2}{\sqrt{1-\gamma_2^2}}x\right)\right)p(x)x^2dx\\ \frac{(1-\gamma_1^2)\gamma_3}{2\pi}\end{bmatrix}.\end{aligned}\tag{161}$$

Denote $g_2(\gamma_1,\gamma_2)=2\left\|\mathbf{M}(\mathbf{w}_1,\mathbf{h}_j)(a_{1,1}\mathbf{w}_1+a_{1,2}\mathbf{w}_2)+\mathbf{M}(\mathbf{w}_2,\mathbf{h}_j)(a_{2,1}\mathbf{w}_1+a_{2,2}\mathbf{w}_2)\right\|_2$. We numerically verify that $g_2(\gamma_1,\gamma_2)^2\leq 1$ and it is maximized at $(0,1)$ and $(1,0)$.

We plot $g_2(\gamma_1,\gamma_2)^2$ in Figure 17. We note that $g_2(\gamma_1,\gamma_2)^2$ is maximized at the boundary. Hence, we plot $g_2(\cos\theta,\sin\theta)$ for $\theta\in\left[-\frac{\pi}{4},\frac{3\pi}{4}\right]$ in Figure 18. We note that $g_2(\cos\theta,\sin\theta)$ is maximized at $\theta=0$ or $\theta=\pi/2$. Therefore, $g_2(\gamma_1,\gamma_2)$ is maximized at $(0,1)$ and $(1,0)$ and the optimal value is 1.

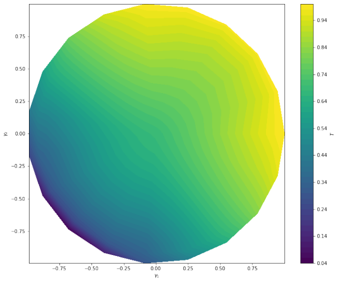


Fig. 17: The contour plot of $g_2(\gamma_1,\gamma_2)^2$.

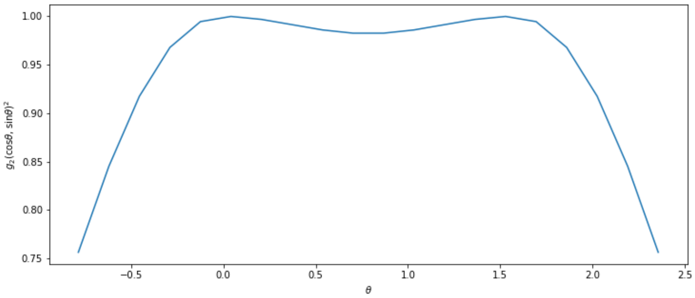


Fig. 18: The plot of $g_2(\cos\theta,\sin\theta)^2$.

C. Proof of Lemma 6

PROOF Consider the rotation matrix $\mathbf{P}\in\mathbb{R}^{d\times d}$ such that $\mathbf{P}\mathbf{w}_1=\mathbf{e}_1$ and $\mathbf{P}\mathbf{w}_2=\mathbf{e}_2$. Then, we have

$$\begin{aligned}&\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)\\ =&\mathbb{E}[\mathbf{x}\mathbf{x}^T\mathbb{I}(\mathbf{x}^T\mathbf{h}_i\geq 0)\mathbb{I}(\mathbf{x}^T\mathbf{h}_j\geq 0)]\\ =&\mathbb{E}[\mathbf{x}\mathbf{x}^T\mathbb{I}(\mathbf{x}^T\mathbf{P}^T\mathbf{P}\mathbf{h}_i\geq 0)\mathbb{I}(\mathbf{x}^T\mathbf{P}^T\mathbf{P}\mathbf{h}_j\geq 0)]\\ =&\mathbf{P}^T\mathbb{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T\mathbb{I}(\tilde{\mathbf{x}}^T\mathbf{P}\mathbf{h}_i\geq 0)\mathbb{I}(\tilde{\mathbf{x}}^T\mathbf{P}\mathbf{h}_j\geq 0)]\mathbf{P}\\ =&\mathbf{P}^T\mathbf{M}(\mathbf{P}\mathbf{h}_i,\mathbf{P}\mathbf{h}_j)\mathbf{P}\end{aligned}\tag{162}$$

where we write $\tilde{\mathbf{x}}=\mathbf{P}\mathbf{x}$. Thus, it is sufficient to compute $\mathbf{M}(\mathbf{P}\mathbf{h}_i,\mathbf{P}\mathbf{h}_j)=\mathbf{M}(\mathbf{e}_1,\mathbf{e}_2)$. We note that

$$\mathbf{M}(\mathbf{e}_1,\mathbf{e}_2)=\begin{bmatrix}\mathbf{M}_{1:2,1:2} & 0\\ 0 & \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(x_2)x_3^2\right]\mathbf{I}_{d-2}\end{bmatrix},\tag{163}$$

where $\mathbf{M}_{1:2,1:2}$ follows

$$\mathbf{M}_{1:2,1:2}=\begin{bmatrix}\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(x_2)x_1^2\right] & \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(x_2)x_1x_2\right]\\ \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(x_2)x_1x_2\right] & \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(x_2)x_2^2\right]\end{bmatrix}=\begin{bmatrix}\frac{1}{4} & \frac{1}{2\pi}\\ \frac{1}{2\pi} & \frac{1}{4}\end{bmatrix}.\tag{164}$$

This implies that $c_1(0) = \mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(x_2)x_3^2\right] = \frac{1}{4}$. We also note that

$$\mathbf{M}_{1:2,1:2} = \begin{bmatrix} u_1(0) + u_3(0) & u_2(0) \\ u_2(0) & u_1(0) + u_3(0) \end{bmatrix}. \tag{165}$$

Hence, we have $c_1(0) = \frac{1}{4}, c_2(0) = \frac{1}{2\pi}$ and $c_3(0) = 0$.

D. *Proof of Lemma 7*

Proof For the first equation, by integrating w.r.t. x_2 first, we immediately obtain that

$$\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1 + \sqrt{1-\gamma^2}x_2)x_1^2\right] = \int_0^\infty \left(1 - F\left(-\frac{\gamma}{\sqrt{1-\gamma^2}}x\right)\right) p(x)x^2 dx. \tag{166}$$

Note that

$$\begin{aligned} &\int_{-\frac{\gamma}{\sqrt{1-\gamma^2}}x_1}^\infty xp(x)dx = \left(\int_{-\frac{\gamma}{\sqrt{1-\gamma^2}}x_1}^0 + \int_0^\infty\right) xp(x)dx \\ =&\frac{1}{\sqrt{2\pi}}\left(\int_{-\frac{\gamma}{\sqrt{1-\gamma^2}}x_1}^0 + \int_0^\infty\right) e^{-\frac{x^2}{2}} d\frac{x^2}{2} = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{\gamma^2}{2(1-\gamma^2)}x^2}. \end{aligned} \tag{167}$$

Therefore, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1 + \sqrt{1-\gamma^2}x_2)x_1x_2\right] = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{\gamma^2}{2(1-\gamma^2)}x^2} p(x)xdx \\ =&\frac{1}{2\pi}\int_0^\infty e^{-\frac{1}{2(1-\gamma^2)}x^2} xdx = \frac{1-\gamma^2}{2\pi}. \end{aligned} \tag{168}$$

E. *Proof of Lemma 8*

Proof We can compute that

$$\begin{aligned} &\mathbb{E}\left[\sigma'(x_1)\sigma'(\gamma_1x_1 + \gamma_2x_2 + \gamma_3x_3 \geq 0)x_1x_2\right] \\ =&\int_0^\infty \int_{-\infty}^\infty \int_{\left|\frac{\gamma_1x_1+\gamma_3x_3}{\gamma_2}\right|}^\infty x_1x_2p(x_1)p(x_2)p(x_3)dx_1dx_3dx_2 \\ =&\int_0^\infty \int_{-\infty}^\infty x_1p(x_1)p(x_3)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\gamma_1x_1+\gamma_3x_3)^2}{2\gamma_2^2}\right) dx_1dx_3. \end{aligned} \tag{169}$$

Note that

$$\begin{aligned} &\int_{-\infty}^\infty p(x_3)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\gamma_1x_1+\gamma_3x_3)^2}{2\gamma_2^2}\right) dx_3 \\ =&\frac{1}{2\pi}\int_{-\infty}^\infty \exp\left(-\frac{(\gamma_1x_1+\gamma_3x_3)^2 - \gamma_2^2x_3^2}{2\gamma_2^2}\right) dx_3 \\ =&\frac{1}{2\pi}\int_{-\infty}^\infty \exp\left(-\frac{(\gamma_3^2+\gamma_2^2)x_3^2 + 2\gamma_1\gamma_3x_1x_3 + \gamma_1^2x_1^2}{2\gamma_2^2}\right) dx_3 \\ =&\frac{1}{2\pi}\int_{-\infty}^\infty \exp\left(-\frac{(\gamma_3^2+\gamma_2^2)\left(x_3 - \frac{\gamma_1\gamma_3}{\gamma_3^2+\gamma_2^2}x_1\right)^2 + \gamma_1^2x_1^2 - \frac{\gamma_1^2\gamma_3^2}{\gamma_3^2+\gamma_2^2}x_1^2}{2\gamma_2^2}\right) dx_3 \\ =&\frac{1}{2\pi}\int_{-\infty}^\infty \exp\left(-\frac{(\gamma_3^2+\gamma_2^2)\left(x_3 - \frac{\gamma_1\gamma_3}{\gamma_3^2+\gamma_2^2}x_1\right)^2 + \frac{\gamma_1^2\gamma_2^2}{\gamma_3^2+\gamma_2^2}x_1^2}{2\gamma_2^2}\right) dx_3 \\ =&\frac{|\gamma_2|}{\sqrt{2\pi}\sqrt{\gamma_3^2+\gamma_2^2}} \exp\left(-\frac{\gamma_1^2}{2(\gamma_3^2+\gamma_2^2)}x_1^2\right). \end{aligned} \tag{170}$$

Then, we have

$$\begin{aligned}
&\int_0^\infty \int_{-\infty}^\infty x_1 p(x_1)p(x_3)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\gamma_1 x_1+\gamma_3 x_3)^2}{2\gamma_2^2}\right)dx_1 dx_3\\
=&\int_0^\infty \frac{|\gamma_2|}{2\pi\sqrt{\gamma_3^2+\gamma_2^2}}\exp\left(-\frac{\gamma_1^2}{2(\gamma_3^2+\gamma_2^2)}x_1^2-\frac{x_1^2}{2}\right)x_1 dx\\
=&\int_0^\infty \frac{|\gamma_2|}{2\pi\sqrt{\gamma_3^2+\gamma_2^2}}\exp\left(-\frac{1}{2(\gamma_3^2+\gamma_2^2)}x_1^2\right)x_1 dx\\
=&\frac{1-\gamma_1^2}{2\pi}\frac{\gamma_2}{\sqrt{\gamma_3^2+\gamma_2^2}}=\frac{\sqrt{1-\gamma_1^2}\gamma_2}{2\pi}.
\end{aligned} \tag{171}$$

For the second equation, we note that

$$\begin{aligned}
&\mathbb{E}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1+\gamma_2 x_2+\gamma_3 x_3)x_2^2\right]\\
=&\int_{-\infty}^\infty\int_{-\infty}^\infty\int_{\max\{0,-\frac{\gamma_2 x_2+\gamma_3 x_3}{\gamma_1}\}}^\infty p(x_1)p(x_2)p(x_3)x_2^2 dx_2 dx_3 dx_1\\
=&\int_{-\infty}^\infty\left(\int_0^\infty\left(1-F\left(-\frac{\gamma_1 x_1+\gamma_2 x_2}{\gamma_3}\right)\right)p(x_1)dx_1\right)p(x_2)x_2^2 dx_2.
\end{aligned} \tag{172}$$

For the third equation, we have

$$\mathbb{E}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1+\gamma_2 x_2+\gamma_3 x_3)x_2 x_3\right]=\int_0^\infty\int_{-\infty}^\infty x_3 p(x_1)p(x_3)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\gamma_1 x_1+\gamma_3 x_3)^2}{2\gamma_2^2}\right)dx_1 dx_3. \tag{173}$$

Following the previous calculation, we can compute that

$$\begin{aligned}
&\int_{-\infty}^\infty x_3 p(x_3)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\gamma_1 x_1+\gamma_3 x_3)^2}{2\gamma_2^2}\right)dx_3\\
=&\frac{1}{2\pi}\int_{-\infty}^\infty \exp\left(-\frac{(\gamma_3^2+\gamma_2^2)\left(x_3-\frac{\gamma_1\gamma_3}{\gamma_3^2+\gamma_2^2}x_1\right)^2+\frac{\gamma_1^2\gamma_2^2}{\gamma_3^2+\gamma_2^2}x_1^2}{2\gamma_2^2}\right)dx_3\\
=&\frac{|\gamma_2|}{\sqrt{2\pi}\sqrt{\gamma_3^2+\gamma_2^2}}\frac{\gamma_1\gamma_3}{\gamma_3^2+\gamma_2^2}x_1.
\end{aligned} \tag{174}$$

Therefore, we have

$$\mathbb{E}\left[\sigma'(x_1)\sigma'(\gamma_1 x_1+\gamma_2 x_2+\gamma_3 x_3)x_2 x_3\right]=\int_0^\infty \frac{|\gamma_2|}{\sqrt{2\pi}\sqrt{\gamma_3^2+\gamma_2^2}}\frac{\gamma_1\gamma_3}{\gamma_3^2+\gamma_2^2}x_1 p(x_1)dx_1=\frac{|\gamma_2|\gamma_1\gamma_3}{2\pi\left(\gamma_3^2+\gamma_2^2\right)^{3/2}}. \tag{175}$$

Appendix K
Proofs in Section VI

A. Proof of Proposition 9

Proof We denote $\mathbf{h}_i=\frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$. For simplicity, we can assume that $\|\mathbf{h}_j\|_2=1$. Hence, it follows that $\gamma=\mathbf{h}_i^T\mathbf{h}_j$. As $n\to\infty$, $\mathbf{X}^T\mathbf{D}_i\mathbf{D}_j\mathbf{X}$ converges in probability to

$$\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)=\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[\mathbf{x}\mathbf{x}^T\mathbb{I}(\mathbf{x}^T\mathbf{h}_i\geq 0)\mathbb{I}(\mathbf{x}^T\mathbf{h}_j\geq 0)]. \tag{176}$$

According to Lemma 7 in [50], the above expectation takes the form

$$\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)=c_1(\gamma)\mathbf{I}_d+c_2(\gamma)(\mathbf{h}_i\mathbf{h}_j^T+\mathbf{h}_j\mathbf{h}_i^T)+c_3(\gamma)(\mathbf{h}_i\mathbf{h}_i^T+\mathbf{h}_j\mathbf{h}_j^T). \tag{177}$$

Here c_1,c_2,c_3 are functions of γ. As $n\to\infty$, we note that $\mathbf{X}^T\mathbf{D}_i\mathbf{X}\xrightarrow{p}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}[\mathbb{I}(\mathbf{x}^T\mathbf{h}_i\geq 0)\mathbf{x}\mathbf{x}^T]=\frac{1}{2}\mathbf{I}_d$. Thus, as $n\to\infty$, we have

$$T\xrightarrow{p}2\left\|\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i\right\|_2. \tag{178}$$

According to the expression (177), we have

$$\mathbf{M}(\mathbf{h}_i,\mathbf{h}_j)\mathbf{h}_i=(c_1(\gamma)+\gamma c_2(\gamma)+c_3(\gamma))\mathbf{h}_i+(c_2(\gamma)+\gamma c_3(\gamma))\mathbf{h}_j. \tag{179}$$

As $\|\mathbf{h}_i\|_2 = \|\mathbf{h}_j\|_2 = 1$, the quantity $\|\mathbf{M}(\mathbf{h}_i, \mathbf{h}_j)\mathbf{h}_i\|_2$ only depends on $\gamma = \mathbf{h}_i^T\mathbf{h}_j$. Denote $g(\gamma) = 4\|\mathbf{M}(\mathbf{h}_i, \mathbf{h}_j)\mathbf{h}_i\|_2^2$. Thus, we can simply consider $\mathbf{h}_i = \mathbf{e}_1$ and $\mathbf{h}_j = \gamma\mathbf{e}_1 + \sqrt{1-\gamma^2}\mathbf{e}_2$. According to Lemma 7, we can compute that

$$\begin{aligned} g(\gamma)^2 =&4\|\mathbf{M}(\mathbf{h}_i, \mathbf{h}_j)\mathbf{h}_i\|_2^2 \\ =&4\left(\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1 + \sqrt{1-\gamma^2}x_2)x_1^2\right]\right)^2 + 4\left(\mathbb{E}_{\mathbf{x}}\left[\sigma'(x_1)\sigma'(\gamma x_1 + \sqrt{1-\gamma^2}x_2)x_1x_2\right]\right)^2 \\ =&4\left(\int_0^\infty \left(1 - F\left(-\frac{\gamma}{\sqrt{1-\gamma^2}}x\right)\right)p(x)x^2dx\right)^2 + 4\left(\frac{1-\gamma^2}{2\pi}\right)^2. \end{aligned} \tag{180}$$

We plot $g(\gamma)^2$ as a function of γ as follows.

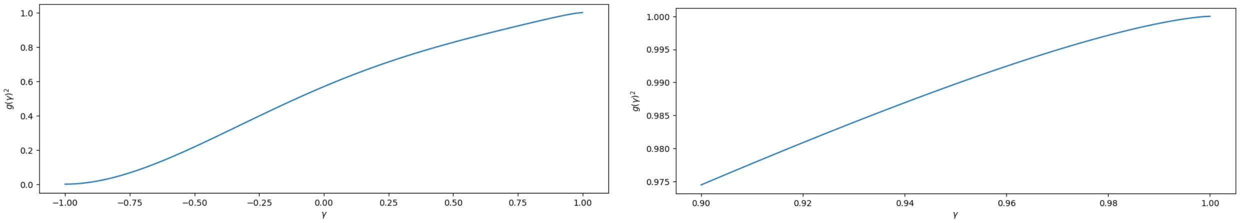


Fig. 19: $g(\gamma)^2$ as a function of γ.

Appendix L
Necessary condition and sufficient condition of the neural isometry condition for ReLU networks

For the recovery of single-neuron ReLU network, we first present a necessary condition for the neural isometry condition (NIC-1).

Definition 7 We say that a diagonal arrangement pattern $\mathbf{D}_i \in H$ satisfies the maximal condition if for all index $j \in [p]$ and $j \neq i$, we have $\mathbf{D}_i\mathbf{D}_j \neq \mathbf{D}_i$.

Lemma 9 *A necessary condition for* (NIC-1) *is that the matrix* $\mathbf{D}_{i^*}$ *satisfies the maximal condition.*

Then, we present a sufficient condition to ensure that the neural isometry condition (NIC-1) holds.

Proposition 22 *Suppose that* $\mathbf{D}_{i^*}$ *satisfies the maximal condition. Assume that the planted neuron* $\mathbf{w}^*$ *satisfies the following conditions*

- $\mathbf{w}^*$ *is the eigenvector corresponding to the largest eigenvalue of* $\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X}$, *i.e.,*
$$\|\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X}\mathbf{w}^*\|_2 = \|\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X}\|_2\|\mathbf{w}^*\|_2.$$
- $\mathbf{x}_l^T\mathbf{w}^* > 0$ *for all* $l \in [n]$ *satisfying* $(\mathbf{D}_{i^*})_{ll} = 1$

Then, the NIC-1 *given in* (NIC-1) *holds.*

Remark 6 The first condition on $\mathbf{w}^*$ requires that $\mathbf{w}^*$ lies in the eigenspace of $\mathbf{X}^T\mathbf{D}_{i^*}\mathbf{X}$ corresponding to its largest eigenvalue. Combining with the maximal condition on $\mathbf{D}_{i^*}$, the second condition on $\mathbf{w}^*$ implies that $\mathbf{w}^*$ lie in the interior of the cone $\{\mathbf{w}|(2\mathbf{D}_{i^*} - \mathbf{I}_n)\mathbf{X}\mathbf{w} \geq 0\}$.

A. Justification of the assumptions in Proposition 22

We justify the assumption on the planted neuron $\mathbf{w}^*$ in Proposition 22 for the Gaussian mixture model. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^d$ and $\sigma > 0$. Suppose that n_1 of the n training samples follow $\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}_d)$ and the rest of $n_2 = n - n_1$ training samples follows follow $\mathcal{N}(\boldsymbol{\mu}_2, \sigma^2\mathbf{I}_d)$. Let $\mathbf{w} \in \{0,1\}^n$ be the vector defined by

$$q_i = \begin{cases} 1, & \text{s.t. } \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}_d), \\ 0, & \text{s.t. } \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_2, \sigma^2\mathbf{I}_d). \end{cases} \tag{181}$$

Proposition 23 *Suppose that* $\sigma^2 > 0$ *and* $b =: \frac{\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_1\|_2\|\boldsymbol{\mu}_2\|_2} < 1$. *With probability at least* $1 - N_1e^{-\frac{(1-b)\|\boldsymbol{\mu}_1\|_2^2}{4\sigma}^2} - N_2e^{-\frac{(1-b)\|\boldsymbol{\mu}_2\|_2^2}{4\sigma^2}}$, *there exists* $i^* \in [p]$ *such that* $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbf{w})$.

Suppose that $\delta \in (0,1)$. From Proposition 23, for $\sigma \leq \frac{\sqrt{1-b}\min\{\|\mu_1\|_2, \|\mu_2\|_2\}}{2\sqrt{\log\left(\frac{n}{\delta}\right)}}$, with probability at least $1-\delta$, there exists $i^* \in [p]$ such that $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbf{w})$. The next question is whether this $\mathbf{D}_{i^*}$ also satisfies the maximal condition.

Proposition 24 *Suppose that there exists $i^* \in [p]$ such that $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbf{w})$. Denote $K^+ = \operatorname{cone}(\{\mathbf{x}_i | q_i = 1, i \in [n]\})$ and $K^- = \operatorname{cone}(\{\mathbf{x}_i | q_i = -1, i \in [n]\})$. Then, D_{i^*} satisfying the maximal condition if and only if $-K^- \subseteq \mathbf{int}\,(K^+)$.*

Finally, we show that the eigenvalue condition will hold for some specific planted neuron.

Proposition 25 *Let $\mathbf{x}_n = \boldsymbol{\mu} + \sigma \mathbf{z}_i$, where $\mathbf{z}_n \sim \mathcal{N}(0, I)$ for $n = 1, \dots, N$. Suppose that $\delta > 0$. Denote $\mathbf{X}^{(1)} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_{n_1}^T \end{bmatrix} \in \mathbb{R}^{n_1 \times d}$. Let $\mathbf{w}$ satisfy that $\|\mathbf{w}\|_2 = 1$ and $\| \left(\mathbf{X}^{(1)}\right)^T \mathbf{X}^{(1)} \mathbf{w}\|_2 = \| \left(\mathbf{X}^{(1)}\right)^T \mathbf{X}^{(1)}\|_2$. Then, with probability at least $1 - \delta$, we have*

$$\left\| \mathbf{w} - \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} \right\|_2^2 \leq c_1 \sigma, \tag{182}$$

where $c_1 = \frac{2d}{n^2 \|\boldsymbol{\mu}\|_2^2} 2n_1 \sqrt{2 \log(n_1 d/\delta)} \|\boldsymbol{\mu}\|_2 + 2d \log(n_1 d/\delta)$ is a constant depending on δ.

The following proposition illustrates that for sufficiently small σ, except for the maximal condition, all conditions in Proposition 22 will hold with probability at least $1 - \delta$.

Theorem 13 *Let $0 < \delta < 2\exp(-d/8)$. Suppose that $0 < \sigma \leq \min\left\{1/(32c_1), \frac{\|\boldsymbol{\mu}\|_2}{2(d + 8\log(4n/\delta))}\right\}$, where c_1 is a constant defined in Proposition 25 Then, there exists $i^* \in \mathcal{P}$ such that $\mathbf{D}_{i^*} = \mathbf{diag}(\mathbf{w})$ with probability at least $1 - \delta$. Let $\mathbf{w}^*$ satisfy that $\|\mathbf{X}^T \mathbf{D}_{i^*} \mathbf{X} \mathbf{w}^*\|_2 = \|\mathbf{X}^T \mathbf{D}_{i^*} \mathbf{X}\|_2 \|\mathbf{w}^*\|_2$. We also have $(2\mathbf{D}_{i^*} - I)\mathbf{X}\mathbf{w}^* > 0$ with probability at least $1 - \delta$. By further assuming that the cone condition in Proposition 24 holds, the neural isometry condition holds, i.e., the problem* (9) *has a unique solution.*

B. Proof of Lemma 9

Proof Suppose that there exists D_j such that $\mathbf{D}_{i^*} \mathbf{D}_j = \mathbf{D}_{i^*}$ and $\mathbf{D}_j \neq \mathbf{D}_{i^*}$. This implies that

$$\mathbf{X}^T \mathbf{D}_j^T \mathbf{D}_i \mathbf{X} = \mathbf{X}^T \mathbf{D}_j \mathbf{D}_{i^*} \mathbf{X} = \mathbf{X}^T \mathbf{D}_{i^*} \mathbf{X}. \tag{183}$$

Thus, we have

$$\left\| \mathbf{X}^T \mathbf{D}_j^T \mathbf{D}_i \mathbf{X} (\mathbf{X}^T \mathbf{D}_{i^*} \mathbf{X})^{-1} \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2} \right\|_2 = \left\| \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2} \right\|_2 = 1. \tag{184}$$

Thus, the irrepresentability condition (NIC-1) is violated.

C. Proof of Proposition 22

Proof Consider any $j \in [p]$ and $j \neq i^*$. Let $\mathcal{I}_j = \{l \in [n] | (\mathbf{D}_{i^*})_{ll} - (\mathbf{D}_j)_{ll} > 0\}$. As $\mathbf{D}_{i^*}$ satisfies the maximal condition, we have $\mathbf{D}_j \mathbf{D}_{i^*} \neq \mathbf{D}_{i^*}$. This implies that $\mathcal{I}_j \neq \varnothing$. Let $k = |\mathcal{I}_j|$. Note that

$$\mathbf{X}^T \mathbf{D}_{i^*} \mathbf{X} = \mathbf{X}^T \mathbf{D}_j \mathbf{D}_{i^*} \mathbf{X} + \sum_{l \in \mathcal{I}} \mathbf{x}_l \mathbf{x}_l^T. \tag{185}$$

For simplicity, we write $\mathbf{w} = \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$, $\mathbf{A} = \mathbf{X}^T \mathbf{D}_{i^*} \mathbf{D}_j \mathbf{X}$ and $\mathbf{B} = \sum_{l \in \mathcal{I}_j} \mathbf{x}_l \mathbf{x}_l^T$. It is sufficient to prove that

$$\|\mathbf{A}(\mathbf{A} + \mathbf{B})^{-1} \mathbf{w}\|_2 < 1. \tag{186}$$

As $\mathbf{w}^*$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{A} + \mathbf{B}$, we have

$$(\mathbf{A} + \mathbf{B})\mathbf{w} = \|\mathbf{A} + \mathbf{B}\|_2 \mathbf{w}, \tag{187}$$

which also implies that $\|\mathbf{A} + \mathbf{B}\|_2^{-1} \mathbf{w} = (\mathbf{A} + \mathbf{B})^{-1} \mathbf{w}$. Therefore, we have

$$\|(\mathbf{A} + \mathbf{B})^{-1} \mathbf{w}\|_2 = \|(\mathbf{A} + \mathbf{B})\|_2^{-1}. \tag{188}$$

As $\mathbf{A}, \mathbf{B}$ are positive semi-definite, we have

$$\|\mathbf{A} + \mathbf{B}\|_2 \geq \|\mathbf{A}\|_2. \tag{189}$$

This implies that

$$\|\mathbf{A}(\mathbf{A} + \mathbf{B})^{-1} \mathbf{w}\|_2 \leq \|\mathbf{A}\|_2 \|(\mathbf{A} + \mathbf{B})^{-1} \mathbf{w}\|_2 = \|\mathbf{A}\|_2 \|\mathbf{A} + \mathbf{B}\|_2^{-1} \leq 1. \tag{190}$$

The equality holds if and only if $\|\mathbf{A} + \mathbf{B}\|_2 = \|\mathbf{A}\|_2$ and $\mathbf{w}$ is the eigenvector of the largest eigenvalue of $\mathbf{A} + \mathbf{B}$. Let $\gamma = \|\mathbf{A} + \mathbf{B}\|_2 = \|\mathbf{A}\|_2$. Then, we have

$$(\mathbf{A} + \mathbf{B})\mathbf{w} = \gamma \mathbf{w}, \mathbf{A}\mathbf{w} = \gamma \mathbf{w}, \tag{191}$$

which implies that $\mathbf{B}\mathbf{w} = 0$. Therefore, we have $\mathbf{w}^T \mathbf{B} \mathbf{w} = 0$, or equivalently $\sum_{l \in \mathcal{I}_j} \mathbf{x}_l^T \mathbf{w} = 0$. As $\mathbf{x}_l^T \mathbf{w} > 0$ for all k satisfying $(\mathbf{D}_{i^*})_{ll} = 1$, we have $\sum_{l \in \mathcal{I}_j} \mathbf{x}_l^T \mathbf{w} > 0$, which leads to a contradiction.

D. Proof of Proposition 23

Proof Let $\mathbf{w}=\frac{\boldsymbol{\mu}_1}{\|\boldsymbol{\mu}_1\|_2}-\frac{\boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_2\|_2}$. We can compute that

$$\|\mathbf{w}\|_2^2=2-2\frac{\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_1\|_2\|\boldsymbol{\mu}_2\|_2}=2(1-b).$$

This implies that $\|\mathbf{w}\|_2=\sqrt{2(1-b)}$. Note that

$$\boldsymbol{\mu}_1^Tw=\|\boldsymbol{\mu}_1\|_2-\frac{\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_2\|_2}=\|\boldsymbol{\mu}_1\|_2-b\|\boldsymbol{\mu}_1\|_2=(1-b)\|\boldsymbol{\mu}_1\|_2>0.$$

Similarly, we have

$$\boldsymbol{\mu}_2^T\mathbf{w}=\frac{\boldsymbol{\mu}_1^T\boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_1\|_2}-\|\boldsymbol{\mu}_2\|_2=-(1-b)\|\boldsymbol{\mu}_2\|_2<0.$$

For index $i\in[n]$ such that $q_i=1$, we can write $\mathbf{x}_i=\boldsymbol{\mu}_1+\sigma\mathbf{z}_i$, where $\mathbf{z}_i\sim\mathcal{N}(0,1)$. We can compute that $\mathbf{x}_i^T\mathbf{w}=\boldsymbol{\mu}_1^T\mathbf{w}+\sigma\mathbf{z}_i^T\mathbf{w}=(1-b)\|\boldsymbol{\mu}_1\|_2+\sigma\mathbf{z}_i^T\mathbf{w}$. According to the tail bound of Gaussian random variable, we have

$$P(\mathbf{x}_i^T\mathbf{w}\leq 0)=P(\sigma\mathbf{z}_i^T\mathbf{w}\leq-(1-b)\|\boldsymbol{\mu}_1\|_2)\leq\exp\left(-\frac{(1-b)^2\|\boldsymbol{\mu}_1\|_2^2}{2\sigma^2\|\mathbf{w}\|_2^2}\right)=\exp\left(-\frac{(1-b)\|\boldsymbol{\mu}_1\|_2^2}{4\sigma^2}\right),$$

which implies that

$$P(\mathbf{x}_n^T\mathbf{w}>0)\geq 1-\exp\left(-\frac{(1-b)\|\boldsymbol{\mu}_1\|_2^2}{4\sigma^2}\right).$$

Therefore, we have

$$P(\mathbf{x}_i^T\mathbf{w}>0,\forall i\text{ with }q_i=1)\geq\left(1-\exp\left(-\frac{(1-b)\|\boldsymbol{\mu}_1\|_2^2}{4\sigma^2}\right)\right)^{n_1}\geq 1-n_1\exp\left(-\frac{(1-b)\|\boldsymbol{\mu}_1\|_2^2}{4\sigma^2}\right).$$

For index $i\in[n]$ such that $q_i=0$, we can write $\mathbf{x}_n=\boldsymbol{\mu}_2+\sigma\mathbf{z}_i$. Similarly, we can compute that

$$P(\mathbf{x}_i^T\mathbf{w}<0,\forall i\text{ with }y_i=0)\geq 1-n_2\exp\left(-\frac{(1-b)\|\boldsymbol{\mu}_2\|_2^2}{4\sigma^2}\right). \tag{192}$$

In summary, we have

$$P((2c_i-1)\mathbf{x}_i^T\mathbf{w}>0,\forall i\in[n])\geq 1-n_1\exp\left(-\frac{(1-b)\|\boldsymbol{\mu}_1\|_2^2}{4\sigma^2}\right)-n_2\exp\left(-\frac{(1-b)\|\boldsymbol{\mu}_2\|_2^2}{4\sigma^2}\right). \tag{193}$$

Under the event $\{q_i\mathbf{x}_i^T\mathbf{w}>0,\forall i\in[n]\}$, the diagonal arrangement pattern $\mathbf{D}_{i^*}$ induced by the vector $\mathbf{w}$ is exactly $\mathbf{diag}(\mathbf{c})$. This completes the proof.

E. Proof of Proposition 24

Proof Suppose that D_{i^*} satisfies the maximal condition. Then, for any $\mathbf{w}$ satisfying that

$$\mathbf{w}^T\mathbf{x}_i\geq 0,\forall i\text{ with }q_i=1, \tag{194}$$

we shall have $\mathbf{w}^Tx_i<0$ for $q_i=0$. We note that the condition (194) is equivalent to $w\in-(K^+)^\circ$. Here $(K^+)^\circ=:\{\mathbf{w}|\mathbf{w}^T\mathbf{x}\leq 0,\forall\mathbf{x}\in K^+\}$ is the polar cone of the cone K^+. We note that $\mathbf{w}^T\mathbf{x}_i<0$ for $c_i=0$ implies that $\mathbf{w}\in\mathbf{int}\left((K^-)^\circ\right)$. Therefore, $(K^+)^\circ\subseteq\mathbf{int}\left((K^-)^\circ\right)$. This is equivalent to $-K^-\subseteq\mathbf{int}\left(K^+\right)$.

Suppose that $-K^-\subseteq\mathbf{int}\left(K^+\right)$. Then, we have $(K^+)^\circ\subseteq\mathbf{int}\left((K^-)^\circ\right)$. This implies that for any $\mathbf{w}\in\mathbb{R}^d$ satisfying (194), we shall have $\mathbf{w}^T\mathbf{x}_i<0$ for $c_i=0$. Therefore, $\mathbf{D}_{i^*}$ satisfies the maximal condition.

F. Proof of Proposition 25

Proof We can write

$$\mathbf{X}^{(1)}=\mathbf{1}\boldsymbol{\mu}^T+\sigma\mathbf{Z}, \tag{195}$$

where $\mathbf{Z}=\begin{bmatrix}\mathbf{z}_1^T\\\vdots\\\mathbf{z}_{n_1}^T\end{bmatrix}\in\mathbb{R}^{n_1\times d}$ and $\mathbf{1}\in\mathbb{R}^{n_1}$ is a vector of 1s. Each element of $\mathbf{Z}$ follows $\mathcal{N}(0,1)$. We note that the extreme eigenvalue vector of $(\mathbf{1}\boldsymbol{\mu}^T)^T\mathbf{1}\boldsymbol{\mu}^T=n_1^2\boldsymbol{\mu}\boldsymbol{\mu}^T$ is $\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}$. According to the tail bound of Gaussian random variable, with probability at least $1-\delta$, we have

$$\max_{i\in[n_1],j\in[d]}|z_{i,j}|\leq\sqrt{2\log(n_1d/\delta)}. \tag{196}$$

Denote $E = \{\max_{i\in[n_1],j\in[d]} |z_{i,j}| \leq \sqrt{2\log(n_1 d/\delta)}\}$. Conditioned on the event E, we have

$$\|(\mathbf{1}\boldsymbol{\mu}^T)^T\mathbf{Z} + \mathbf{Z}^T\mathbf{1}\boldsymbol{\mu}^T + \mathbf{Z}^T\mathbf{Z}\|_F \leq 2\sqrt{2\log(n_1 d/\delta)}\|\mathbf{1}\boldsymbol{\mu}^T\|_F + 2d\log(n_1 d/\delta) = 2n_1\sqrt{2\log(n_1 d/\delta)}\|\boldsymbol{\mu}\|_2 + 2d\log(n_1 d/\delta) =: c_0. \tag{197}$$

Let $\left(\mathbf{X}^{(1)}\right)^T\mathbf{X}^{(1)} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T$ be the eigenvalue decomposition. According the Weil's theorem, we note that

$$|n^2\|\mu\|_2^2 - \sigma_1| \leq \sigma\|(\mathbf{1}\mu^T)^T\mathbf{Z} + \mathbf{Z}^T\mathbf{1}\mu^T + \mathbf{Z}^T\mathbf{Z}\|_F \leq \sigma c_0. \tag{198}$$

The other eigenvalues σ_i of $\left(\mathbf{X}^{(1)}\right)^T\mathbf{X}^{(1)}$ satisfies that

$$|\sigma_i| \leq \sigma\|(\mathbf{1}\mu^T)^T\mathbf{Z} + \mathbf{Z}^T\mathbf{1}\mu^T + \mathbf{Z}^T\mathbf{Z}\|_F \leq \sigma c_0. \tag{199}$$

We note that

$$\begin{aligned}\left\|\left(\mathbf{X}^{(1)}\right)^T\mathbf{X}^{(1)}\boldsymbol{\mu}\right\|_2 \geq& n^2\|\boldsymbol{\mu}\|_2^3 - \sigma\|(\mathbf{1}\boldsymbol{\mu}^T)^T\mathbf{Z} + \mathbf{Z}^T\mathbf{1}\boldsymbol{\mu}^T + \mathbf{Z}^T\mathbf{Z}\|_2\|\boldsymbol{\mu}\|_2\\ \geq& n^2\|\boldsymbol{\mu}\|_2^3 - \sigma\|(\mathbf{1}\boldsymbol{\mu}^T)^T\mathbf{Z} + \mathbf{Z}^T\mathbf{1}\boldsymbol{\mu}^T + \mathbf{Z}^T\mathbf{Z}\|_F\|\boldsymbol{\mu}\|_2\\ =& (n^2\|\boldsymbol{\mu}\|_2^2 - \sigma c_0)\|\boldsymbol{\mu}\|_2.\end{aligned} \tag{200}$$

As $\left(\mathbf{X}^{(1)}\right)^T\mathbf{X}^{(1)}\boldsymbol{\mu} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\mu} = \sum_{i=1}^d \sigma_i\mathbf{v}_i(\mathbf{v}_i^T\boldsymbol{\mu})$, we have

$$\sigma_1|\mathbf{v}_1^T\boldsymbol{\mu}| \geq \left\|\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\mu} - \sum_{i=2}^d \sigma_i\mathbf{v}_i(\mathbf{v}_i^T\boldsymbol{\mu})\right\|_2 = \|\left(\mathbf{X}^{(1)}\right)^T\mathbf{X}^{(1)}\boldsymbol{\mu}\|_2 - \sum_{i=2}^d \sigma_i|\mathbf{w}_i^T\boldsymbol{\mu}| \geq (n^2\|\boldsymbol{\mu}\|_2^2 - (d-1)c_0\sigma)\|\boldsymbol{\mu}\|_2. \tag{201}$$

Note that $\mathbf{w} = \mathbf{v}_1$. This implies that

$$\left|\mathbf{w}^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right| \geq \frac{n^2\|\boldsymbol{\mu}\|_2^2 - (d-1)c_0\sigma}{n^2\|\boldsymbol{\mu}\|_2^2 + c_0\sigma}. \tag{202}$$

Note that

$$\left\|\mathbf{w} - \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right\|_2^2 = 2\left(1 - \left|\mathbf{w}^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right|\right) \leq \frac{2dc_0\sigma}{n^2\|\boldsymbol{\mu}\|_2^2 + c_0\sigma} \leq c_1\sigma, \tag{203}$$

where we let $c_1 = \frac{2dc_0}{n^2\|\boldsymbol{\mu}\|_2^2}$. This completes the proof.

G. Proof of Theorem 13

Proof Denote $\hat{\mathbf{w}}^* = \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|_2}$. According to Proposition 25, there exists a constant $c_1 > 0$ such that $\hat{\mathbf{w}}^*$ satisfies that

$$\left\|\hat{\mathbf{w}}^* - \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right\|_2^2 \leq c_1\sigma, \tag{204}$$

with probability at least $1 - \delta/2$. For $\sigma \leq \frac{\|\boldsymbol{\mu}\|_2}{2(d + 8\log(4n/\delta))}$, the events

$$\sigma\|\mathbf{z}_i\|_2 \leq \|\boldsymbol{\mu}\|_2/2, \tag{205}$$

holds with probability at least $1 - \delta/(2n)$ respectively for all $i \in [n]$. Denote $E = \{\sigma\|\mathbf{z}_i\|_2 \leq \|\boldsymbol{\mu}\|_2/2, \forall i \in [n]\}$. Then, $P(E) \geq 1 - \delta/4$. For indices $i \in [n]$ such that $q_i = 1$, conditioned on E, we have $\sigma\left|\mathbf{z}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right| \leq \|\boldsymbol{\mu}\|_2/2$ and

$$\mathbf{x}_i^T\hat{\mathbf{w}}^* \geq \mathbf{x}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} - \|\mathbf{x}_i\|\left\|\hat{\mathbf{w}}^* - \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right\|_2 \geq \mathbf{x}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} - \sqrt{c_1\sigma}\|\mathbf{x}_i\|_2 \geq \|\boldsymbol{\mu}\|_2 + \sigma\mathbf{z}_n^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} - 2\sqrt{c_1\sigma}\|\boldsymbol{\mu}\|_2 > 0 \tag{206}$$

Here we utilize that $\sigma \leq \frac{1}{32c_1}$. Similarly, for indices $i \in [n]$ such that $q_i = -1$, conditioned on E, we have $\sigma\left|\mathbf{z}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right| \leq \|\boldsymbol{\mu}\|_2/2$ and

$$\mathbf{x}_i^T\hat{\mathbf{w}}^* \leq \mathbf{x}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \|\mathbf{x}_i\|\left\|\hat{\mathbf{w}}^* - \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\right\|_2 \leq \mathbf{x}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + \sqrt{c_1\sigma}\|\mathbf{x}_i\|_2\mathbf{x}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} \leq -\|\boldsymbol{\mu}\|_2 - \sigma\mathbf{z}_i^T\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} + 2\sqrt{c_1\sigma}\|\boldsymbol{\mu}\|_2 < 0. \tag{207}$$

This implies that $(2\mathbf{D}_{i^*} - I)\mathbf{X}\hat{\mathbf{w}}^* > 0$. Overall, for sufficiently small $a > 0$, the event $(2\mathbf{D}_{i^*} - I)\mathbf{X}\hat{\mathbf{w}}^* > 0$ holds with probability at least

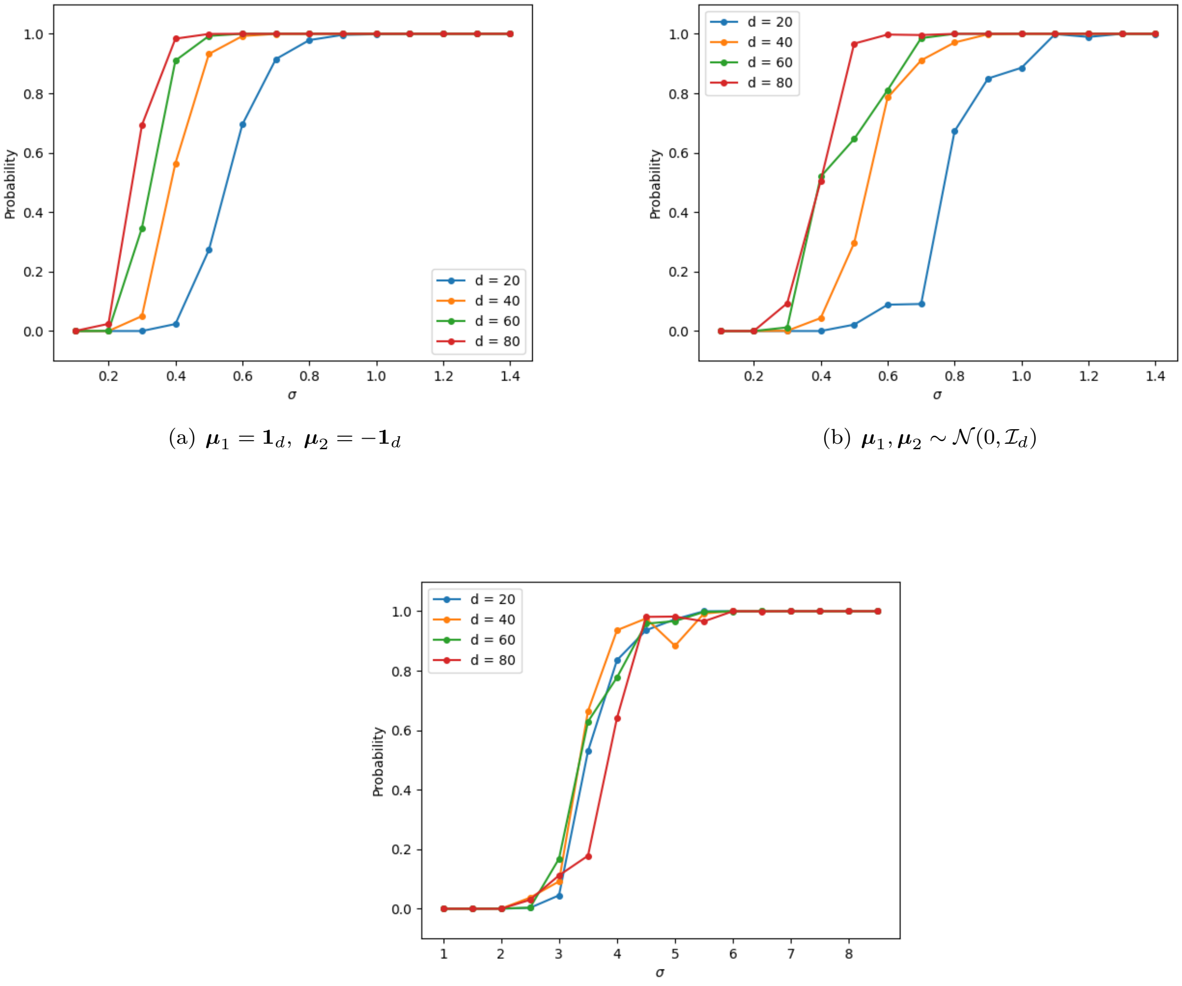
$$1 - \delta/2 - \delta/2 = 1 - \delta. \tag{208}$$

H. Numerical verification

In this subsection, we numerically verify Proposition 23. We take $n = 100, n_1 = n_2 = 50$ and $d = 20, 40, 60, 80$, and test for three types of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$.

- $\boldsymbol{\mu}_1 = \mathbf{1}_d,\ \boldsymbol{\mu}_2 = -\mathbf{1}_d$.
- $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \sim \mathcal{N}(0, \mathbf{I}_d)$.
- $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \sim \mathcal{U}(\mathbb{S}^{d-1})$.

For each d, we compute the probability that the diagonal arrangement pattern induced by the vector $\mathbf{w} = \frac{\boldsymbol{\mu}_1}{\|\boldsymbol{\mu}_1\|_2} - \frac{\boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_2\|_2}$ equals to $\mathbf{diag}(\mathbf{w})$ for σ in a certain range. For each σ, we establish 5000 independent trials to compute the probability. The S-shaped curves shown in Figure 20 correspond with formulation of the lower bound given in Proposition 23.



(a) $\boldsymbol{\mu}_1 = \mathbf{1}_d,\ \boldsymbol{\mu}_2 = -\mathbf{1}_d$

(b) $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \sim \mathcal{N}(0, \mathcal{I}_d)$

(c) $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \sim \mathcal{U}(\mathbb{S}^{d-1})$

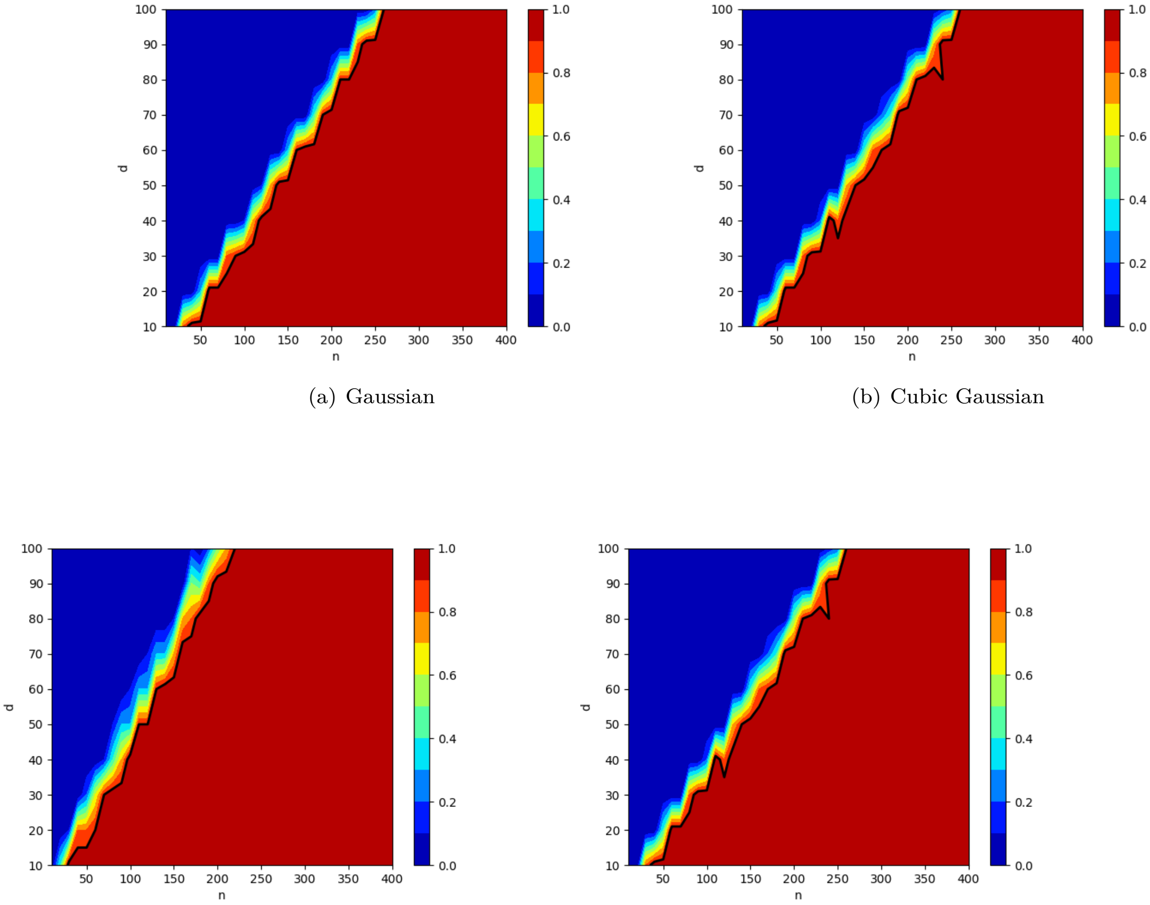
Fig. 20: The probability that the statement in Proposition 23 holds over 5000 independent trials.

Appendix M
Additional numerical experiments

In this section, we present additional numerical results mentioned in Section VII. These results serve as a complement to our main numerical results.

A. ReLU networks with skip connection

In Figure 6, we show phase transition graph for the probability of successful recovery of the planted linear neuron by solving the group ℓ_1-minimization problem (12) when the planted neuron $\mathbf{w}^*$ is randomly generated from $\mathcal{N}(0, \mathbf{I}_d)$. In Figure 21 below, we find similar results when the planted neuron $\mathbf{w}^*$ is the smallest right singular vector of $\mathbf{X}$.



(a) Gaussian

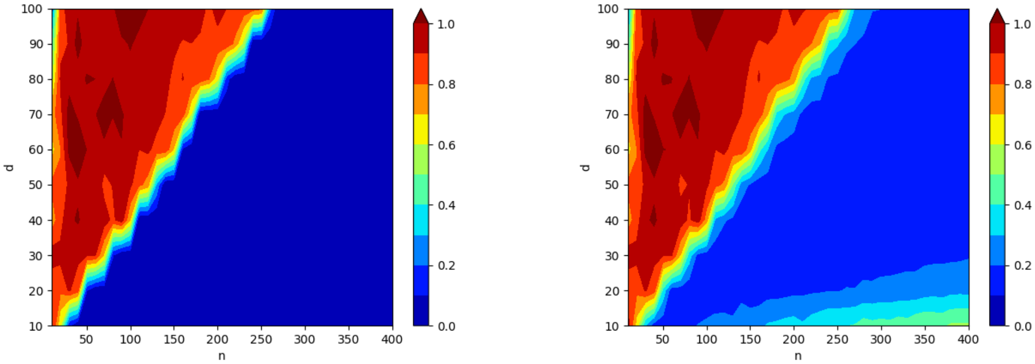
(b) Cubic Gaussian

(c) Haar

(d) Cubic Gaussian + whitened

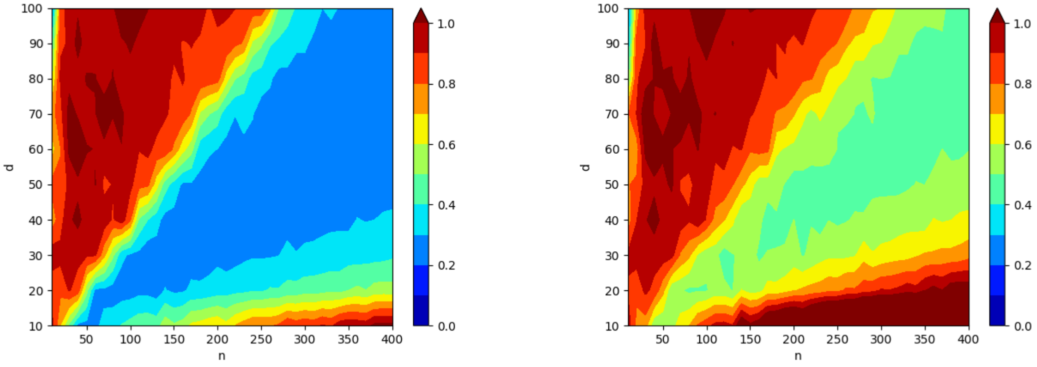
Fig. 21: The probability of successful recovery of the planted linear neuron by solving the group ℓ_1-minimization problem (12) over 5 independent trials. The black lines represent the boundaries of successful recovery with probability 1. Here the planted neuron $\mathbf{w}^*$ is the smallest right singular vector of $\mathbf{X}$.

In Figure 7, we show phase transition graph for absolute distance by solving the group ℓ_1-minimization problem (12) derived from training ReLU networks with skip connection. In Figure 22 below, we find similar pattern of the phase transition.



(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$



(c) $\sigma = 0.1$

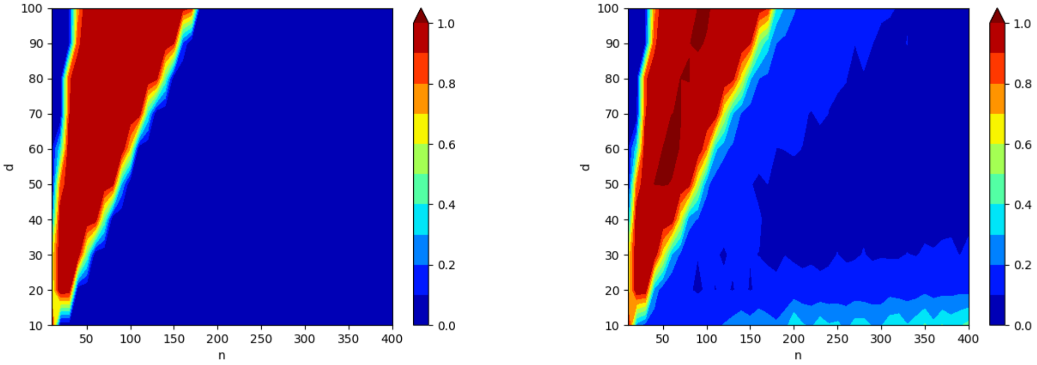
(d) $\sigma = 0.2$

Fig. 22: Averaged test distance by solving the group ℓ_1-minimization problem (12) derived from training ReLU networks with skip connection over 5 independent trials.

As a complement to Figure 8, we study the generalization property of ReLU networks with skip connections using convex training methods. We solve the following regularized training problem with small regularization parameter $\beta = 10^{-6}$ as an approximation of the minimal-norm problem (3).

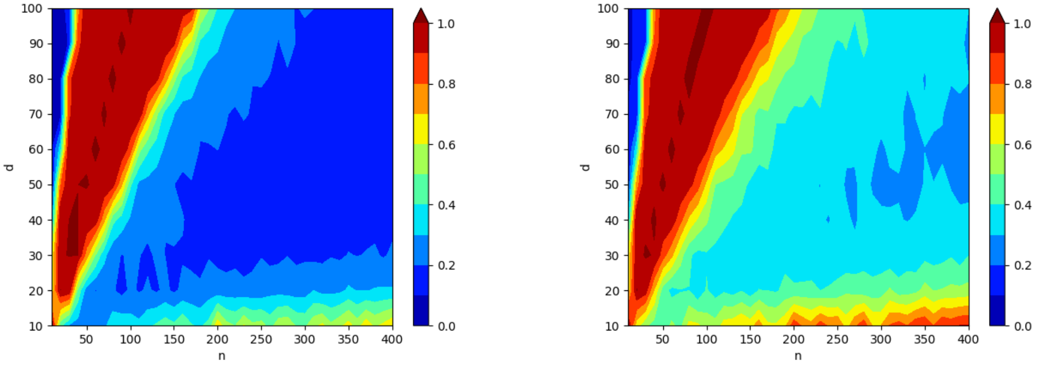
$$\begin{aligned}\min_{\mathbf{w}_0, \{\mathbf{w}_j, \mathbf{w}'_j\}_{j=1}^{p}} \quad & \left\| \mathbf{X}\mathbf{w}_0 + \sum_{j=1}^{p} \mathbf{D}_j \mathbf{X} \left(\mathbf{w}_j - \mathbf{w}'_j \right) - \mathbf{y} \right\|_2^2 + \beta \left(\|\mathbf{w}_0\|_2 + \sum_{j=1}^{p} \left(\|\mathbf{w}_j\|_2 + \|\mathbf{w}'_j\|_2 \right) \right), \\ \text{s.t.} \quad & (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{w}_j \geq 0, (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{w}'_j \geq 0, j \in [p].\end{aligned} \tag{209}$$

Then we compute the corresponding absolute distance and test distance. The results shown in Figures 23 and 24 have the same sharp $n = 2d$ boundary of the red region. This validates Proposition 1. Namely, the recovery of the group ℓ_1-minimization problem (12) implies the recovery of the convex program (3).



(a) $\sigma = 0$(noiseless)

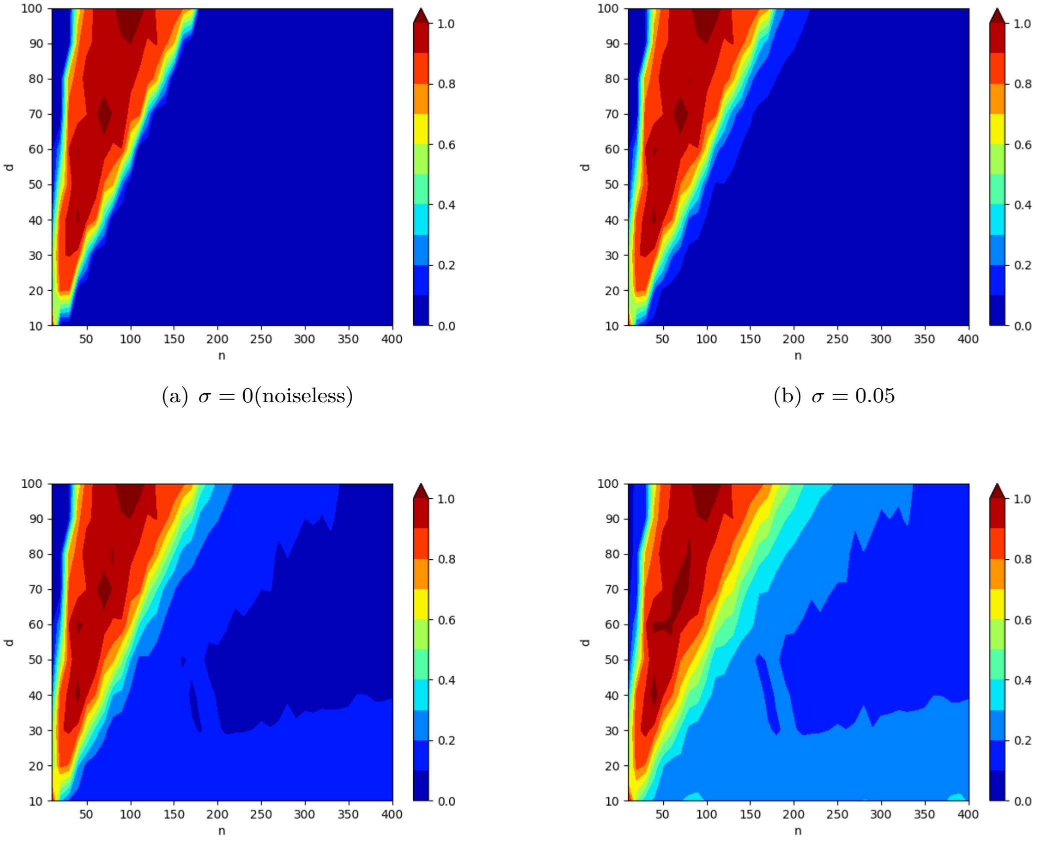
(b) $\sigma = 0.05$



(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 23: Averaged absolute distance to the planted linear neuron by solving the convex regularized training problem (209) for ReLU networks with skip connection over 5 independent trials.



(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 24: Averaged test distance by solving the convex regularized training problem (209) for ReLU networks with skip connection over 5 independent trials.

B. ReLU networks with normalization layer

We first present the phase transition graph for successful recovery of the planted normalized ReLU neuron by solving the group ℓ_1-minimization (14). Analogous to Section VII-A, we compute the recovery rate for d ranging from 10 to

100 and n ranging from 10 to 400 and establish 5 independent trials for each pair of (n, d). Here each element $x_{i,j}$ of the dataset $\mathbf{X} \in \mathbb{R}^{n\times d}$ is i.i.d. random variable following $\mathcal{N}(0, 1/n)$. The planted neuron $\mathbf{w}^*$ is randomly generated from $\mathcal{N}(0, \mathbf{I}_d)$.

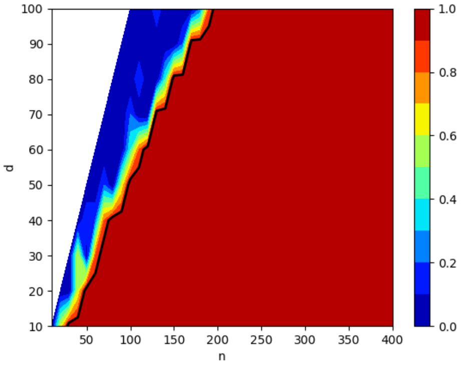
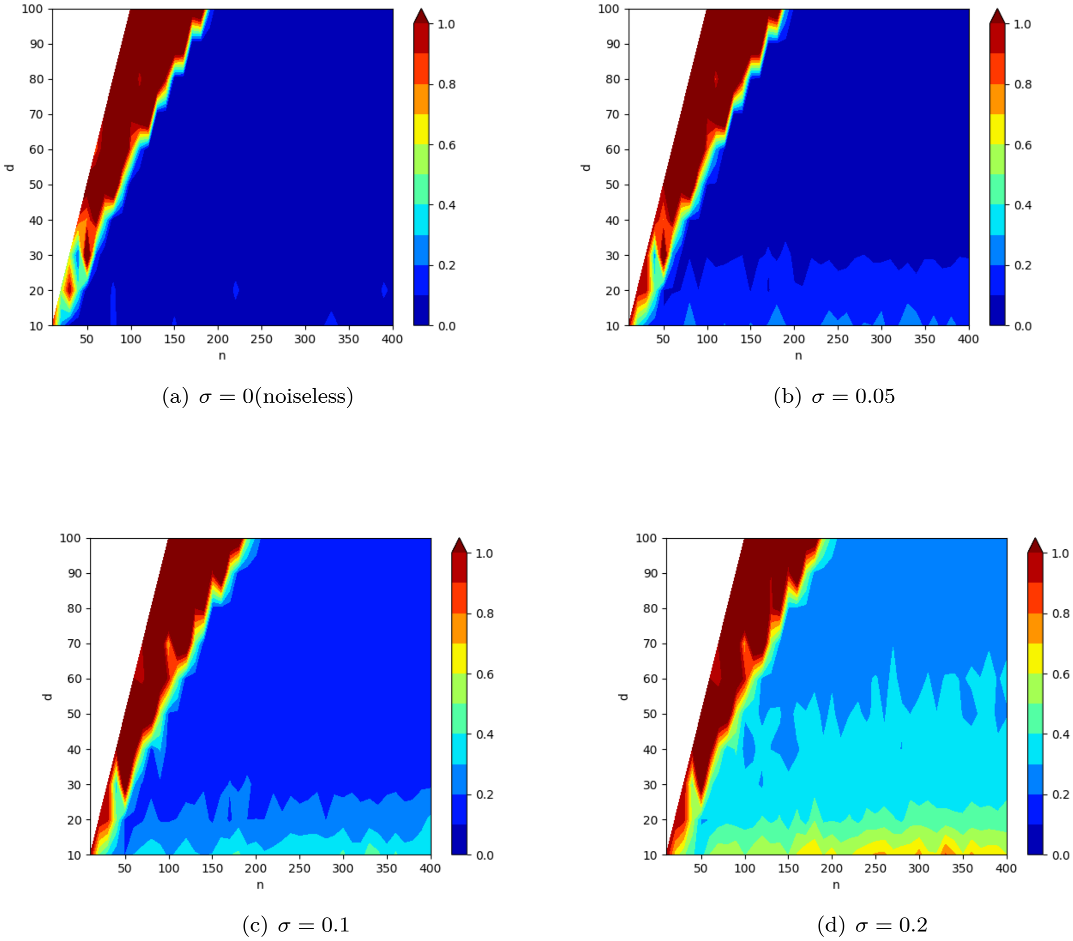


Fig. 25: The probability of successful recovery of the planted normalized ReLU neuron by solving the group ℓ_1-minimization problem (14) over 5 independent trials. The black line represents the boundaries of successful recovery with probability 1. Here the planted neuron $\mathbf{w}^*$ is randomly generated from $\mathcal{N}(0, \mathbf{I}_d)$.

The second part is phase transition under noisy observation, i.e., $\mathbf{y} = \frac{(\mathbf{X}\mathbf{w}^*)_+}{\|(\mathbf{X}\mathbf{w}^*)_+\|_2} + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \sigma^2/n)$. In Figure 26, we focus on absolute distance, which is defined as the ℓ_2 distance between the optimal solution $\mathbf{w}_{i^*}$ and $\tilde{\mathbf{w}}^*$. A sharp $n = 2d$ boundary can be observed from the phase transition graphs, which corresponds with Theorem 11.



(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

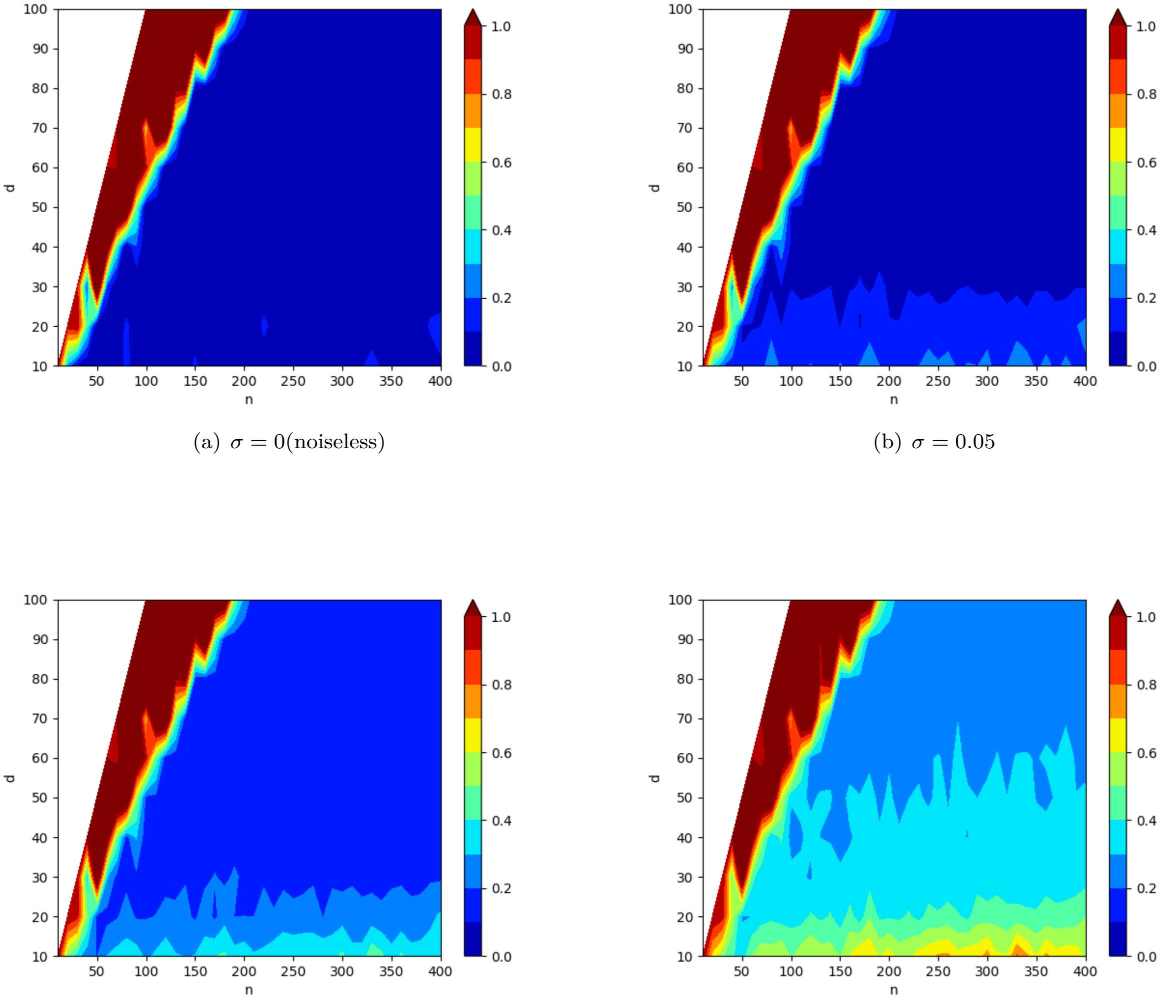
(d) $\sigma = 0.2$

Fig. 26: Averaged absolute distance to the planted normalized ReLU neuron by solving the group ℓ_1-minimization problem (14) from training ReLU networks with normalization layer over 5 independent trials.

In the third part, we study the generalization property of ReLU networks with normalization layer using both the convex and non-convex training procedures. For the non-convex training procedure, we minimize the regularized training objective with small regularization parameter $\beta = 10^{-6}$ to approximate the minimal-norm problem (13).

$$\begin{aligned} \min_{\{\mathbf{w}_j,\mathbf{w}'_j\}_{j=1}^p} \quad & \left\| \sum_{j=1}^p \mathbf{U}_j \left(\mathbf{w}_j - \mathbf{w}'_j\right) - \mathbf{y} \right\|_2^2 + \beta \sum_{j=1}^p \left(\left\|\mathbf{w}_j\right\|_2 + \left\|\mathbf{w}'_j\right\|_2 \right) \\ \text{s.t.} \quad & (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{V}_j^T\mathbf{\Sigma}_j^{-1}\mathbf{w}_j \geq 0, (2\mathbf{D}_j - \mathbf{I}_n)\mathbf{X}\mathbf{V}_j^T\mathbf{\Sigma}_j^{-1}\mathbf{w}'_j \geq 0, j \in [p], \end{aligned} \tag{210}$$

Then we compute the corresponding absolute distance. The phase transition graphs are shown in Figure 27. We can also observe a sharp $n = 2d$ transition similar to the group ℓ_1-minimization (14).



(a) $\sigma = 0$(noiseless)

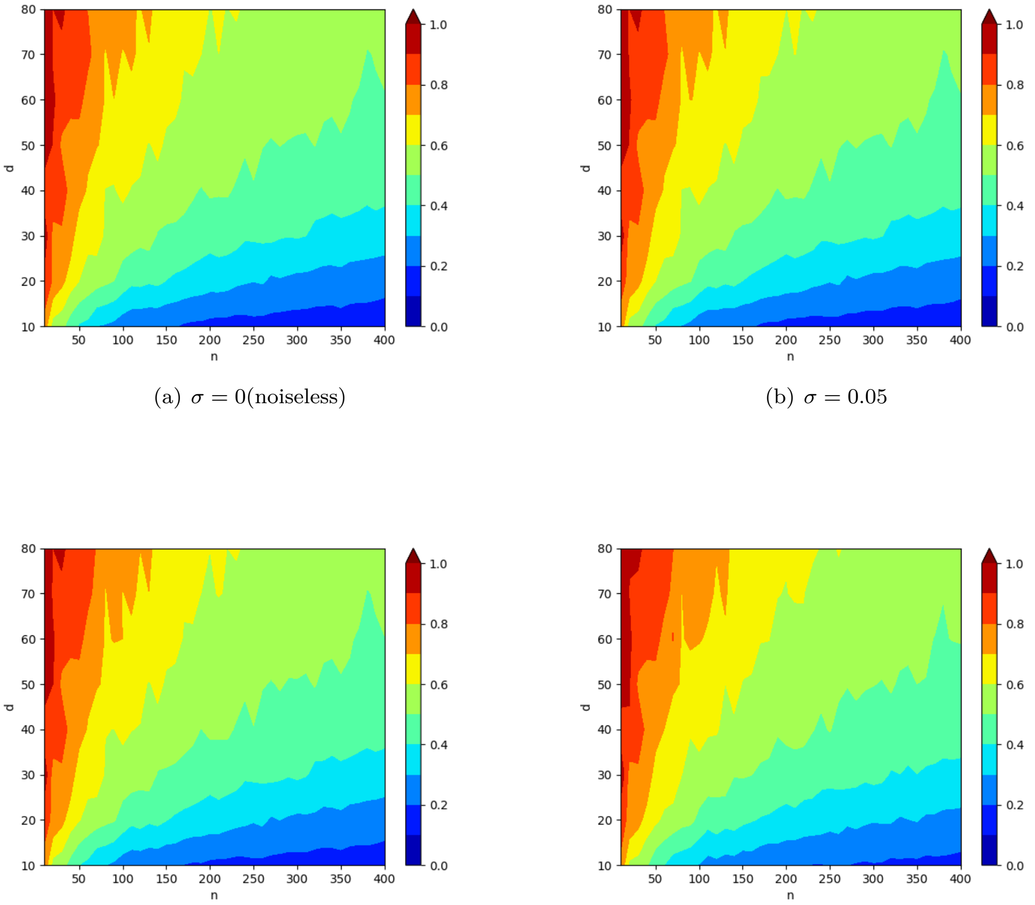
(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 27: Averaged absolute distance by training ReLU networks with normalization layer on the regularized convex problem (6) over 5 independent trials.

For nonconvex training method, we solve the regularized non-convex training problem (6) with the same setting as Section VII-A. We set the number of neurons to be $m = n + 1$ and train the ReLU neural network with normalization layer for 400 epochs. We use the AdamW optimizer with weight decay $\beta = 10^{-6}$. We note that the nonconvex training may still reach local minimizers. Thus, the absolute distance do not show clear phase transition as the convex training. However, the transitions of test error generally follow the patterns of the group ℓ_1-minimization problem. In Figure 28, we show that the test error increases as n/d increases, and the rate of increase becomes sharp around $n = 2d$ (the boundary of orange region).



(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

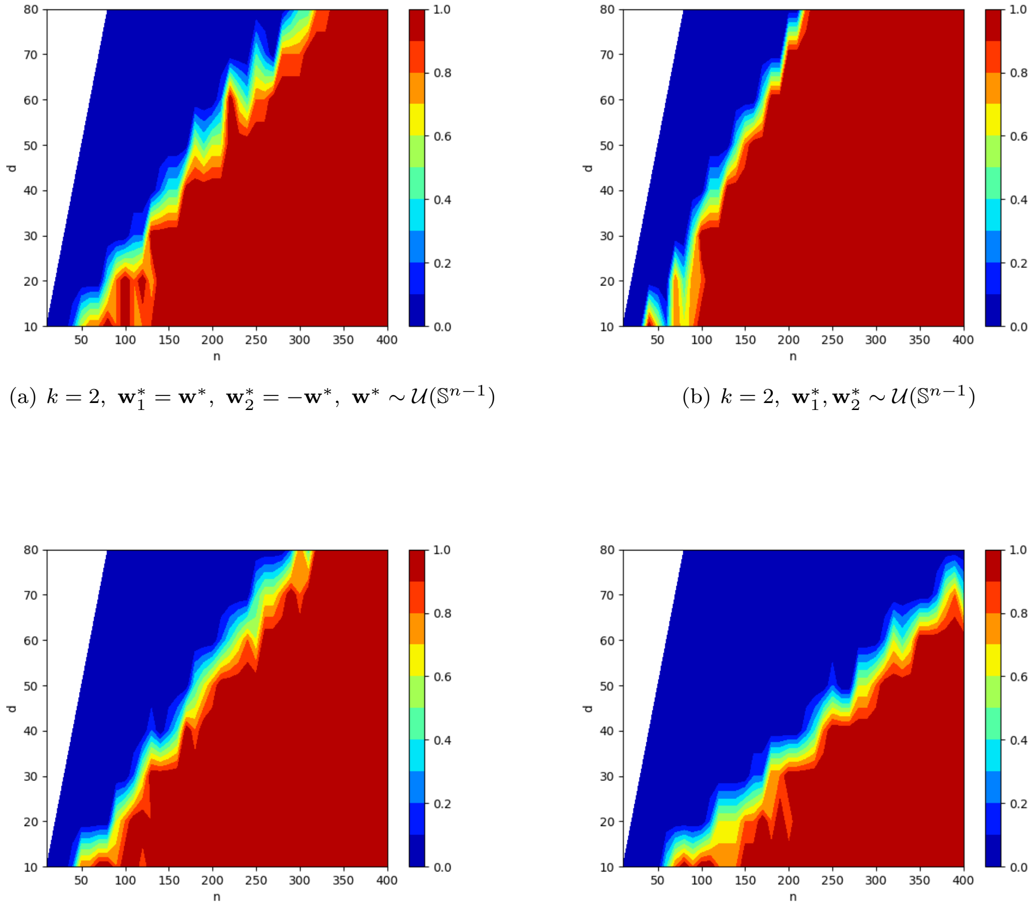
(d) $\sigma = 0.2$

Fig. 28: Averaged test error by training ReLU networks with normalization layer on the regularized non-convex problem (6) over 10 independent trials.

C. *Multi-neuron recovery*

In this subsection, we present further numerical results on the recovery of ReLU networks with normalization layer when the label vector is the combination of several normalized ReLU neurons. Readers can refer to the details of the experiments in Section VII-B and Appendix M-B.

We first show phase transition graphs for successful recovery of the planted normalized ReLU neurons. We compute the recovery rate for d ranging from 10 to 80 and n ranging from 10 to 400 and establish 5 independent trials for each pair of (n, d).



(a) $k = 2,\ \mathbf{w}_1^* = \mathbf{w}^*,\ \mathbf{w}_2^* = -\mathbf{w}^*,\ \mathbf{w}^* \sim \mathcal{U}(\mathbb{S}^{n-1})$

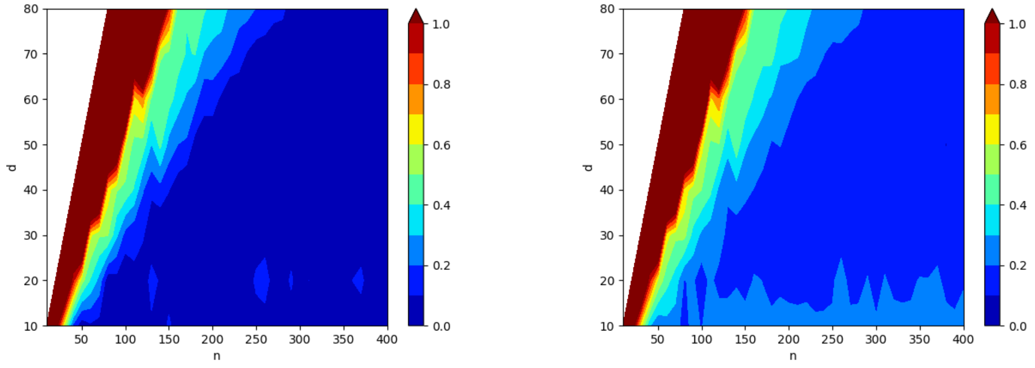
(b) $k = 2,\ \mathbf{w}_1^*, \mathbf{w}_2^* \sim \mathcal{U}(\mathbb{S}^{n-1})$

(c) $k = 2,\ \mathbf{w}_1^* = \mathbf{e}_1,\ \mathbf{w}_2^* = \mathbf{e}_2,$

(d) $k = 3,\ \mathbf{w}_i^* = \mathbf{e}_i (i = 1, 2, 3)$

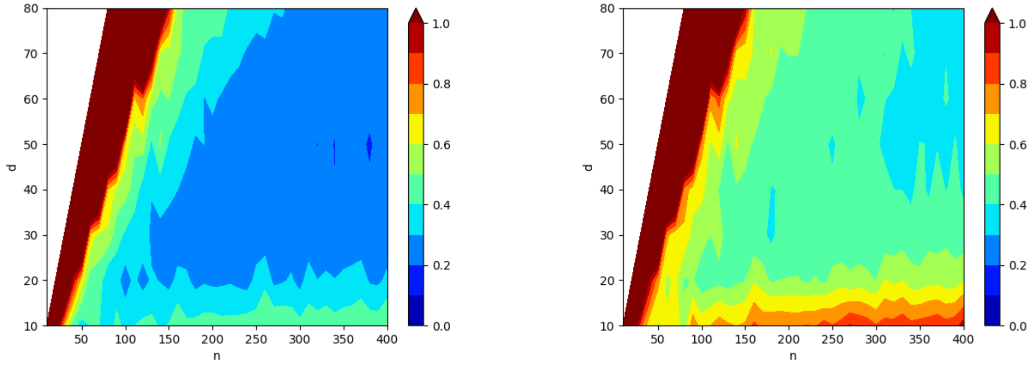
Fig. 29: The probability of successful recovery of the planted normalized ReLU neurons by solving the group ℓ_1-minimization problem (14) over 5 independent trials. The label vector $\mathbf{y}$ is the combination of several normalized ReLU neurons.

The second part is noisy observation model. In Figure 9, we show the phase transition graph when the planted neurons satisfy $\mathbf{w}_1^* = \mathbf{w}^*,\ \mathbf{w}_2^* = -\mathbf{w}^*,\ \mathbf{w}^* \sim \mathcal{U}(\mathbb{S}^{n-1})$. In Figure 30, 31 and 32 below, we show results when $\mathbf{w}_1^*, \mathbf{w}_2^* \sim \mathcal{U}(\mathbb{S}^{n-1})$, $\mathbf{w}_1^* = \mathbf{e}_1,\ \mathbf{w}_2^* = \mathbf{e}_2$ and $\mathbf{w}_i^* = \mathbf{e}_i (i = 1, 2, 3)$, respectively.



(a) $\sigma = 0$(noiseless)

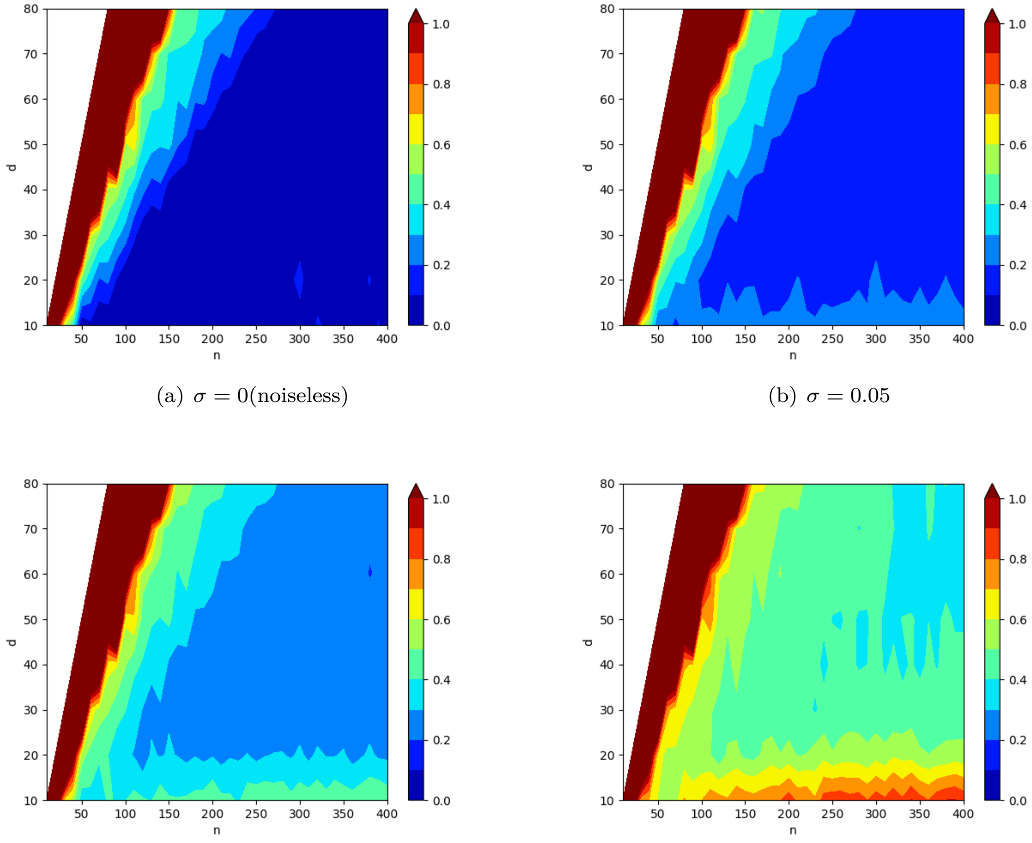
(b) $\sigma = 0.05$



(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 30: Averaged absolute distance to the planted normalized ReLU neurons by solving the group ℓ_1-minimization problem (14) from training ReLU networks with normalization layer over 5 independent trials. Here we set $k = 2$ planted neurons which satisfy $\mathbf{w}_1^*, \mathbf{w}_2^* \sim \mathcal{U}(\mathbb{S}^{n-1})$.



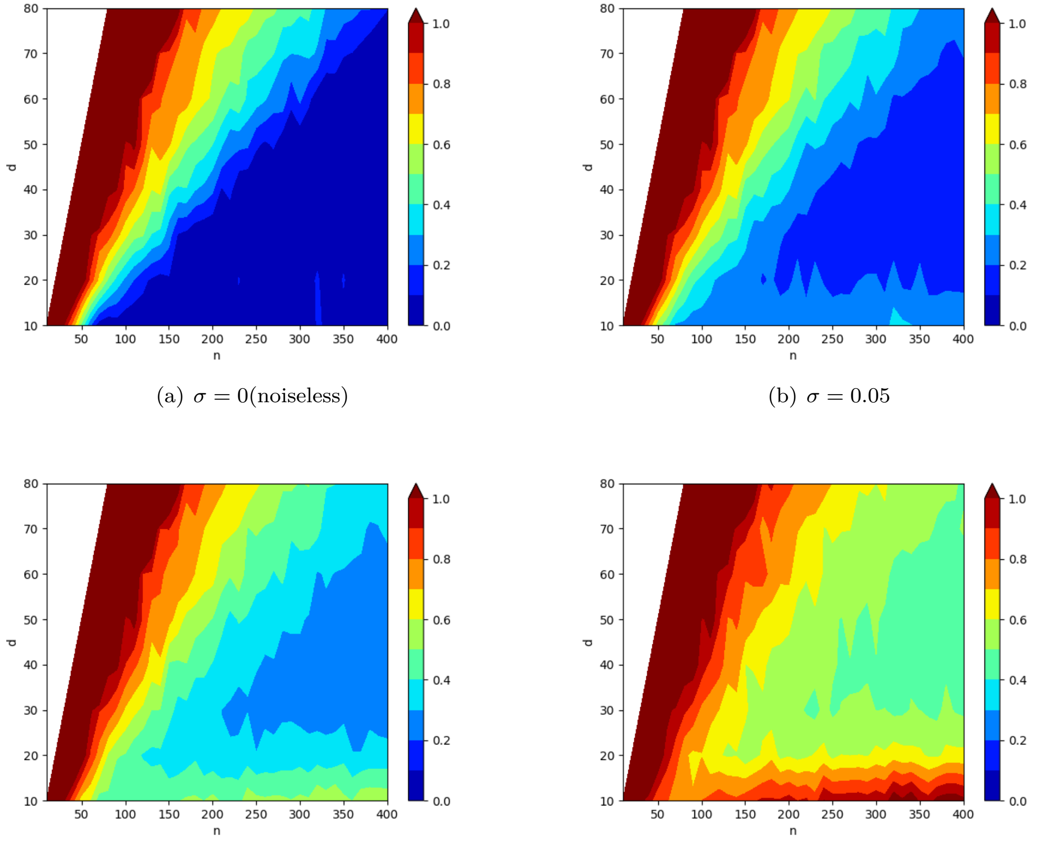
(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 31: Averaged absolute distance to the planted normalized ReLU neurons by solving the group ℓ_1-minimization problem (14) from training ReLU networks with normalization layer over 5 independent trials. Here we set $k = 2$ planted neurons which satisfy $\mathbf{w}_1^* = \mathbf{e}_1,\ \mathbf{w}_2^* = \mathbf{e}_2$.



(a) $\sigma = 0$(noiseless)

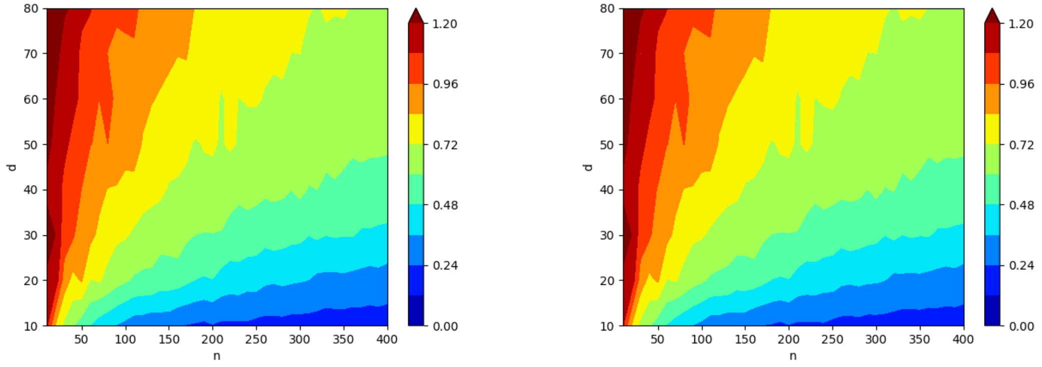
(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

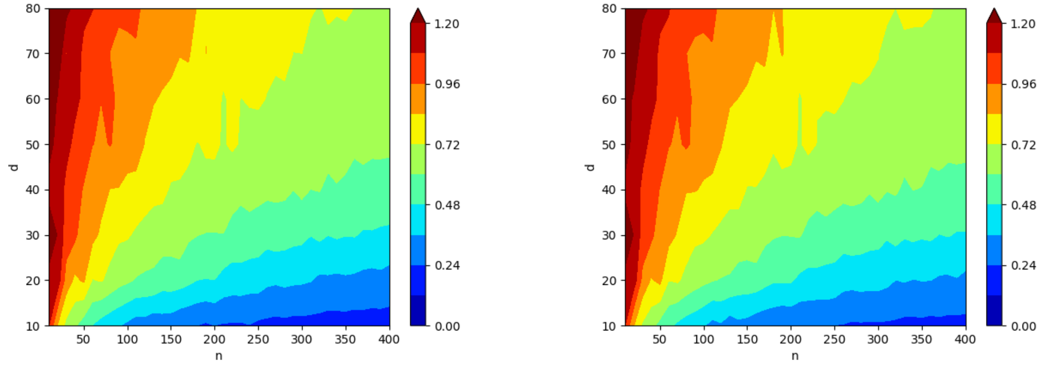
Fig. 32: Averaged absolute distance to the planted normalized ReLU neurons by solving the group ℓ_1-minimization problem (14) from training ReLU networks with normalization layer over 10 independent trials. Here we set $k = 3$ planted neurons which satisfy $\mathbf{w}_i^* = \mathbf{e}_i (i = 1, 2, 3)$.

In the third part, we directly study the generalization property of ReLU networks with normalization layer using non-convex training methods. We note that the transitions of test error generally follow the patterns of the group ℓ_1-minimization problem, analogous to our observation in Appendix M-B.



(a) $\sigma = 0$(noiseless)

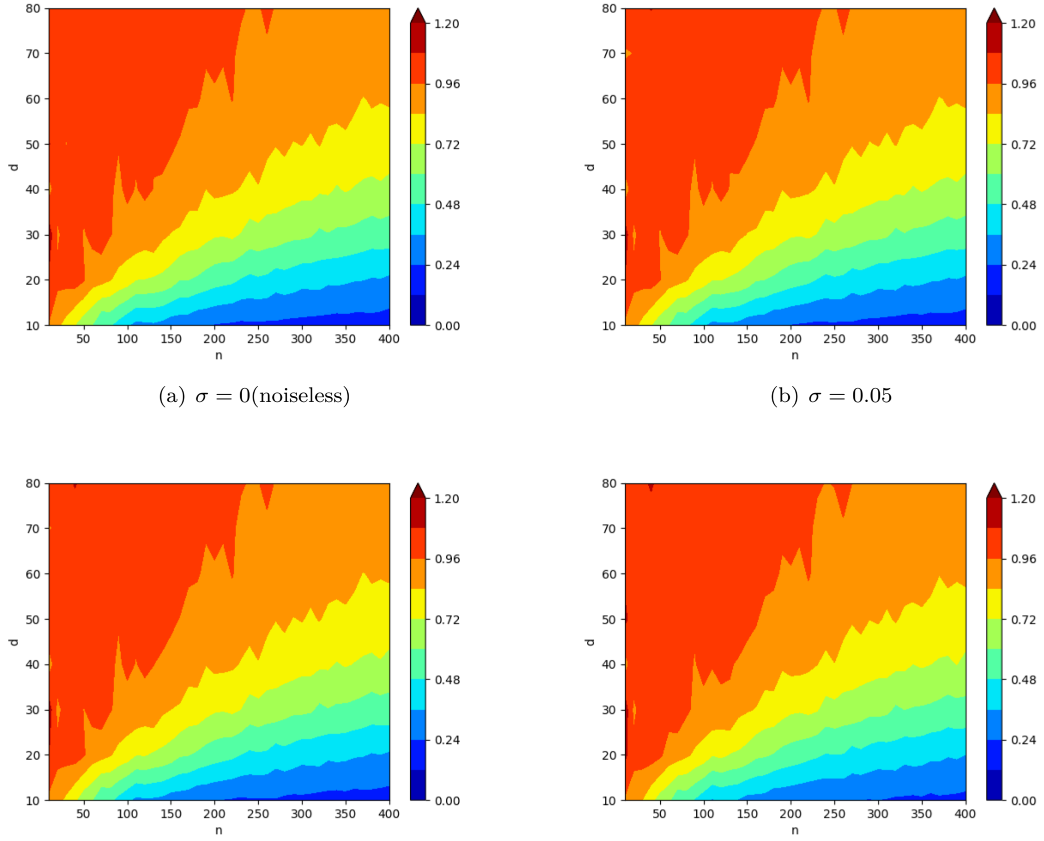
(b) $\sigma = 0.05$



(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 33: Averaged test error by training ReLU networks with normalization layer on the regularized non-convex problem (6) over 10 independent trials. Here we set $k = 2$ planted neurons which satisfy $\mathbf{w}_1^* = \mathbf{w}^*,\ \mathbf{w}_2^* = -\mathbf{w}^*,\ \mathbf{w}^* \sim \mathcal{U}(\mathbb{S}^{n-1})$.



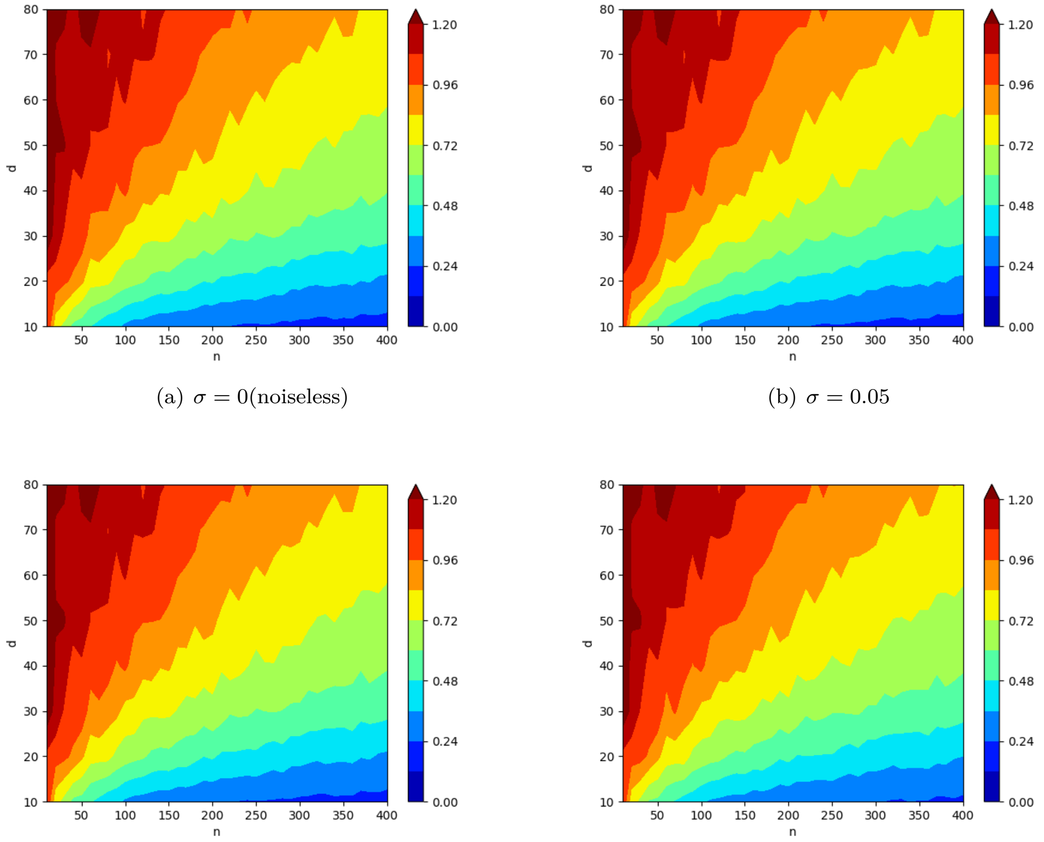
(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 34: Averaged test error by training ReLU networks with normalization layer on the regularized non-convex problem (6) over 10 independent trials. Here we set $k = 2$ planted neurons which satisfy $\mathbf{w}_1^*, \mathbf{w}_2^* \sim \mathcal{U}(\mathbb{S}^{n-1})$.



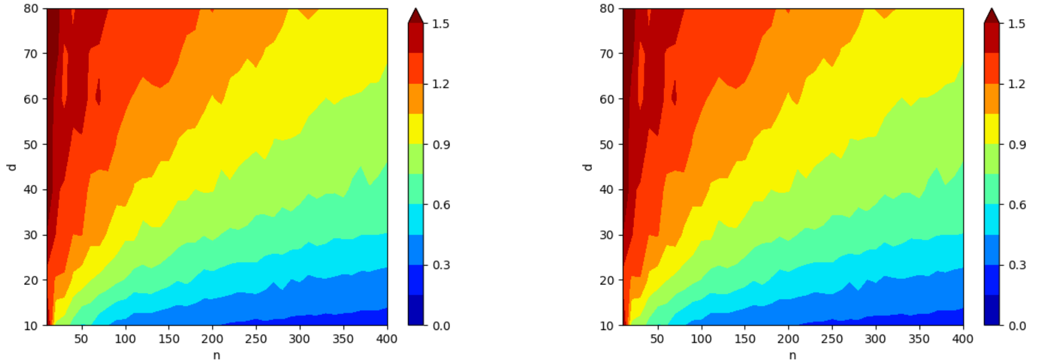
(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$

(c) $\sigma = 0.1$

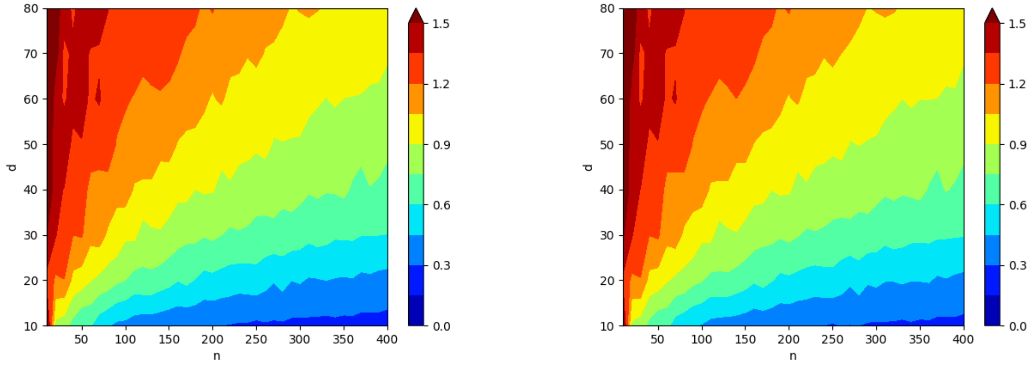
(d) $\sigma = 0.2$

Fig. 35: Averaged test error by training ReLU networks with normalization layer on the regularized non-convex problem (6) over 10 independent trials. Here we set $k = 2$ planted neurons which satisfy $\mathbf{w}_1^* = \mathbf{e}_1, \ \mathbf{w}_2^* = \mathbf{e}_2$.



(a) $\sigma = 0$(noiseless)

(b) $\sigma = 0.05$



(c) $\sigma = 0.1$

(d) $\sigma = 0.2$

Fig. 36: Averaged test error by training ReLU networks with normalization layer on the regularized non-convex problem (6) over 10 independent trials. Here we set $k = 3$ planted neurons which satisfy $\mathbf{w}_i^* = \mathbf{e}_i (i = 1, 2, 3)$.

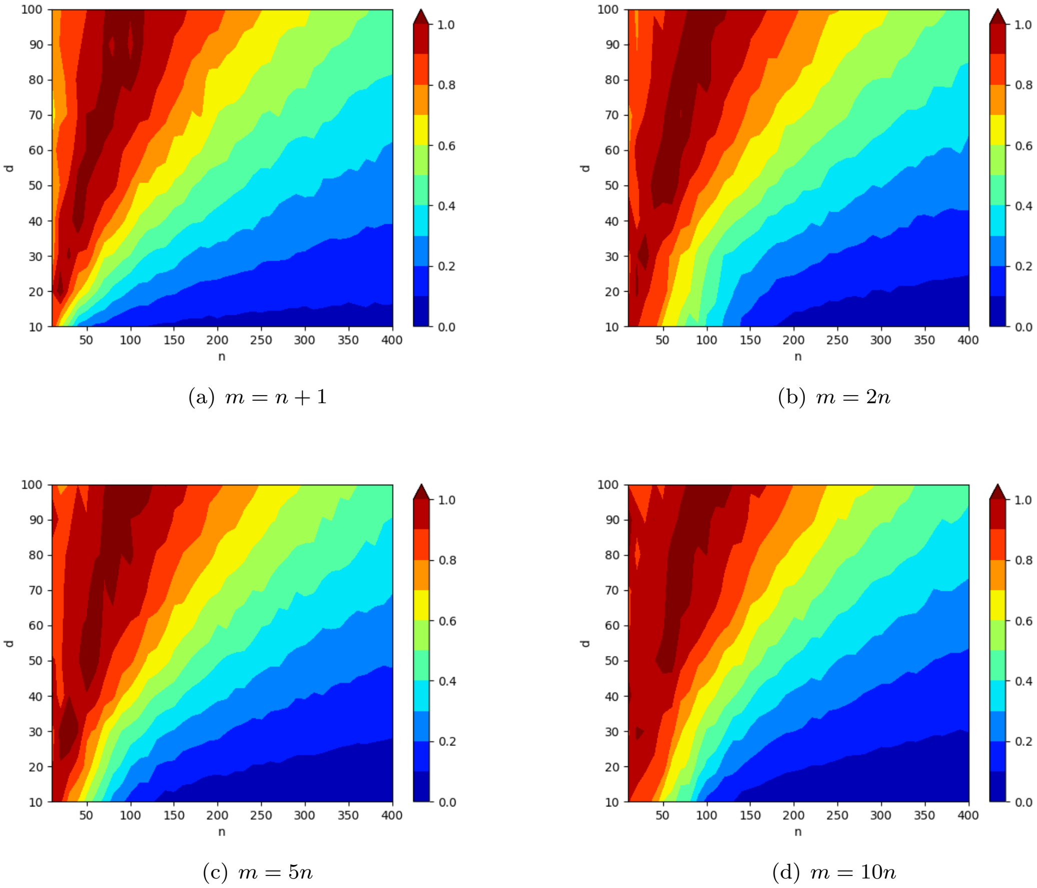
D. Ablation study with respect to the number of neurons in NN

In this subsection, we solve the regularized non-convex training problem (6) with more neurons in the NN. According to Theorem 1 in [2], for all $m \geq m^*$, the convex program (6) and the non-convex problem (8) have identical optimal values. Furthermore, it is indicated in the paper that $m^* \leq n + 1$. Therefore, this experiment serves as a complement of Figure 8 and implies that setting $m = n + 1$ neurons in the NN is sufficient to support our theorem.

We run additional training processes with the following settings. The training data is noiseless. We apply the AdamW optimizer with weight decay $\beta = 10^{-6}$. Then we compute the averaged test error for d ranging from 10 to 100 and n ranging from 10 to 400 and establish 10 independent trials for each pair of (n, d).

Number of neurons m	Number of training epochs	Learing rate
$m = 2n$	600	5e-5
$m = 5n$	1000	2e-5
$m = 10n$	2000	5e-6

Numerical results show that the transitions of test error generally follow the patterns of the group ℓ_1-minimization problem, and increasing the number of neurons in the NN does not affect these patterns.



(a) $m = n + 1$

(b) $m = 2n$

(c) $m = 5n$

(d) $m = 10n$

Fig. 37: Averaged test error by training ReLU networks with skip connection and with different number of neurons on the regularized non-convex problem (6) over 10 independent trials.

Acknowledgment

This work was partially supported by the National Science Foundation (NSF) under grants ECCS- 2037304, DMS-2134248, NSF CAREER award CCF-2236829, the U.S. Army Research Office Early Career Award W911NF-21-1-0242, Stanford Precourt Institute, and the ACCESS – AI Chip Center for Emerging Smart Systems, sponsored by InnoHK funding, Hong Kong SAR.

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[2] M. Pilanci and T. Ergen, "Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks," *International Conference on Machine Learning (ICML)*, 2020.

[3] T. Ergen and M. Pilanci, "Convex geometry and duality of over-parameterized neural networks," *Journal of Machine Learning Research*, vol. 22, no. 212, pp. 1–63, 2021.

[4] ——, "Global optimality beyond two layers: Training deep relu networks via convex programs," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2993–3003.

[5] Y. Wang and M. Pilanci, "The convex geometry of backpropagation: Neural network gradient flows converge to extreme points of the dual convex program," *International Conference on Learning Representations (ICLR)*, 2022.

[6] Y. Wang, J. Lacotte, and M. Pilanci, "The hidden convex optimization landscape of two-layer relu neural networks: an exact characterization of the optimal solutions," *International Conference on Learning Representations (ICLR)*, 2022.

[7] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[8] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[9] E. J. Candes, “The restricted isometry property and its implications for compressed sensing,” *Comptes rendus mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[10] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International Conference on Machine Learning.* PMLR, 2015, pp. 2256–2265.

[11] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[12] T. Ergen, A. Sahiner, B. Ozturkler, J. Pauly, M. Mardani, and M. Pilanci, “Demystifying batch normalization in relu networks: Equivalent convex optimization models and implicit regularization,” *International Conference on Learning Representations (ICLR)*, 2022.

[13] T. Ergen and M. Pilanci, “Implicit convex regularizers of cnn architectures: Convex optimization of two- and three-layer networks in polynomial time,” *International Conference on Learning Representations (ICLR)*, 2021.

[14] B. Bartan and M. Pilanci, “Neural spectrahedra and semidefinite lifts: Global convex optimization of polynomial activation neural networks in fully polynomial-time,” *arXiv preprint arXiv:2101.02429*, 2021.

[15] A. Sahiner, T. Ergen, B. Ozturkler, J. Pauly, M. Mardani, and M. Pilanci, “Unraveling attention via convex duality: Analysis and interpretations of vision transformers,” *International Conference on Machine Learning*, 2022.

[16] A. Sahiner, T. Ergen, B. Ozturkler, B. Bartan, J. Pauly, M. Mardani, and M. Pilanci, “Hidden convexity of wasserstein gans: Interpretable generative models with closed-form solutions,” *International Conference on Learning Representations*, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[18] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” 2018. [Online]. Available: https://arxiv.org/abs/1806.07572

[19] S. S. Du, X. Zhai, B. Poczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” 2018. [Online]. Available: https://arxiv.org/abs/1810.02054

[20] Z. Allen-Zhu, Y. Li, and Y. Liang, “Learning and generalization in overparameterized neural networks, going beyond two layers,” *arXiv preprint arXiv:1811.04918*, 2018.

[21] C. Liu, L. Zhu, and M. Belkin, “On the linearity of large non-linear models: when and why the tangent kernel is constant,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 954–15 964, 2020.

[22] D. Zou, Y. Cao, D. Zhou, and Q. Gu, “Stochastic gradient descent optimizes over-parameterized deep relu networks,” *arXiv preprint arXiv:1811.08888*, 2018.

[23] Z. Allen-Zhu, Y. Li, and Z. Song, “A convergence theory for deep learning via over-parameterization,” in *International Conference on Machine Learning.* PMLR, 2019, pp. 242–252.

[24] Z. Allen-Zhu, Y. Li, and Y. Liang, “Learning and generalization in overparameterized neural networks, going beyond two layers,” *Advances in neural information processing systems*, vol. 32, 2019.

[25] L. Chizat, E. Oyallon, and F. Bach, “On lazy training in differentiable programming,” 2018. [Online]. Available: https://arxiv.org/abs/1812.07956

[26] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, “Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks,” in *International Conference on Machine Learning.* PMLR, 2019, pp. 322–332.

[27] Z. Chen, Y. Cao, D. Zou, and Q. Gu, “How much over-parameterization is sufficient to learn deep relu networks?” *arXiv preprint arXiv:1911.12360*, 2019.

[28] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, “The role of over-parametrization in generalization of neural networks,” in *International Conference on Learning Representations*, 2018.

[29] S. Hochreiter and J. Schmidhuber, “Flat minima,” *Neural computation*, vol. 9, no. 1, pp. 1–42, 1997.

[30] S. Jastrzkebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, “Three factors influencing minima in sgd,” *arXiv preprint arXiv:1711.04623*, 2017.

[31] L. Wu, Z. Zhu *et al.*, “Towards understanding generalization of deep learning: Perspective of loss landscapes,” *arXiv preprint arXiv:1706.10239*, 2017.

[32] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” *arXiv preprint arXiv:1706.02677*, 2017.

[33] E. Hoffer, I. Hubara, and D. Soudry, “Train longer, generalize better: closing the generalization gap in large batch training of neural networks,” *arXiv preprint arXiv:1705.08741*, 2017.

[34] V. Luo and Y. Wang, “How many factors influence minima in sgd?” *arXiv preprint arXiv:2009.11858*, 2020.

[35] R. Ge, R. Kuditipudi, Z. Li, and X. Wang, “Learning two-layer neural networks with symmetric inputs,” *arXiv preprint arXiv:1810.06793*, 2018.

[36] A. Bakshi, R. Jayaram, and D. P. Woodruff, “Learning two layer rectified neural networks in polynomial time,” in *Conference on Learning Theory.* PMLR, 2019, pp. 195–268.

[37] Y. Tian, “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.* PMLR, 2017, pp. 3404–3413.

[38] M. Soltanolkotabi, "Learning relus via gradient descent," *arXiv preprint arXiv:1705.04591*, 2017.

[39] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," *arXiv preprint arXiv:1705.09886*, 2017.

[40] X. Zhang, Y. Yu, L. Wang, and Q. Gu, "Learning one-hidden-layer relu networks via gradient descent," in *The 22nd international conference on artificial intelligence and statistics.* PMLR, 2019, pp. 1524–1534.

[41] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, "Recovery guarantees for one-hidden-layer neural networks," in *International conference on machine learning.* PMLR, 2017, pp. 4140–4149.

[42] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.

[43] S. Kim and M. Pilanci, "Convex relaxations of relu neural networks approximate global optima in polynomial time," in *Forty-first International Conference on Machine Learning.*

[44] A. Mishkin, A. Sahiner, and M. Pilanci, "Fast convex optimization for two-layer relu networks: Equivalent model classes and cone decompositions," in *International Conference on Machine Learning.* PMLR, 2022, pp. 15 770–15 816.

[45] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[46] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge university press, 2019, vol. 48.

[47] R. Vershynin, *High-dimensional probability: An introduction with applications in data science.* Cambridge university press, 2018, vol. 47.

[48] S. A. Van De Geer and P. Bühlmann, "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[49] V. De la Pena and E. Giné, *Decoupling: from dependence to independence.* Springer Science & Business Media, 2012.

[50] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "Linearized two-layers neural networks in high dimension," *The Annals of Statistics*, vol. 49, no. 2, pp. 1029–1054, 2021.