PROVABLE IDENTIFIABILITY OF RELU NEURAL NETWORKS VIA LASSO REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LASSO regularization is a popular regression tool to enhance the prediction accuracy of statistical models by performing variable selection through the ℓ_1 penalty, initially formulated for the linear model and its variants. In this paper, the territory of LASSO is extended to the neural network model, a fashionable and powerful nonlinear regression model. Specifically, given a neural network whose output y depends only on a small subset of input $\boldsymbol{x}$, denoted by $\mathcal{S}^\star$, we prove that the LASSO estimator can stably reconstruct the neural network and identify $\mathcal{S}^\star$ when the number of samples scales logarithmically with the input dimension. This challenging regime has been well understood for linear models while barely studied for neural networks. Our theory lies in an extended Restricted Isometry Property (RIP)-based analysis framework for two-layer ReLU neural networks, which may be of independent interest to other LASSO or neural network settings. Based on the result, we further propose a neural network-based variable selection method. Experiments on simulated and real-world datasets show the promising performance of our variable selection approach compared with classical techniques.

1 INTRODUCTION

Given n observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, we often model them with the regression form of $y_i = f(\boldsymbol{x}_i) + \xi_i$, with an unknown function f, $\boldsymbol{x}_i \in \mathbb{R}^p$ being the input variables, and ξ_i representing statistical errors. A general goal is to estimate a regression function $\widehat{f}_n$ close to f for prediction or interpretation. This is a challenging problem when the input dimension p is comparable or even much larger than the data size n. For linear regressions, namely $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) regularization has been established as a standard tool to estimate f. The LASSO has also been successfully used and studied in many nonlinear models such as generalized linear models (Van de Geer et al., 2008), proportional hazards models (Tibshirani, 1997), and neural networks (Goodfellow et al., 2016). In particular, the LASSO regularization has been added into the standard deep learning toolbox of many open-source libraries, e.g., Tensorflow (Abadi et al., 2016) and Pytorch (Paszke et al., 2019). Despite the practical success of LASSO, its theoretical efficacy in neural networks is barely studied. In particular, it remains unclear whether LASSO may be used for variable selection and subsequent interpretations of a learned model.

Meanwhile, in the theoretical study of neural networks, there has been remarkable progress towards understanding their approximation errors (Barron, 1993; 1994) and generalization errors (Barron & Klusowski, 2019; Schmidt-Hieber, 2017; Bauer & Kohler, 2019). Nevertheless, the identifiability issue of neural networks has been an unsolved challenge. Specifically, supposing that data observations are generated from a neural network with only a few nonzero coefficients (or its proximity), the identifiability concerns the possibility of identifying those coefficients. In practice, such sparsity of neural coefficients may be interpreted as a sparse set of input variables that are genuinely relevant to the response, which may be of scientific interest.

In this paper, we consider the following class of two-layer ReLU neural networks.

$$\mathcal{F}_r = \left\{ f : \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j), \text{ where } a_j, b_j \in \mathbb{R}, \boldsymbol{w}_j \in \mathbb{R}^p \right\}.$$

Here, p and r denote the input dimension and the number of neurons, respectively. We will assume that a neural network model of parsimoniousness generates the data. In other words, some of the

input signals are irrelevant to explain y, or some of the network structure in f is redundant for modeling $(y, \boldsymbol{x})$. Different forms of parsimoniousness were assumed in (Schmidt-Hieber, 2017; Bauer & Kohler, 2019; Barron & Klusowski, 2019) to derive tight neural network risk bounds. We raise the following two questions to understand the nonlinear nature of neural networks.

First, if the underlying system f admits a parsimonious representation, meaning that only a small set of input variables, $\mathcal{S}^\star$, is relevant, can we identify them with high probability given possibly noisy measurements $(y_i, \boldsymbol{x}_i)$, for $i = 1, \ldots, n$? Second, is such a $\mathcal{S}^\star$ estimable, meaning that it can be solved from an optimization problem with high probability, even in small-n and large-p regimes?

To address the above questions, we will establish a theory for neural networks with the LASSO regularization by considering the minimization problem

$$\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \|\boldsymbol{W}\|_1 \qquad \text{subject to} \qquad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 \le \sigma^2, \tag{1}$$

which is an alternative version of the L_1-regularization. More notational details will be introduced in Subsection 3.2.

We theoretically show that the LASSO-type estimator can stably identify ReLU neural networks with sparse input signals, up to a permutation of hidden neurons. Our result is rather general as it applies to noisy observations of y and dimension regimes where the sample size n is much smaller than the number of input variables p. Our theory gives positive answers to the above questions. The theory was derived based on new concentration bounds and function analysis that may be interesting in their own right.

Inspired by the developed theory, we also propose a neural network-based variable selection method. The idea is to use the neural system as a vehicle to model nonlinearity and extract significant variables. To the best of our knowledge, the identifiability perspective of neural networks and its subsequent variable selection method have not been seen in the literature. Through various experimental studies, we show encouraging performance of the technique in identifying a sparse set of significant variables from large dimensional data, even if the underlying data are not generated from a neural network. Compared with popular approaches based on tree ensembles and linear-LASSO, the developed method is suitable for variable selection from nonlinear, large-dimensional, and low-noise systems.

The rest of the paper is outlined as follows. Section 2 reviews the related work. Section 3 introduces the main theoretical results and develops a practical algorithm to perform variable selection. Section 4 uses simulated and real-world datasets to demonstrate the proposed theory and algorithm. Section 5 concludes the paper.

2 Related work

Linear models. The variable selection problem is also known as support recovery or feature selection in different literature. Selection consistency requires that the probability of $\mathrm{supp}(\widehat{\boldsymbol{w}}) = \mathrm{supp}(\boldsymbol{w})$ converges to one as $n \to \infty$. The mainstream approach to select a parsimonious sub-model is to either solve a penalized regression problem or iteratively pick up significant variables. The existing methods differ in how they incorporate unique domain knowledge (e.g., sparsity, multicollinearity, group behavior) or what desired properties (e.g., consistency in coefficient estimation, consistency in variable selection) to achieve. For instance, consistency of the LASSO method (Tibshirani, 1996) in estimating the significant variables has been extensively studied under various technical conditions, including sparsity, mutual coherence (Donoho & Huo, 2001), restricted isometry (Candes & Tao, 2005), irrepresentable condition (Zhao & Yu, 2006), and restricted eigenvalue (Bickel et al., 2009).

Neural network models. Neural networks have been practically successful in modeling a wide range of nonlinear systems. Analytically, a universal approximation theorem was established that shows any continuous multivariate function can be represented precisely by a polynomial-sized two-layer network (Kolmogorov, 1957). It was later shown that any continuous function could be approximated arbitrarily well by a two-layer perceptron with sigmoid activation functions (Cybenko, 1989), and an approximation error bound of using two-layer neural networks to fit arbitrary smooth functions has been established (Barron, 1993; 1994). Statistically, generalization error bounds for two-layer neural networks (Barron, 1994) and multi-layer networks (Neyshabur et al., 2015; Golowich et al.,

2017) have been developed. From an optimization perspective, the parameter estimation of neural networks could be cast into a tensor decomposition problem where a provably global optimum can be obtained (Janzamin et al., 2015; Ge et al., 2017; Mondelli & Montanari, 2018). Very recently, a dimension-free Rademacher complexity to bound the generalization error for deep ReLU neural networks was developed to avoid the curse of dimensionality (Barron & Klusowski, 2019). It was proved that certain deep neural networks with few non-zero network parameters could achieve minimax rates of convergence (Schmidt-Hieber, 2017). A tight error bound free from the input dimension was developed by assuming that the data is generated from a generalized hierarchical interaction model (Bauer & Kohler, 2019). Overall, theoretical studies have primarily focused on the prediction risk bounds or generalization error bounds of estimated neural networks.

3 Main Results

3.1 Notation

Let $\boldsymbol{u}_{\mathcal{S}}$ denote the vector whose entries indexed in the set $\mathcal{S}$ remain the same as those in $\boldsymbol{u}$, and the remaining entries are zero. For any matrix $\boldsymbol{W} \in \mathbb{R}^{p\times r}$, we define

$$\|\boldsymbol{W}\|_1 = \sum_{1\le k\le p, 1\le j\le r} |w_{kj}|, \quad \|\boldsymbol{W}\|_{\mathrm{F}} = \left(\sum_{1\le k\le p, 1\le j\le r} w_{kj}^2\right)^{1/2}.$$

Similar notations apply to vectors. The inner product of two vectors is denoted as $\langle \boldsymbol{u}, \boldsymbol{v}\rangle$. Let $\boldsymbol{w}_j$ denote the j-th column of $\boldsymbol{W}$. The sparsity of a matrix $\boldsymbol{W}$ refers to the number of nonzero entries in $\boldsymbol{W}$. Let $\mathcal{N}(0, \boldsymbol{I}_p)$ denote the standard p-dimensional Gaussian distribution, and $\mathbb{1}(\cdot)$ denote the indicator function. The rectified linear unit (ReLU) function is defined by $\mathrm{relu}(v) = \max\{v, 0\}$ for all $v \in \mathbb{R}$.

3.2 Formulation

Given n independently and identically distributed (i.i.d.) observations $\{\boldsymbol{x}_i, y_i\}_{1\le i\le n}$ satisfying

$$y_i = \sum_{j=1}^{r} a_j^\star \cdot \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star) + \xi_i \qquad \text{with } \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_p), \tag{2}$$

where r is the number of neurons, $a_j^\star \in \{1, -1\}$, $\boldsymbol{w}_j^\star \in \mathbb{R}^p$, $b_j^\star \in \mathbb{R}$, and ξ_i denotes the random noise or approximation error obeying

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i^2 \le \sigma^2. \tag{3}$$

In the above formulation, the assumption $a_j^\star \in \{1, -1\}$ does not lose generality since $a \cdot \mathrm{relu}(b) = ac \cdot \mathrm{relu}(b/c)$ for any $c > 0$. The setting Equation (3) is for simplicity. If ξ_i's are unbounded random variables, our theoretical result later on still holds, and more explanations are in the supplement. The ξ_i's are not necessarily i.i.d. and σ is allowed to be zero, which reduces to the noiseless scenario.

Let $\boldsymbol{W}^\star = [\boldsymbol{w}_1^\star, \ldots, \boldsymbol{w}_r^\star] \in \mathbb{R}^{p\times r}$ denote the data-generating coefficients. The question we aim to address is whether we can stably identify those nonzero elements, given that most entries in $\boldsymbol{W}^\star$ are zero. The study of neural networks from an identifiability perspective is exciting and essential. Unlike the generalizability problem that studies the predictive performance of machine learning models, the identifiability may be used to interpret modeling results and help scientists make trustworthy decisions. To illustrate this point, we will propose to use neural networks for variable selection in Subsection 3.4.

To answer the above question, we propose to study the following LASSO-type optimization. Let $(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}})$ be a solution to the following optimization problem,

$$\min_{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}} \|\boldsymbol{W}\|_1 \qquad \text{subject to} \qquad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \cdot \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 \le \sigma^2, \tag{4}$$

within the feasible range $\boldsymbol{a} \in \{1, -1\}^r$, $\boldsymbol{W} \in \mathbb{R}^{p\times r}$, and $\boldsymbol{b} \in \mathbb{R}^r$.

Intuitively, the optimization operates under the constraint that the training error is not too large and the objective function tends to sparsify $\boldsymbol{W}$. Under some regularity conditions, we will prove that the solution is indeed sparse and close to the truth.

Assumption 1. *Suppose that for some constant $B \geq 1$,*

$$1 \leq \|\boldsymbol{w}_j^\star\|_2 \leq B \qquad \textit{and} \qquad |b_j^\star| \leq B \qquad \forall 1 \leq j \leq r. \tag{5}$$

In addition, we assume that for some constant $\omega > 0$,

$$\max_{j\neq k} \frac{\left|\langle \boldsymbol{w}_j^\star, \boldsymbol{w}_k^\star\rangle\right|}{\|\boldsymbol{w}_j^\star\|_2 \|\boldsymbol{w}_k^\star\|_2} \leq \frac{1}{r^\omega}. \tag{6}$$

The condition in (5) is a normalization only for technical convenience, since we can re-scale $\boldsymbol{w}_j, b_j, y_i, \sigma$ proportionally without loss of generality. Though this condition implicitly requires $\boldsymbol{w}_j^\star \neq \boldsymbol{0}$ for all $j = 1, \ldots, r$, it is reasonable since it means the neuron j is not used/activated. The condition in (6) requires that the angle of any two different coefficient vectors is large enough. We will provide an alternative assumption in the supplementary document.

3.3 Main theorem

Our main result shows that if $\boldsymbol{W}^\star$ is sparse, one can stably reconstruct a neural network when the number of samples (n) scales logarithmically with the input dimension (p). We only focus on the varying n and p and implicitly assume that the sparsity of $\boldsymbol{W}^\star$ and the number of neurons r are fixed. A skeptical reader may ask how the constants exactly depend on the sparsity and r. We will provide a more elaborated theorem in the supplementary document.

Theorem 1. *Under the Assumption 1, there exist some universal constants $c_1, c_2, c_3 > 0$ depending only (polynomially) on the sparsity of $\boldsymbol{W}^\star$, such that: for any $\delta > 0$, one has with probability at least $1 - \delta$,*

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star \qquad \textit{and} \qquad \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^\star\boldsymbol{\Pi}^\top\|_{\mathrm{F}} + \|\widehat{\boldsymbol{b}} - \boldsymbol{\Pi}\boldsymbol{b}^\star\|_2 \leq c_1\sigma \tag{7}$$

for some permutation matrix $\boldsymbol{\Pi}$, provided that

$$n > c_2 \log^4 \frac{p}{\delta} \qquad \textit{and} \qquad \sigma < c_3. \tag{8}$$

Remark 1 (Interpretations of Theorem 1)**.** *The permutation matrix $\boldsymbol{\Pi}$ is necessary since the considered neural networks produce identical predictive distributions (of y conditional $\boldsymbol{x}$) under any permutation of the hidden neurons. The result says that the underlying neural coefficients can be stably estimated even when the sample size n is much smaller than the number of variables p. Also, the estimation error bound is at the order of σ, the specified noise level in (3).*

Suppose that we define the signal-to-noise ratio (SNR) to be $\mathbb{E}\|\boldsymbol{x}\|^2/\sigma^2$. An alternative way to interpret the theorem is that a large SNR ensures the global minimizer to be close to the ground truth with high probability. One may wonder what if the $\sigma < c_3$ condition is not met. We note that if σ is too large, the error bound in (7) would be loose, and it is not of much interest anyway. In other words, if the SNR is small, we may not be able to estimate parameters stably. This point will be demonstrated by experimental studies in Section 4.

The estimation results in Theorem 1 can be translated into variable selection results as shown in the following Corollary 1. The connection is based on the fact that if i-th variable is redundant, the underlying coefficients associated with it should be zero. Let $\boldsymbol{w}_{i,\cdot}^\star$ denote the i-th row of $\boldsymbol{W}^\star$. Then,

$$\mathcal{S}^\star = \{1 \leq i \leq p : \|\boldsymbol{w}_{i,\cdot}^\star\|_2 > 0\}$$

characterizes the "significant variables." Corollary 1 says that the set of variables with non-vanished coefficient estimates contains all the significant variables. The corollary also shows with a suitable shrinkage of the coefficient estimates, one can achieve variable selection consistency.

Corollary 1 (Variable selection)**.** *Let $\widehat{\mathcal{S}}_0$ and $\widehat{\mathcal{S}}_{c_1\sigma} \subseteq \{1, \ldots, p\}$ denote the sets of i's such that $\|\widehat{\boldsymbol{w}}_{i,\cdot}\|_2 > 0$ and $\|\widehat{\boldsymbol{w}}_{i,\cdot}\|_2 > c_1\sigma$, respectively. Under the same assumption as in Theorem 1, and $\inf\|\boldsymbol{w}_{i,\cdot}^\star\|_2 > c_1\sigma$, for any $\delta > 0$, one has*

$$\mathbb{P}(\mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_0) \geq 1 - \delta \qquad \textit{and} \qquad \mathbb{P}(\mathcal{S}^\star = \widehat{\mathcal{S}}_{c_1\sigma}) \geq 1 - \delta,$$

provided that $n > c_2 \log^4 \frac{p}{\delta}$ and $\sigma < c_3$.

Considering the noiseless scenario $\sigma = 0$, Theorem 1 also implies the following corollary.

Corollary 2 (Unique parsimonious representation)**.** *Under the Assumption 1, there exist universal constants $c_1, c_2 > 0$ depending only on the sparsity of $\boldsymbol{W}^\star$ such that: for any $\delta > 0$, one has with probability at least $1 - \delta$,*

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star, \qquad \textit{and} \qquad \widehat{\boldsymbol{W}} = \boldsymbol{W}^\star\boldsymbol{\Pi}^\top, \qquad \textit{and} \qquad \widehat{\boldsymbol{b}} = \boldsymbol{\Pi}\boldsymbol{b}^\star$$

for some permutation matrix $\boldsymbol{\Pi}$, provided that $n > c_2 \log^4 \frac{p}{\delta}$.

Corollary 2 says that among all the possible representations $\boldsymbol{W}$ of (2) (with $\xi_i = 0$), the one(s) with the smallest L_1-norm must be identical to $\boldsymbol{W}^\star$ up to a column permutation with high probability. In other words, the most parsimonious representation (in the sense of L_1 norm) of two-layer ReLU neural networks is unique. This observation addresses the questions raised in Section 1.

Remark 2 (Sketch proof of Theorem 1)**.** *The proof of Theorem 1 is highly nontrivial, and it is included in the supplementary document. Next, we briefly explain the sketch of the proof. First, we will define what we refer to as D_1-distance and D_2-distance between $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ and $(\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)$. These distances can be regarded as the counterpart of the classical L_1 and L_2 distances between two vectors, but allow the invariance under any permutation of neurons (Remark 1). Then, we let*

$$\Delta_n := \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2,$$

here $\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}$ is the solution of Equation (4), and develop the following upper and lower bounds of it.

$$\Delta_n \le c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right)D_1^2 + c_6\sigma^2 \qquad \textit{and} \qquad \Delta_n \ge c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \tag{9}$$

hold with probability at least $1 - \delta$, provided that $n \ge c_5 S^3 r^4 \log^4 \frac{p}{\delta}$, for some constants c_4, c_5, c_6, and S to be specified. Here, the upper bound will be derived from a series of elementary inequalities. The lower bound is reminiscent of the Restricted Isometry Property (RIP) (Candes & Tao, 2005) for linear models. We will derive it from the lower bound of the population counterpart by concentration arguments, namely

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star)\right]^2 \ge c \min\left\{\frac{1}{r}, D_2^2\right\},$$

for some constant $c > 0$. The bounds in (9) imply that with high probability,

$$c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \le c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right)D_1^2 + c_6\sigma^2,$$

Using this and an inequality connecting D_1 and D_2, we can prove the final result.

3.4 Variable selection

To solve Equation (4) in practice, we consider the following alternative problem,

$$\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \cdot \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 + \lambda\|\boldsymbol{W}\|_1. \tag{10}$$

The above optimization problem can be numerically solved using algorithms such as stochastic gradient descent (Bottou, 2010) and ADAM (Kingma & Ba, 2014), available from many open-source libraries. We discuss some details regarding the variable selection using LASSO regularized neural networks.

Tuning parameters. Given a labeled dataset in practice, we will need to tune several hyper-parameters, including the penalty term λ, number of neurons r, learning rate, and number of epochs. We suggest the usual approach that splits the training data into training and validation parts. The

training data are used to estimate neural networks for a set of candidate hyper-parameters. The most suitable candidate will be identified based on the predictive performance on the validation data. We point out that there are gaps between the developed theory and the selection method in practice. For example, the selected number of hidden neurons r based on the training data may violate the constant bounds in Assumption 1. Fortunately, from our experimental studies, the results are not very sensitive to the choice of r.

Variable importance. Inspired by Corollary 1, we assign the norm of $\widehat{\boldsymbol{w}}_{i,\cdot}$ as the importance of the i-th variable, for $i = 1, \ldots, p$. As Corollary 1 implies, we can accurately identify all the significant variables in $\mathcal{S}^\star$ with high probability if we correctly set the cutoff value $c_1\sigma$.

Setting the cutoff value. In practice, we have no idea of the threshold $c_1\sigma$. But it is conceivable that variables with large importance are preferred over those with near-zero importance. This inspires us to cluster the variables into two groups based on their importance. Here, we suggest two possible approaches. The first is to use a data-driven approach such as *k*-means and Gaussian mixture model (GMM). The second is to manually set a threshold value according to domain knowledge on the number of important variables.

4 EXPERIMENTS

We perform experimental studies to show the promising performance of the proposed variable selection method. We compare the variable selection accuracy and prediction performance of the proposed algorithm ('NN') with several baseline methods, including the LASSO ('LASSO'), orthogonal matching pursuit ('OMP'), random forest ('RF'), and gradient boosting ('GB'). The implementation follows Subsection 3.4. In particular, we used ADAM to optimize and GMM to select significant variables. The parameters grid of 'NN' is set as the penalty term $\lambda \in \{0.1, 0.05, 0.01, 0.005\}$, the number of neurons $r \in \{20, 50, 100\}$, the learning rate in set $\{0.05, 0.01, 0.005\}$, and the number of epochs in set $\{200, 500, 1000\}$. We use the absolute value of the estimated coefficient as the variable importance for 'LASSO' and 'OMP', and use the self-produced feature importance for the tree-based methods. All the computation is done on the 2.3GHz Quad-Core Intel Core i5 with Intel Iris Plus Graphics 655.

4.1 SYNTHETIC DATASETS

4.1.1 NN-GENERATED DATASET

The first experiment uses the data generated from Equation (2) with $p = 100$ variables and $r = 16$ neurons. The first 10 rows of neural coefficients $\boldsymbol{W}$ are independently generated from the standard uniform distribution and the remaining rows are zeros, representing 10 significant variables. The neural biases $\boldsymbol{b}$ are also generated from the standard uniform distribution. The signs of neurons, $\boldsymbol{a}$, follow an independent Bernoulli distribution. The training size is $n = 500$ and the test size is 2000. The noise level σ is set to be 0, 0.5, 1, and 5. For each σ, we evaluate the number of correctly selected variables ('TP') and wrongly selected variables ('FP'), along with the test error. The procedure is independently replicated 100 times. The average numbers of selected features are reported in Table 1. The test errors are reported in Table 2.

The results show that 'NN' has the best performance on both the selection and prediction. The performance of tree-based methods is surprisingly undesirable. Also, when the noise level σ increases, or the SNR decreases, all the methods perform worse. Another observation is that selection accuracy and prediction performance are positively associated for 'NN', but this is not the case for other methods.

4.1.2 LINEAR DATASET

This experiment considers data generated from a linear model $y = \boldsymbol{x}^\top \boldsymbol{\beta} + \xi$, where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$, $\xi \sim \mathcal{N}(0, \sigma^2)$, and $\boldsymbol{x}$ follows a multivariate Gaussian distribution whose (i, j)-th correlation is $0.5^{|i-j|}$. Among the $p = 8$ features, only three of them are significant. The training size is $n = 60$ and the test size is 200. The other settings are the same as Subsubsection 4.1.1. The results are presented in Tables 3 and 4.

Table 1: Performance comparison on the NN-generated data, in terms of the number of correctly ('TP') and wrongly ('FP') selected features for different σ. The standard errors are within 0.3, except for the 'FP' of 'LASSO', which is 0.6.

Method		$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
NN	TP	**10.0**	**9.7**	**9.8**	**6.7**
	FP	0.0	0.1	1.8	1.3
LASSO	TP	9.5	8.8	8.6	6.5
	FP	12.4	10.8	10.5	9.3
OMP	TP	8.4	8.0	8.6	5.8
	FP	0.1	0.4	0.0	0.4
RF	TP	6.3	6.8	7.4	4.2
	FP	0.1	0.2	0.7	0.8
GB	TP	7.9	7.8	8.4	5.6
	FP	1.2	1.5	3.1	3.5

Table 2: Performance comparison on the NN-generated data, in terms of the average mean squared error for different σ. The standard errors of 'NN' are within 0.1, while linear methods are around 0.4 and tree-based methods are 0.8.

Method	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
NN	**0.55**	**0.72**	**1.18**	**4.75**
LASSO	5.05	5.71	5.07	5.50
OMP	5.27	4.75	5.01	6.17
RF	10.01	8.86	9.22	9.70
GB	5.67	5.84	6.58	10.92

Table 3: Performance comparison on the linear data, in terms of the number of correctly ('TP') and wrongly ('FP') selected features for different σ. The standard errors are within 0.1.

Method		$\sigma = 0$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$
NN	TP	**3.0**	2.7	2.2	1.6
	FP	0.0	0.0	0.1	0.3
LASSO	TP	2.7	**3.0**	**2.5**	**2.1**
	FP	0.0	0.0	0.1	0.3
OMP	TP	3.0	2.8	2.5	1.7
	FP	0.0	0.0	0.3	0.9
RF	TP	1.5	1.5	1.7	1.4
	FP	0.0	0.0	0.0	0.3
GB	TP	1.3	1.5	1.3	1.0
	FP	0.0	0.0	0.0	0.1

Table 4: Performance comparison on the linear data, in terms of the number of average mean squared error for different σ. The standard errors are within 0.2 when $\sigma < 5$, and about 0.4 when $\sigma = 5$.

Method	$\sigma = 0$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$
NN	0.11	0.43	2.11	5.42
LASSO	**0.00**	0.13	**1.32**	**4.97**
OMP	0.00	**0.09**	1.47	6.61
RF	3.54	3.52	4.98	10.00
GB	2.68	3.04	5.76	14.20

The results show that the linear model-based methods 'LASSO' and 'OMP' have the best overall performance, which is expected since the underlying data are from a linear model. The proposed approach 'NN' is almost as good as the linear methods. On the other hand, the tree-based methods 'RF' and 'GB' perform significantly worse. We think that this is because the sample size $n = 60$ is quite small, so the tree-based methods have a large variance. Meanwhile, the 'NN' uses L_1 penalty to alleviate the over-parameterization and consequently spots the relevant variables. Additionally, 'NN' exhibits a positive association between the selection accuracy and prediction performance, while the tree-based methods do not.

4.1.3 FRIEDMAN DATASET

This experiment uses the Friedman dataset with the following nonlinear data-generating process, $y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \xi$. We generate standard Gaussian predictors $\boldsymbol{x}$ with a dimension of $p = 50$. The training size is $n = 500$ and the test size is 2000. Other settings are the same as before. The results are summarized in Tables 5 and 6. For this nonlinear dataset, 'NN' almost always finds the significant variables and excludes redundant ones, which is better than tree-based methods. At the same time, the linear methods fail to select the quadratic factor x_3. Moreover, we find that when different methods are compared, the method with a better selection accuracy does not necessarily exhibit a better prediction and vice versa.

4.2 BGSBOY DATASET

The BGSBoy dataset involves 66 boys from the Berkeley guidance study (BGS) of children born in 1928-29 in Berkeley, CA (Tuddenham, 1954). The dataset includes the height ('HT'), weight ('WT'),

Table 5: Performance comparison on the Friedman data, in terms of the number of correctly ('TP') and wrongly ('FP') selected features for different σ. The standard errors are within 0.15, except for 'FP' of 'LASSO', which is around 0.3.

Method		$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
NN	TP	4.8	**5.0**	**5.0**	**4.9**
	FP	0.0	0.0	0.0	0.1
LASSO	TP	4.04	4.06	4.1	4.13
	FP	2.03	2.24	2.22	4.72
OMP	TP	4.0	4.0	4.0	4.0
	FP	0.17	0.13	0.09	0.17
RF	TP	4.64	4.54	4.72	3.87
	FP	0.03	0.02	0.02	0.22
GB	TP	**4.98**	4.94	4.94	4.5
	FP	0.03	0.0	0.02	0.56

Table 6: Performance comparison on the Friedman data, in terms of the average mean squared error for different σ. The standard errors are within 0.1.

Method	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
NN	1.89	2.08	2.32	**5.69**
LASSO	6.28	6.08	6.2	6.94
OMP	5.8	5.98	5.45	6.31
RF	5.22	5.43	5.36	7.82
GB	**1.75**	**1.87**	**2.18**	7.57

Table 7: Experiment results of different methods on the BGSBoy dataset. RMSE: the mean of the root mean squared error(standard error). Top 3 features: the feature name(number of selection, out of 100 times).

Method	**NN**	**LASSO**	**OMP**	**RF**	**GB**
RMSE	0.04 (0.003)	0.05 (0.002)	0.05 (0.002)	3.07 (0.154)	2.4 (0.142)
Top 3 frequently selected features	WT18(100) HT18(81) N/A	WT18(100) HT18(71) HT9(51)	WT18(100) HT18(64) HT9(16)	WT18(91) LG18(86) LG9(2)	WT18(90) LG18(59) HT18(8)

leg circumference ('LG'), strength ('ST') at different ages (2, 9, 18 years), and body mass index ('BMI18'). We choose 'BMI18' as the response, which is defined as follows.

$$\text{BMI18} = \text{WT18}/(\text{HT18}/100)^2, \tag{11}$$

where WT18 and HT18 denote the weight and height at the age of 18, respectively. In other words, 'WT18' and 'HT18' are sufficient for modeling the response among $p = 10$ variables. Other variables are correlated but redundant. The training size is $n = 44$ and the test size is 22. Other settings are the same as before. We compare the prediction performance and explore the three features which are most frequently selected by each method. The results are summarized in Table 7.

From the results, all of the methods can identify 'WT18' most of the time. Nevertheless, 'NN' only selects 'WT18' and 'HT18' in all the replications, while other methods sometimes select features that are redundant but correlated with the response. For example, tree-based methods usually miss 'HT18' but select 'LG18' instead. The results indicate that only 'NN' can stably identify the underlying significant variables. Interestingly, we find that the linear methods still predict well in this experiment. The reason is that Equation (11) can be well-approximated by a first-order Taylor expansion on 'HT18' at the value around 180, and the range of 'HT18' is within a small interval around 180.

4.3 UJIINDOORLOC DATASET

The UJIIndoorLoc dataset aims to solve the indoor localization problem via WiFi fingerprinting and other variables such as the building and floor numbers. A detailed description can be found in (Torres-Sospedra et al., 2014). Specifically, we have 520 Wireless Access Points (WAPs) signals (which are continuous variables) and 'FLOOR', 'BUILDINGID', 'SPACEID', 'RELATIVEPOSITION', 'USERID', and 'PHONEID' as categorical variables. The response variable is a user's longitude ('Longitude'). The dataset has 19937 observations. We randomly sample 3000 observations and split them into $n = 2000$ for training and 1000 for test. As part of the pre-processing, we create binary dummy variables for the categorical variables, which results in $p = 681$ variables in total. We explore the ten features that are most frequently selected by each method. We set the cutoff value as the tenth-largest variable importance. The procedure is independently replicated 100 times. The results are reported in Table 8.

Table 8: Experiment results of different methods on the UJIINdoor dataset. RMSE: the mean of the root mean squared error(standard error). Top 10 features: the feature name(number of selection, out of 100 times).

Method	NN	LASSO	OMP	RF	GB
RMSE	9.6(0.067)	14.23(0.046)	16.58(0.052)	9.49(0.053)	10.3(0.043)
Top 10 frequently selected features	BUILDINGID_2(100)	BUILDINGID_1(100)	BUILDINGID_1(100)	BUILDINGID_1(100)	BUILDINGID_2(100)
	BUILDINGID_1(100)	USERID_16(100)	BUILDINGID_2(100)	BUILDINGID_2(100)	BUILDINGID_1(100)
	USERID_16(97)	BUILDINGID_2(100)	WAP099(81)	WAP120(82)	WAP141(91)
	SPACEID_202(86)	USERID_9(94)	USERID_10(70)	WAP141(76)	WAP120(87)
	USERID_8(76)	WAP099(90)	USERID_16(60)	WAP117(75)	WAP099(68)
	USERID_9(74)	USERID_10(72)	USERID_7(58)	WAP173(74)	WAP113(67)
	PHONEID_14(65)	USERID_7(67)	WAP124(55)	WAP118(58)	WAP117(60)
	FLOOR_3(61)	WAP121(49)	USERID_9(46)	WAP167(57)	PHONEID_14(58)
	SPACEID_201(52)	WAP118(34)	WAP120(31)	WAP035(52)	WAP114(48)
	SPACEID_203(41)	WAP124(28)	WAP117(29)	WAP113(33)	WAP167(47)

Based on the results, the 'NN' achieves similar prediction performance as 'RF' and significantly outperforms other methods. As for variable selection, since 'BUILDING' greatly influences the location from our domain knowledge, it is non-surprisingly selected by all methods in every replication. However, except for 'BUILDING', different methods select different variables. Some overlaps, e.g., 'PHONEID_14' selected by 'NN' and 'GB', 'USERID_16' selected by 'NN' and 'LASSO', indicate the potentially important variables. Nevertheless, those methods do not achieve an agreement for variable selection. 'NN' implies that all the WAPs signals are weak while categorical variables provide more information about the user location. Given the extremely high missing rate of WAPs signals (97% on average, as reported in (Torres-Sospedra et al., 2014)), we think that the interpretation of 'NN' is reasonable.

4.4 Summary

The experiment results show the following points. First, 'NN' can stably identify the important variables and have competitive prediction performance compared with the baselines. Second, the increase of the noise level will hinder both the selection and prediction performance. Third, the LASSO regularization is crucial for 'NN' to avoid over-fitting, especially for small data. Fourth, the selection and prediction performances are often positively associated for 'NN', but may not be the case for baseline methods.

5 Concluding Remarks

We established a theory for the use of LASSO in two-layer ReLU neural networks. In particular, we showed that the LASSO estimator could stably reconstruct the neural network coefficients and identify the critical underlying variables under reasonable conditions. We also proposed a practical method to solve the optimization and perform variable selection.

We briefly remark on some interesting further work. First, the algorithm can be directly extended to deeper neural networks. It will be exciting to generalize the main theorem to the multi-layer cases. Second, the developed theory may be extended to study the variable selection for general nonlinear functions due to the universal approximation theorem.

The supplementary material includes detailed proofs and Python codes used for the experiments.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *Proc. USENIX*, pp. 265–283, 2016.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.

Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.*, 14(1):115–133, 1994.

Andrew R Barron and Jason M Klusowski. Complexity, statistical risk, and metric entropy of deep nets using total path variation. *arXiv preprint arXiv:1902.00800*, 2019.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Stat.*, 47(4):2261–2285, 2019.

Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *Ann. Stat.*, 37(4):1705–1732, 2009.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT*, pp. 177–186. Springer, 2010.

EJ Candes and T Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

George Cybenko. Approximations by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2:183–192, 1989.

Vu Dinh and Lam Si Tung Ho. Consistent feature selection for analytic deep neural networks. *arXiv preprint arXiv:2010.08097*, 2020.

David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inf. Theory*, 47(7):2845–2862, 2001.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

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

Andrei Nikolaevich Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*, volume 114, pp. 953–956. Russian Academy of Sciences, 1957.

Marco Mondelli and Andrea Montanari. On the connection between learning two-layers neural networks and tensor decomposition. *arXiv preprint arXiv:1802.07301*, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. *Conf. Learn. Theory*, pp. 1376–1401, 2015.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B*, 58(1): 267–288, 1996.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Stat. Med.*, 16(4): 385–395, 1997.

Joaquín Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Joan P Avariento, Tomás J Arnau, Mauri Benedito-Bordonau, and Joaquín Huerta. Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems. In *Int. Conf. IPIN*, pp. 261–270. IEEE, 2014.

Read D Tuddenham. Physical growth of california boys and girls from birth to eighteen years. *Univ. Calif. Publ. Child Dev.*, 1:183–364, 1954.

Sara A Van de Geer et al. High-dimensional generalized linear models and the LASSO. *Ann. Stat.*, 36(2):614–645, 2008.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7(Nov): 2541–2563, 2006.

Supplementary Document

A Main Results

We first restate the main assumptions and results in the following for ease of understanding. Given n i.i.d. observations $\{\boldsymbol{x}_i, y_i\}_{1\le i\le n}$ satisfying

$$y_i = \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star) + \xi_i, \qquad \text{with } \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{12}$$

where ξ_i denotes the random noise and/or approximation error obeying

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 \le \sigma^2, \tag{13}$$

let $\left(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}\right)$ be the solution to the following optimization problem

$$\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \|\boldsymbol{W}\|_1 \qquad \text{subject to} \qquad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 \le \sigma^2. \tag{14}$$

Here, $a_j \in \{1, -1\}$, $\|\boldsymbol{W}\|_1 \coloneqq \sum_{j,k}|w_{jk}|$.

Let ψ be the largest value such that

$$\mathbb{E}\Big[\langle \boldsymbol{a}, \mathrm{relu}(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b})\rangle - \langle \boldsymbol{a}^\star, \mathrm{relu}(\boldsymbol{W}^{\star\top}\boldsymbol{x} + \boldsymbol{b}^\star)\rangle\Big]^2 \ge \psi D_2\left[(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}), (\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)\right]^2. \tag{15}$$

With similar analysis as Dinh & Ho (2020)[Lemma 3.2], one can see that

$$\psi > 0. \tag{16}$$

In addition, we make the following assumptions[1].

Assumption 2. *Suppose that for some constant $B > 0$,*

$$\|\boldsymbol{w}_j^\star\|_2 \le B \qquad \textit{and} \qquad |b_j^\star| \le B \qquad \textit{for all } 1 \le j \le r. \tag{17}$$

Since ψ may depend on model dimensions saliently, we demonstrate that the above assumption can be replaced with the following condition under Gaussian input.

Assumption 3. *Suppose that for some constant $B > 0$,*

$$1 \le \|\boldsymbol{w}_j^\star\|_2 \le B \qquad \textit{and} \qquad |b_j^\star| \le B \qquad \textit{for all } 1 \le j \le r. \tag{18}$$

Here, we consider the normalized setting $\|\boldsymbol{w}_j^\star\|_2 \ge 1$ for simplicity. In addition, we assume that[2]

$$\max_{j\neq k} \frac{\left|\langle \boldsymbol{w}_j^\star, \boldsymbol{w}_k^\star\rangle\right|}{\|\boldsymbol{w}_j^\star\|_2\|\boldsymbol{w}_k^\star\|_2} \le \frac{1}{r^{0.1}}. \tag{19}$$

Then if $\boldsymbol{W}^\star$ has at most s nonzero entries, one can stably reconstruct the neural network stated in the following result when the sample size scales logarithmically with the input dimension. The following theorem is a **more elaborated version of Theorem 1** in the main paper.

Theorem 2. *There exist some universal constants $c_1, c_2, c_3 > 0$, such that for any $\delta > 0$, one has with probability at least $1 - \delta$,*

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star \qquad \textit{and} \qquad \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^\star\boldsymbol{\Pi}^\top\|_{\mathrm{F}} + \|\widehat{\boldsymbol{b}} - \boldsymbol{\Pi}\boldsymbol{b}^\star\|_2 \le c_1\sigma \tag{20}$$

[1] From the technical proofs, it can be seen that the Gaussian input assumption can be replaced with sub-Gaussian input. We consider the Gaussian for simplicity.

[2] Actually, we only need $\max_{j\neq k}\frac{\left|\langle \boldsymbol{w}_j^\star, \boldsymbol{w}_k^\star\rangle\right|}{\|\boldsymbol{w}_j^\star\|_2\|\boldsymbol{w}_k^\star\|_2} \le \frac{1}{r^\omega}$ for some constant $\omega > 0$. Here, we choose $\omega = 0.1$ for ease of understanding.

for some permutation $\mathbf{\Pi} \in \{0,1\}^{r\times r}$*, provided that under the Assumption 2,*

$$n > \frac{c_2}{\psi} s^3 r^3 \log^4 \frac{p}{\delta}, \tag{21}$$

or under the Assumption 3,

$$n > c_2 s^3 r^{13} \log^4 \frac{p}{\delta} \qquad \textit{and} \qquad \sigma < \frac{c_3}{r}. \tag{22}$$

In addition, there exists some $\lambda \in \mathbb{R}$, such that $\left(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}\right)$ is also the solution to the following optimization problem

$$\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 + \lambda\|\boldsymbol{W}\|_1. \tag{23}$$

Specifically, by tuning λ, we can let the optimizer of (23) have the same loss as $\left(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}\right)$, which is exactly the solution to (14).

B Analysis: proof of Theorem 2

Let $\mathcal{S}$ be the index set with cardinality S consisting of the support for $\boldsymbol{W}^\star$ and top entries of $\widehat{\boldsymbol{W}}$. Define

$$\boldsymbol{W} \coloneqq \widehat{\boldsymbol{W}}_{\mathcal{S}} \in \mathbb{R}^{p\times r},$$

and $a_j = \widehat{a}_j$, $b_j = \widehat{b}_j$. Define

$$d_1(\boldsymbol{w}_1, a_1, b_1, \boldsymbol{w}_2, a_2, b_2) = \begin{cases} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 + |b_1 - b_2| & \text{if } a_1 = a_2; \\ \|\boldsymbol{w}_1\|_1 + \|\boldsymbol{w}_2\|_1 + |b_1| + |b_2| & \text{if } a_1 \neq a_2, \end{cases} \tag{24}$$

and

$$d_2(\boldsymbol{w}_1, a_1, b_1, \boldsymbol{w}_2, a_2, b_2) = \begin{cases} \sqrt{\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + |b_1 - b_2|^2} & \text{if } a_1 = a_2; \\ 1 & \text{if } a_1 \neq a_2. \end{cases} \tag{25}$$

In addition, for permutation π on $[r]$, let

$$D_1 \coloneqq \min_{\pi} \sum_{j=1}^{r} d_1(\boldsymbol{w}_{\pi(j)}, a_{\pi(j)}, b_{\pi(j)}, \boldsymbol{w}_j^\star, a_j^\star, b_j^\star), \tag{26a}$$

$$D_2 \coloneqq \min_{\pi} \sqrt{\sum_{j=1}^{r} d_2(\boldsymbol{w}_{\pi(j)}, a_{\pi(j)}, b_{\pi(j)}, \boldsymbol{w}_j^\star, a_j^\star, b_j^\star)^2} \tag{26b}$$

denote the D_1-distance and D_2-distance between $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ and $(\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)$, respectively. Then one has the following bounds.

Lemma 1. *For any* $\boldsymbol{W} \in \mathbb{R}^{p\times r}$ *with* $\|\boldsymbol{W}\|_0 \leq S$*, there exists some universal constants* $c_4, c_5 > 0$ *such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i + b_j^\star)\right]^2 \geq c_4 \psi D_2^2 \tag{27}$$

holds with probability at least $1 - \delta$ *provided that*

$$n \geq c_5 \psi^2 S^3 \log^4 \frac{p}{\delta}. \tag{28}$$

In addition, one has

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i + b_j^\star)\right]^2 \geq c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \tag{29}$$

holds with probability at least $1 - \delta$ *provided that*

$$n \geq c_5 S^3 r^4 \log^4 \frac{p}{\delta}. \tag{30}$$

Lemma 2. *Then there exists some universal constants $c_6 > 0$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i + b_j^\star)\right]^2 \leq c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2 \tag{31}$$

holds with probability at least $1-\delta$.

By comparing the bounds given in Lemma 1 and 2, one has

$$c_4\psi D_2^2 \leq c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2,$$

provided that

$$n > c_5\psi^2 S^3 \log^4\frac{p}{\delta}.$$

Let $\widehat{\mathcal{S}}^\star$ be the index set with cardinality $2s$ consisting of the support for $\boldsymbol{W}^\star$ and top entries of $\widehat{\boldsymbol{W}}$. In addition, let $D_1^\star$ and $D_2^\star$ denote the D_1-distance and D_2-distance between $\left(\widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star}, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}\right)$ and $(\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)$ in a similar way as (26). Notice the fact that

$$D_2^\star \leq D_2 \qquad \text{and} \qquad D_1 \leq 2D_1^\star. \tag{32}$$

Combined with Lemma 3, the above results give

$$D_2^\star \leq \frac{2c_6}{c_4\psi}\sigma,$$

provided that for some constant $c_7 > 0$

$$n \geq c_5\psi^2 S^3 \log^4\frac{p}{\delta} \qquad \text{with } S \geq \frac{c_7 sr}{\psi},$$

such that

$$c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^{\star 2} \leq \frac{c_4\psi}{8} D_2^{\star 2}.$$

Similarly, one has

$$c_4\min\left\{\frac{1}{r}, D_2^2\right\} \leq c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2,$$

provided that

$$n > c_5 S^3 r^4 \log^4\frac{p}{\delta}.$$

Combined with Lemma 3, the above results give

$$D_2^\star \leq \frac{2c_6}{c_4}\sigma,$$

provided that for some constant $c_7 > 0$

$$n \geq c_5 S^3 r^4 \log^4\frac{p}{\delta} \qquad \text{and} \qquad \sigma^2 \leq \frac{c_4}{2c_6 r} \qquad \text{with } S \geq c_7 sr^3,$$

such that

$$c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2 < \frac{c_4}{r} \qquad \text{and} \qquad c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^{\star 2} \leq \frac{c_4}{8} D_2^{\star 2}.$$

Then we conclude the proof since after appropriate permutation

$$\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^\star\|_{\mathrm{F}} \leq 2\|\widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star} - \boldsymbol{W}^\star\|_{\mathrm{F}}.$$

C PROOF OF LEMMA 1 (LOWER BOUND)

This can be seen from the following three properties.

- Consider the case that $D_1 \leq \epsilon = \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}$. With probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2$$
$$= \frac{D_1^2}{\epsilon^2}\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{x}_i + \widetilde{b}_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2, \tag{33}$$

  where $\widetilde{\boldsymbol{w}}_j = \boldsymbol{w}_j^\star + \frac{\epsilon}{D_1}\left(\boldsymbol{w}_j - \boldsymbol{w}_j^\star\right)$ and $\widetilde{b}_j = b_j^\star + \frac{\epsilon}{D_1}\left(b_j - b_j^\star\right)$.

- For any $\epsilon > 0$ and
$$D_1 \geq \frac{\epsilon}{\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}},$$
  there exists some universal constant $C_1 > 0$, such that with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2$$
$$\geq \mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star)\right]^2$$
$$- C_1 D_1^2 \log\frac{pn}{\delta}\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}. \tag{34}$$

- For some universal constant $C_2 > 0$

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star)\right]^2 \geq C_2 \min\left\{\frac{1}{r}, D_2^2\right\}. \tag{35}$$

Putting all together. Let

$$\epsilon = C_3\frac{\delta}{nr}\sqrt{\frac{S}{n}\log\frac{BnS}{\delta}},$$

for some universal constant $C_3 > 0$ such that

$$\frac{\epsilon}{\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}} < \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

Inserting (15) into (34) gives

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2$$
$$\geq \psi D_2^2 - C_1 D_1^2 \log\frac{pn}{\delta}\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}} \geq \frac{\psi}{2}D_2^2, \tag{36}$$

holds with probability at least $1-\delta$ provided that for some constant $C_4 > 0$

$$n \geq C_4\psi^2 S^3 \log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}\log^2\frac{pn}{\delta} \qquad \text{and} \qquad D_1 \geq \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

Here, the last line holds due to Lemma 3 and we assume that $\max\{\|\boldsymbol{W}\|_\infty, \|\boldsymbol{b}\|_\infty\}$ is bounded by some constant. On the other hand, if $D_1 < \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}$, (33) and (36) tell us that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \geq \frac{D_1^2}{\epsilon^2}\frac{\psi}{2}\widetilde{D}_2^2 = \frac{\psi}{2}D_2^2, \tag{37}$$

where $\widetilde{D}_2$ denotes the D_2-distance between $\left(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}}\right)$ and $(\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)$ in a similar way as (26).

Inserting (35) into (34) gives

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \\
&\geq C_2 \min\left\{\frac{1}{r}, D_2^2\right\} - C_1 D_1^2 \log\frac{pn}{\delta}\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}} \\
&\geq \frac{C_2}{2}\min\left\{\frac{1}{r}, D_2^2\right\},
\end{aligned} \tag{38}$$

holds with probability at least $1-\delta$ provided that for some constant $C_4 > 0$

$$n \geq C_4 S^3 r^4 \log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}\log^2\frac{pn}{\delta} \qquad \text{and} \qquad D_1 \geq \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

Here, the last line holds due to Lemma 3 and we assume that $\max\{\|\boldsymbol{W}\|_\infty, \|\boldsymbol{b}\|_\infty\}$ is bounded by some constant. On the other hand, if $D_1 < \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}$, (33) and (38) tell us that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \geq \frac{D_1^2}{\epsilon^2}\frac{C_2}{2}\min\left\{\frac{1}{r}, \widetilde{D}_2^2\right\} = \frac{C_2}{2}D_2^2. \tag{39}$$

Summing up, we conclude the proof by verifying the claims in the following.

C.1 Proof of (33)

Without loss of generality, we assume that $a_j = a_j^\star$ for $1 \leq j \leq r$, and

$$D_1 = \sum_{j=1}^{r}\left(\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_1 + |b_j - b_j^\star|\right) \leq \epsilon.$$

By taking union bound, with probability at least $1 - \frac{\delta}{2}$, one has for all $1 \leq i \leq n$ and $1 \leq j \leq r$,

$$\left|\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star\right| > \frac{\delta}{2nr}\sqrt{\frac{\pi}{2}},$$

since $\|\boldsymbol{w}_j^\star\|_2 \geq 1$ and $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. In addition, for all $1 \leq i \leq n$ and $1 \leq j \leq r$,

$$\left|\boldsymbol{w}_j^\top\boldsymbol{x}_i + b_j - \boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i - b_j^\star\right| \leq \|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_1\|\boldsymbol{x}_i\|_\infty + |b_j - b_j^\star| \leq \epsilon\sqrt{2\log\frac{4pn}{\delta}}$$

holds with probability at least $1 - \frac{\delta}{2}$. Here, the last inequality comes from the fact that with probability at least $1 - \frac{\delta}{2}$,

$$\|\boldsymbol{x}_i\|_\infty \leq \sqrt{2\log\frac{4pn}{\delta}} \qquad \text{for all } 1 \leq i \leq n. \tag{40}$$

Putting together, we have with probability at least $1-\delta$,

$$u(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) = u(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i + b_j^\star), \tag{41}$$

with the proviso that $\epsilon \leq \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}$. Note that $u(x)=1$ if $x>0$, and $u(x)=0$ if $x\leq 0$. Then combining with the definition of $\widetilde{\boldsymbol{w}}_j$ and $\widetilde{b}_j$, the above property yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left[\sum_{j=1}^r a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^r a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left[\sum_{j=1}^r a_j^\star u(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j - \boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i - b_j^\star)\right]^2 \\
&= \frac{D_1^2}{\epsilon^2}\frac{1}{n}\sum_{i=1}^n \left[\sum_{j=1}^r a_j^\star u(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)(\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{x}_i + \widetilde{b}_j - \boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i - b_j^\star)\right]^2 \\
&= \frac{D_1^2}{\epsilon^2}\frac{1}{n}\sum_{i=1}^n \left[\sum_{j=1}^r a_j \mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{x}_i + \widetilde{b}_j) - \sum_{j=1}^r a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2,
\end{aligned}$$

and the claim is proved.

C.2 Proof of (34)

Notice that

$$\begin{aligned}
&\left|a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right| \\
&\leq \begin{cases} \|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_1 \|\boldsymbol{x}\|_\infty + |b_j - b_j^\star| & \text{if } a_j = a_j^\star, \\ \left(\|\boldsymbol{w}_j\|_1 + \|\boldsymbol{w}_j^\star\|_1\right)\|\boldsymbol{x}\|_\infty + |b_j| + |b_j^\star| & \text{if } a_j \neq a_j^\star, \end{cases}
\end{aligned}$$

which leads to

$$\left|\sum_{j=1}^r a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^r a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star)\right| \leq D_1 \max\left\{\|\boldsymbol{x}\|_\infty, 1\right\}. \tag{42}$$

For any fixed $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$, let

$$z_i \coloneqq \sum_{j=1}^r a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^r a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star),$$

and define the following event set

$$\mathcal{E} \coloneqq \left\{\|\boldsymbol{x}_i\|_\infty \leq \sqrt{2\log\frac{4pn}{\delta}} \qquad \text{for all } 1\leq i\leq n\right\}.$$

Then with probability at least $1-\delta$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \left(z_i^2 - \mathbb{E}\left[z_i^2\right]\right) &= \frac{1}{n}\sum_{i=1}^n \left\{z_i^2 \mathbb{1}(\mathcal{E}) - \mathbb{E}\left[z_i^2 \mathbb{1}(\mathcal{E})\right] - \mathbb{E}\left[z_i^2 \mathbb{1}(\overline{\mathcal{E}})\right]\right\} \\
&\geq -4D_1^2 \log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2}{\delta}} - D_1^2\frac{\delta}{n} \\
&\geq -5D_1^2 \log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2}{\delta}}.
\end{aligned} \tag{43}$$

Here, the first line holds due to (40); the last line comes from Hoeffding's inequality, and the fact that

$$\begin{aligned}\left|\mathbb{E}\left[z_i^2\mathbb{1}(\overline{\mathcal{E}})\right]\right| &\le D_1^2\left|\mathbb{E}\left[\|\boldsymbol{x}_i\|_\infty^2\mathbb{1}(\|\boldsymbol{x}_i\|_\infty > \sqrt{2\log\frac{4pn}{\delta}})\right]\right| \\ &\le D_1^2\int_{\sqrt{2\log\frac{4pn}{\delta}}}^{\infty} x^2 d\mathbb{P}(\|\boldsymbol{x}_i\|_\infty < x) \\ &\le D_1^2\int_{\sqrt{2\log\frac{4pn}{\delta}}}^{\infty} 4xp\exp(-\frac{x^2}{2})dx \le D_1^2\frac{\delta}{n}.\end{aligned}$$

In addition, consider the following ϵ-net

$$\begin{aligned}\mathcal{N}_\epsilon \coloneqq \Bigg\{(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}) : |W_{ij}| \in \frac{\epsilon}{r+S}\left[\lceil\frac{B(r+S)}{\epsilon}\rceil\right], \|\boldsymbol{W}\|_0 \le S, \\ |b_j| \in \frac{\epsilon}{r+S}\left[\lceil\frac{B(r+S)}{\epsilon}\rceil\right], |a_j| = 1\Bigg\},\end{aligned}$$

where $[n] \coloneqq \{1,2,\ldots,n-1\}$. Then for all $(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b})$ with $\|\boldsymbol{W}\|_1 \le B$ and $\|\boldsymbol{b}\|_1 \le B$, there exists some point, denoted by $\left(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{a}},\widetilde{\boldsymbol{b}}\right)$, in $\mathcal{N}_\epsilon$ whose D_1-distance from $(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b})$ is less than ϵ. For simplicity, define

$$\begin{aligned}z_i &\coloneqq \sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}_i + b_j) - \sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star), \\ \widetilde{z}_i &\coloneqq \sum_{j=1}^r \widetilde{a}_j\mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top\boldsymbol{x}_i + \widetilde{b}_j) - \sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star).\end{aligned}$$

Similar to (42), we can derive that

$$\left|\sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x} + b_j) - \sum_{j=1}^r \widetilde{a}_j\mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top\boldsymbol{x} + \widetilde{b}_j)\right| \le \epsilon\max\left\{\|\boldsymbol{x}\|_\infty, 1\right\},$$

which implies

$$\left|z_i^2 - \widetilde{z}_i^2\right| \le \epsilon\left(\epsilon + D_1\right)\max\left\{\|\boldsymbol{x}_i\|_\infty^2, 1\right\},$$

and then with probability at least $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n\left(z_i^2 - \mathbb{E}\left[z_i^2\right]\right) - \frac{1}{n}\sum_{i=1}^n\left(\widetilde{z}_i^2 - \mathbb{E}\left[\widetilde{z}_i^2\right]\right) \ge -4\epsilon\left(\epsilon + D_1\right)\log\frac{4pn}{\delta}. \tag{44}$$

In addition, a little algebra gives

$$\log|\mathcal{N}_\epsilon| \le C_5 S\log\frac{pr}{S}\log\frac{BS}{\epsilon}, \tag{45}$$

for some universal constant $C_5 > 0$. Combining (43), (44), and (45) leads to

$$\frac{1}{n}\sum_{i=1}^n\left(z_i^2 - \mathbb{E}\left[z_i^2\right]\right) \ge -5\left(\epsilon + D_1\right)^2\log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2\left|\mathcal{N}_\epsilon\right|}{\delta}} - 4\epsilon\left(\epsilon + D_1\right)\log\frac{4pn}{\delta}.$$

Then, (34) is obvious.

C.3 Proof of (35)

We first consider a simple case that $b_j = 0$ and $b_j^\star = 0$ for $1 \le j \le r$, and show that for some small constant $C_6 > 0$,

$$\mathbb{E}\left[\sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}) - \sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\right]^2 \ge C_6\min\left\{\frac{1}{r}, D_2^2\right\}. \tag{46}$$

In the following, we will focus on the case

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x})\right]^2 \leq \frac{C_6}{r}.$$

According to Lemma 4, one has for any constant $k \geq 0$, there exists some constant $\alpha_k > 0$ such that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x})\right]^2 \\
&\geq \alpha_k \left\| \sum_{j=1}^{r} a_j \|\boldsymbol{w}_j\|_2 \big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\big)^{\otimes 2k} - \sum_{j=1}^{r} a_j^\star \|\boldsymbol{w}_j^\star\|_2 \big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes 2k} \right\|_F^2. \qquad (47)
\end{aligned}$$

Assumption 1 tells us that for any integer $k \geq \frac{2}{\omega}$,

$$\left|\langle \boldsymbol{v}_{j_1}^\star, \boldsymbol{v}_{j_2}^\star \rangle\right| \leq \frac{1}{r^2}. \qquad (48)$$

where

$$\boldsymbol{v}_j \coloneqq \mathrm{vec}\left(\big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\big)^{\otimes k}\right) \qquad \text{with} \qquad \beta_j \coloneqq a_j \|\boldsymbol{w}_j\|_2,$$

and

$$\boldsymbol{v}_j^\star \coloneqq \mathrm{vec}\left(\big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes k}\right) \qquad \text{with} \qquad \beta_j^\star \coloneqq a_j^\star \|\boldsymbol{w}_j^\star\|_2.$$

Then (47) gives

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x})\right]^2 \geq \alpha_{3k} \left\| \sum_{j=1}^{r} \beta_j \boldsymbol{v}_j^{\otimes 6} - \sum_{j=1}^{r} \beta_j^\star \boldsymbol{v}_j^{\star\otimes 6} \right\|_F^2.$$

Define

$$\mathbb{S}_+ \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j\right\}_{j:\beta_j>0} \qquad \mathbb{S}_- \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j\right\}_{j:\beta_j<0},$$

and

$$\mathbb{S}_+^\star \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j^\star\right\}_{j:\beta_j^\star>0} \qquad \mathbb{S}_-^\star \coloneqq \mathrm{span}\left\{\boldsymbol{v}_j^\star\right\}_{j:\beta_j^\star<0}.$$

Let $\boldsymbol{P}_{\mathbb{S}}$ and $\boldsymbol{P}_{\mathbb{S}}^\perp$ denote the projection onto and perpendicular to the subspace $\mathbb{S}$, respectively. By noticing that $\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j = \boldsymbol{0}$ for j obeying $\beta_j < 0$, and $\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp \boldsymbol{v}_j^\star = \boldsymbol{0}$ for j obeying $\beta_j^\star > 0$, one has

$$\begin{aligned}
&\left\| \sum_{j=1}^{r} \beta_j \boldsymbol{v}_j^{\otimes 6} - \sum_{j=1}^{r} \beta_j^\star \boldsymbol{v}_j^{\star\otimes 6} \right\|_F^2 \\
&\geq \left\| \sum_{j:\beta_j>0} \beta_j \big(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j\big)^{\otimes 2} \otimes \big(\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp \boldsymbol{v}_j\big)^{\otimes 4} - \sum_{j:\beta_j^\star<0} \beta_j^\star \big(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j^\star\big)^{\otimes 2} \otimes \big(\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp \boldsymbol{v}_j^\star\big)^{\otimes 4} \right\|_F^2 \\
&\geq \sum_{j:\beta_j^\star<0} \left\| \beta_j^\star \big(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j^\star\big)^{\otimes 2} \otimes \big(\boldsymbol{P}_{\mathbb{S}_+^\star}^\perp \boldsymbol{v}_j^\star\big)^{\otimes 4} \right\|_F^2 \geq \frac{1}{2} \sum_{j:\beta_j^\star<0} \left\| \boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j^\star \right\|_2^4,
\end{aligned}$$

where the penultimate inequality holds since the inner product between every pair of terms is positive, and the last inequality comes from the facts that $|\beta_j^\star| \geq 1$ and (48).

Moreover, by means of AM-GM inequality and (48), one can see that

$$\sum_{j:\beta_j^\star<0} \left\| \boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j^\star \right\|_2^4 \geq \frac{1}{r}\Big(\sum_{j:\beta_j^\star<0} \left\| \boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j^\star \right\|_2^2 \Big)^2 = \frac{1}{r} \left\| \boldsymbol{P}_{\mathbb{S}_-}^\perp \left[\boldsymbol{v}_j^\star\right]_{j:\beta_j^\star<0} \right\|_F^4 \geq \frac{1}{2r} \left\| \boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{P}_{\mathbb{S}_-^\star} \right\|_F^4.$$

Then combining with (46), the above result and the counterpart for $\beta_j^\star > 0$ lead to

$$\dim(\mathbb{S}_-) \geq \dim(\mathbb{S}_-^\star) \qquad \text{and} \qquad \dim(\mathbb{S}_+) \geq \dim(\mathbb{S}_+^\star),$$

which gives

$$\dim(\mathbb{S}_-) = \dim(\mathbb{S}_-^\star) \qquad \text{and} \qquad \dim(\mathbb{S}_+) = \dim(\mathbb{S}_+^\star).$$

Furthermore, for some small constant $C_6 > 0$, we have

$$\operatorname{dist}(\mathbb{S}_-, \mathbb{S}_-^\star) \leq C_6 \qquad \text{and} \qquad \operatorname{dist}(\mathbb{S}_+, \mathbb{S}_+^\star) \leq C_6.$$

Let $\boldsymbol{P}_i^\perp$ denote the projection perpendicular to

$$\operatorname{span}\left\{\boldsymbol{v}_j^\star\right\}_{j\neq i:\beta_j^\star>0},$$

and

$$\gamma_j := \frac{\beta_j \langle \boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j, P_{\mathbb{S}_-}^\perp \boldsymbol{v}_i^\star\rangle^2 \langle \boldsymbol{P}_i^\perp \boldsymbol{v}_i, P_{\mathbb{S}_-}^\perp \boldsymbol{v}_i^\star\rangle^2}{\left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_i^\star\right\|_2^2 \left\|\boldsymbol{P}_i^\perp \boldsymbol{v}_i^\star\right\|^2}.$$

Then for any i,

$$\begin{aligned}
&\left\|\sum_{j=1}^r \beta_j \boldsymbol{v}_j^{\otimes 6} - \sum_{j=1}^r \beta_j^\star \boldsymbol{v}_j^{\star\otimes 6}\right\|_F^2 \geq \left\|\sum_{j:\beta_j>0} \beta_j \left(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j\right)^{\otimes 2} \otimes \boldsymbol{v}_j^{\otimes 4} - \sum_{j=1}^r \beta_j^\star \left(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j^\star\right)^{\otimes 2} \otimes \boldsymbol{v}_j^{\star\otimes 4}\right\|_F^2 \\
&\geq \frac{1}{2}\left\|\sum_{j:\beta_j>0} \beta_j \left(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j\right)^{\otimes 2} \otimes \boldsymbol{v}_j^{\otimes 4} - \sum_{j:\beta_j^\star>0} \beta_j^\star \left(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j^\star\right)^{\otimes 2} \otimes \boldsymbol{v}_j^{\star\otimes 4}\right\|_F^2 \\
&\geq \frac{1}{2}\left\|\sum_{j:\beta_j>0} \beta_j \left(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_j\right)^{\otimes 2} \otimes \left(\boldsymbol{P}_i^\perp \boldsymbol{v}_i\right)^{\otimes 2} \otimes \boldsymbol{v}_j^{\otimes 2} - \beta_i^\star \left(\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_i^\star\right)^{\otimes 2} \otimes \left(\boldsymbol{P}_i^\perp \boldsymbol{v}_i^\star\right)^{\otimes 2} \otimes \boldsymbol{v}_i^{\star\otimes 2}\right\|_F^2 \\
&\geq \frac{1}{2}\left\|\sum_{j:\beta_j>0} \gamma_j \boldsymbol{v}_j^{\otimes 2} - \beta_i^\star \left\|\boldsymbol{P}_{\mathbb{S}_-}^\perp \boldsymbol{v}_i^\star\right\|_2^2 \left\|\boldsymbol{P}_i^\perp \boldsymbol{v}_i^\star\right\|^2 \boldsymbol{v}_i^{\star\otimes 2}\right\|_F^2,
\end{aligned}$$

which, together with (46), implies that there exists some j such that

$$\|\sqrt{\beta_j}\boldsymbol{v}_j - \sqrt{\beta_i^\star}\boldsymbol{v}_i^\star\|_2^2 \leq \frac{1}{r}.$$

Without loss of generality, assume that

$$\|\sqrt{\beta_j}\boldsymbol{v}_j - \sqrt{\beta_j^\star}\boldsymbol{v}_j^\star\|_2^2 \leq \frac{1}{r}, \qquad \text{for all } 1 \leq j \leq r. \tag{49}$$

Then

$$\begin{aligned}
&\mathbb{E}\big[\sum_{j=1}^r a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^r a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x})\big]^2 \geq \alpha_k \left\|\sum_{j=1}^r \beta_j \boldsymbol{v}_j \boldsymbol{v}_j^\top - \sum_{j=1}^r \beta_j^\star \boldsymbol{v}_j^\star \boldsymbol{v}_j^{\star\top}\right\|_F^2 \\
&\geq \alpha_k \sum_{j=1}^r \left\|\beta_j \boldsymbol{v}_j \boldsymbol{v}_j^\top - \beta_j^\star \boldsymbol{v}_j^\star \boldsymbol{v}_j^{\star\top}\right\|_F^2 - \frac{\alpha_k}{2r}\left(\sum_{j=1}^r \left\|\beta_j \boldsymbol{v}_j \boldsymbol{v}_j^\top - \beta_j^\star \boldsymbol{v}_j^\star \boldsymbol{v}_j^{\star\top}\right\|_F\right)^2 \\
&\geq \frac{\alpha_k}{2} \sum_{j=1}^r \left\|\beta_j \boldsymbol{v}_j \boldsymbol{v}_j^\top - \beta_j^\star \boldsymbol{v}_j^\star \boldsymbol{v}_j^{\star\top}\right\|_F^2.
\end{aligned}$$

Here, the first line comes from (47); the second line holds through the following claim

$$\begin{aligned}
&\left|\langle \beta_{j_1} \boldsymbol{v}_{j_1} \boldsymbol{v}_{j_1}^\top - \beta_{j_1}^\star \boldsymbol{v}_{j_1}^\star \boldsymbol{v}_{j_1}^{\star\top}, \beta_{j_2} \boldsymbol{v}_{j_2} \boldsymbol{v}_{j_2}^\top - \beta_{j_2}^\star \boldsymbol{v}_{j_2}^\star \boldsymbol{v}_{j_2}^{\star\top}\rangle\right| \\
&\leq \frac{1}{2r}\|\beta_{j_1} \boldsymbol{v}_{j_1} \boldsymbol{v}_{j_1}^\top - \beta_{j_1}^\star \boldsymbol{v}_{j_1}^\star \boldsymbol{v}_{j_1}^{\star\top}\|_2 \|\beta_{j_2} \boldsymbol{v}_{j_2} \boldsymbol{v}_{j_2}^\top - \beta_{j_2}^\star \boldsymbol{v}_{j_2}^\star \boldsymbol{v}_{j_2}^{\star\top}\|_2
\end{aligned}$$

since for $\boldsymbol{\delta}_j \coloneqq \sqrt{\beta_j}\boldsymbol{v}_j - \sqrt{\beta_j^\star}\boldsymbol{v}_j^\star$,

$$\beta_j \boldsymbol{v}_j \boldsymbol{v}_j^\top - \beta_j^\star \boldsymbol{v}_j^\star \boldsymbol{v}_j^{\star\top} = \boldsymbol{\delta}_j \boldsymbol{\delta}_j^\top + \sqrt{\beta_j^\star}\boldsymbol{\delta}_j \boldsymbol{v}_j^{\star\top} + \sqrt{\beta_j^\star}\boldsymbol{v}_j^\star \boldsymbol{\delta}_j^{\star\top}.$$

Then the conclusion is obvious by noticing that

$$\left\|\beta_j \boldsymbol{v}_j \boldsymbol{v}_j^\top - \beta_j^\star \boldsymbol{v}_j^\star \boldsymbol{v}_j^{\star\top}\right\|_F \geq \|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_2.$$

Finally, we analyze the general case with $b_j, b_j^\star \neq 0$, which is similar to the above argument. For simplicity, we only explain the different parts here. According to Lemma 4, one has for any constant $k \geq 0$, there exists some constant $\alpha_k > 0$ and some function $f_k : \mathbb{R} \to \mathbb{R}$ such that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^r a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^r a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star)\right]^2 \\
&\geq \sum_{k \geq \frac{12}{\omega}}^\infty \left\| \sum_{j=1}^r a_j f_k(\frac{b_j}{\|\boldsymbol{w}_j\|_2})\|\boldsymbol{w}_j\|_2 \big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\big)^{\otimes k} - \sum_{j=1}^r a_j^\star f_k(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2})\|\boldsymbol{w}_j^\star\|_2 \big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes k}\right\|_F^2 \\
&\gtrsim \sum_{j=1}^r \sum_{k \geq \frac{12}{\omega}}^\infty \left\| a_j f_k(\frac{b_j}{\|\boldsymbol{w}_j\|_2})\boldsymbol{w}_j - a_j^\star f_k(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2})\boldsymbol{w}_j^\star\right\|_F^2 \\
&\gtrsim \sum_{j=1}^r \inf_{R_l(\boldsymbol{x})} \mathbb{E}\left[a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star) - R_l(\boldsymbol{x})\right]^2 \\
&\gtrsim \sum_{j=1}^r \left(\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_2^2 + |b_j - b_j^\star|^2\right).
\end{aligned}$$

Here, $l = \left[\frac{12}{\omega}\right]$, and the second inequality holds in a similar way to above analysis. Then the general conclusion is handy.

D Proof of Lemma 2 (upper bound)

For simplicity, let

$$\begin{aligned}
z_i &\coloneqq \sum_{j=1}^r a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^r a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star), \\
\widehat{z}_i &\coloneqq \sum_{j=1}^r a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^r \widehat{a}_j \mathrm{relu}(\widehat{\boldsymbol{w}}_j^\top \boldsymbol{x}_i + \widehat{b}_j).
\end{aligned}$$

Recall the optimality of $\left(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}\right)$ w.r.t. (4). According to the triangle inequality, one has

$$\sqrt{\frac{1}{n}\sum_{i=1}^n z_i^2} \leq \sqrt{\frac{1}{n}\sum_{i=1}^n \widehat{z}_i^2} + 2\sigma. \tag{50}$$

We can bound the first term in the right hand side by

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \widehat{z}_i^2 &= \frac{1}{n}\sum_{i=1}^n \left[\sum_{j=1}^r a_j \left(\mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \mathrm{relu}(\widehat{\boldsymbol{w}}_j^\top \boldsymbol{x}_i + \widehat{b}_j)\right)\right]^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n \left[\sum_{j=1}^r \left|(\boldsymbol{w}_j - \widehat{\boldsymbol{w}}_j)^\top \boldsymbol{x}_i\right|\right]^2 \\
&\leq \frac{r}{n}\sum_{i=1}^n \sum_{j=1}^r \left|(\boldsymbol{w}_j - \widehat{\boldsymbol{w}}_j)^\top \boldsymbol{x}_i\right|^2,
\end{aligned}$$

where the second line holds due to the contraction property of ReLu function, and the last line comes from the AM-GM inequality. Lemma 5 further gives for some constant $C_7 > 0$,

$$\sum_{j=1}^{r} \frac{1}{n} \sum_{i=1}^{n} \left|(\boldsymbol{w}_j - \widehat{\boldsymbol{w}}_j)^\top \boldsymbol{x}_i\right|^2 \leq C_7 \sum_{j=1}^{r} \|\boldsymbol{w}_j - \widehat{\boldsymbol{w}}_j\|_2^2 + C_7 \frac{\log^3 \frac{p}{n\delta}}{n} \sum_{j=1}^{r} \|\boldsymbol{w}_j - \widehat{\boldsymbol{w}}_j\|_1^2$$

holds with probability at least $1 - \delta$. In addition,

$$\sum_{j=1}^{r} \|\boldsymbol{w}_j - \widehat{\boldsymbol{w}}_j\|_1^2 \leq \left\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\right\|_1^2 \leq \left(\|\boldsymbol{W}^\star\|_1 - \|\widehat{\boldsymbol{W}}\|_1\right)^2 \leq D_1^2,$$

and

$$\begin{aligned}\sum_{j=1}^{r} \|\boldsymbol{w}_j - \widehat{\boldsymbol{w}}_j\|_2^2 = \left\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\right\|_1^2 &\leq \left\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\right\|_1 \left\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\right\|_\infty \\ &\leq \frac{\left(\|\boldsymbol{W}^\star\|_1 - \|\widehat{\boldsymbol{W}}\|_1\right)\left(\|\boldsymbol{W}^\star\|_1 - \|\widehat{\boldsymbol{W}}^\star\|_1\right)}{S/2} \leq \frac{4}{S} D_1^2.\end{aligned}$$

Here, $\widehat{\boldsymbol{W}}^\star$ denote the entries of $\widehat{\boldsymbol{W}}$ on the support set for $\boldsymbol{W}^\star$, and we make use of the fact that $\|\widehat{\boldsymbol{W}}\|_1 \leq \|\boldsymbol{W}^\star\|_1$ and

$$\left\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\right\|_\infty \leq \frac{\|\widehat{\boldsymbol{W}}^\star - \widehat{\boldsymbol{W}}\|_1}{S - s} \leq \frac{\|\boldsymbol{W}^\star\|_1 - \|\widehat{\boldsymbol{W}}^\star\|_1}{S/2}.$$

Putting everything together gives the desired result.

E TECHNICAL LEMMAS

Lemma 3. *For any $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ with $\|\boldsymbol{W}\|_0 + \|\boldsymbol{b}\|_0 + \|\boldsymbol{W}^\star\|_0 + \|\boldsymbol{b}^\star\|_0 \leq S$. Assume that $\|\boldsymbol{W}\|_1 + \|\boldsymbol{b}\|_1 \leq \|\boldsymbol{W}^\star\|_1 + \|\boldsymbol{b}^\star\|_1$ and $\|\boldsymbol{w}_j^\star\|_2^2 + |b_j^\star|^2 \leq 1$. Then one has*

$$D_1 \leq 2\sqrt{S} D_2, \tag{51}$$

where D_1, D_2 are defined in (26)*.*

Proof. For simplicity, assume that

$$D_2^2 = \sum_{j \in \mathcal{J}} \left(\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_2^2 + |b_j - b_j^\star|^2\right) + \sum_{j \notin \mathcal{J}} \left(\|\boldsymbol{w}_j^\star\|_2^2 + |b_j^\star|^2\right).$$

Here, $j \in \mathcal{J}$ means that $a_j = a_j^\star$ and

$$\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_2^2 + |b_j - b_j^\star|^2 \leq \|\boldsymbol{w}_j^\star\|_2^2 + |b_j^\star|^2.$$

Then according to the AM-GM inequality, one has

$$\begin{aligned}\sqrt{S} D_2 &\geq \sum_{j \in \mathcal{J}} \left(\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_1 + |b_j - b_j^\star|\right) + \sum_{j \notin \mathcal{J}} \left(\|\boldsymbol{w}_j^\star\|_1 + |b_j^\star|\right) \\ &\geq \sum_{j \in \mathcal{J}} \left(\|\boldsymbol{w}_j^\star\|_1 - \|\boldsymbol{w}_j\|_1 + |b_j^\star| - |b_j|\right) + \|\boldsymbol{W}^\star\|_1 + \|\boldsymbol{b}^\star\|_1 - \sum_{j \in \mathcal{J}} \left(\|\boldsymbol{w}_j^\star\|_1 + |b_j^\star|\right) \\ &\geq \sum_{j \notin \mathcal{J}} \left(\|\boldsymbol{w}_j\|_1 + |b_j|\right),\end{aligned}$$

which implies

$$2\sqrt{S} D_2 \geq \sum_{j \in \mathcal{J}} \left(\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_1 + |b_j - b_j^\star|\right) + \sum_{j \notin \mathcal{J}} \left(\|\boldsymbol{w}_j^\star\|_1 + |b_j^\star| + \|\boldsymbol{w}_j\|_1 + |b_j|\right).$$

Thus we conclude the proof. □

Lemma 4. *For any constant $k \geq 0$, there exists some universal function $f_k : \mathbb{R} \to \mathbb{R}$ such that*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x} + b_j^\star)\right]^2 \\
&= \sum_{k=0}^{\infty} \Big\| \sum_{j=1}^{r} a_j f_k(\frac{b_j}{\|\boldsymbol{w}_j\|_2}) \|\boldsymbol{w}_j\|_2 \Big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\Big)^{\otimes k} \\
&\qquad - \sum_{j=1}^{r} a_j^\star f_k(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2}) \|\boldsymbol{w}_j^\star\|_2 \big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes k} \Big\|_F^2,
\end{aligned} \tag{52}$$

with

$$\alpha_k \coloneqq f_{2k}(0) > 0, \qquad \textit{for all } k > 0. \tag{53}$$

In addition, we have

$$\begin{aligned}
&\inf_{R_l(\boldsymbol{x})} \mathbb{E}\left[a\mathrm{relu}(\boldsymbol{w}^\top \boldsymbol{x} + b) - \sum_{j=1}^{r} a^\star \mathrm{relu}(\boldsymbol{w}^{\star\top} \boldsymbol{x} + b^\star) - R_l(\boldsymbol{x})\right]^2 \\
&= \sum_{k>l}^{\infty} \Big\| a f_k(\frac{b}{\|\boldsymbol{w}\|_2}) \|\boldsymbol{w}\|_2 \big(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\big)^{\otimes k} - a^\star f_k(\frac{b^\star}{\|\boldsymbol{w}^\star\|_2}) \|\boldsymbol{w}^\star\|_2 \big(\frac{\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2}\big)^{\otimes k} \Big\|_F^2,
\end{aligned} \tag{54}$$

where $R_l(\boldsymbol{x})$ denote the polynomial with order less than l.

Lemma 5. *There exists some universal constant $c > 0$, such that for all $\boldsymbol{w} \in \mathbb{R}^p$,*

$$\frac{1}{n}\sum_{i=1}^{n} \left|\boldsymbol{w}^\top \boldsymbol{x}_i\right|^2 \leq c\,\|\boldsymbol{w}\|_2^2 + c\frac{\log^3 \frac{p}{n\delta}}{n}\,\|\boldsymbol{w}\|_1^2, \tag{55}$$

holds with probability at least $1 - \delta$.

Proof. Before proceeding, we introduce some useful techniques about Restricted Isometry Property (RIP). Let $\boldsymbol{X} \coloneqq \frac{1}{\sqrt{n}}[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$. For some constant $c_0 > 0$, if $n \geq c_0\left(s \log \frac{p}{s} + \log \frac{1}{\delta}\right)$, then with probability at least $1 - \delta$,

$$\left\|\boldsymbol{X}^\top \boldsymbol{w}\right\|_2^2 \leq 2\|\boldsymbol{w}\|_2^2 \tag{56}$$

holds for all $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\|_0 \leq s$.

We divide the entries of $\boldsymbol{w}$ into several groups $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots \cup \mathcal{S}_L$ with equal size s (except for $\mathcal{S}_L$), such that the entries in $\mathcal{S}_j$ are no less than $\mathcal{S}_k$ for any $j < k$. Then, according (56), one has

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{w}^\top \boldsymbol{x}_i)^2 = \boldsymbol{w}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w} &= \sum_{j,k} \boldsymbol{w}_{\mathcal{S}_j}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w}_{\mathcal{S}_k} \\
&\leq 2\sum_{j,k} \|\boldsymbol{w}_{\mathcal{S}_j}\|_2 \|\boldsymbol{w}_{\mathcal{S}_k}\|_2 = 2\Big(\sum_{l=1}^{L} \|\boldsymbol{w}_{\mathcal{S}_l}\|_2\Big)^2.
\end{aligned}$$

In addition, the order of $\boldsymbol{w}_{\mathcal{S}_l}$ yields for $l > 1$,

$$\|\boldsymbol{w}_{\mathcal{S}_l}\|_2 \leq \sqrt{s}\|\boldsymbol{w}_{\mathcal{S}_l}\|_\infty \leq \frac{1}{(l-1)\sqrt{s}}\|\boldsymbol{w}\|_1,$$

which leads to

$$\Big(\sum_{l=1}^{L} \|\boldsymbol{w}_{\mathcal{S}_l}\|_2\Big)^2 \leq 2\|\boldsymbol{w}_{\mathcal{S}_1}\|_2^2 + 2\Big(\sum_{l=2}^{L} \frac{1}{(l-1)\sqrt{s}}\|\boldsymbol{w}\|_1\Big)^2 \leq 2\|\boldsymbol{w}\|_2^2 + \frac{2\log^2 L}{s}\|\boldsymbol{w}\|_1^2.$$

Then the result is obvious by taking above relations together. □