Provable Identifiability of Two-Layer ReLU Neural Networks via LASSO Regularization

Gen Li, Ganghua Wang, and Jie Ding, *Member, IEEE*

***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 two-layer ReLU neural networks, a fashionable and powerful nonlinear regression model. Specifically, given a neural network whose output y depends only on a small subset of input 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 advocate a neural network-based variable selection method. Experiments on simulated and real-world datasets show promising performance of the variable selection approach compared with existing techniques.**

***Index Terms*— Lasso, identifiability, neural network, nonlinear regression, variable selection.**

I. 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) [1] 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 [2], proportional hazards models [3], and neural networks [4]. For LASSO-regularized neural networks, existing works have studied different properties, such as convergence of training [5], model pruning [6], [7], and feature selection [8], [9]. The LASSO regularization has also been added into the standard deep learning toolbox of many open-source libraries, e.g., Tensorflow [10] and Pytorch [11].

Despite the practical success of LASSO in improving the generalizability and sparsification of neural networks, whether one can use LASSO for identifying significant variables is underexplored. For 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 $\text{supp}(\widehat{\boldsymbol{w}}) = \text{supp}(\boldsymbol{w})$ converges to one as $n \to \infty$. The standard approach to selecting a parsimonious sub-model is to either solve a penalized regression problem or iteratively pick up significant variables [12]. 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 [13]. For instance, consistency of the LASSO method [1] in estimating the significant variables has been extensively studied under various technical conditions, including sparsity, mutual coherence [14], restricted isometry [15], irrepresentable condition [16], and restricted eigenvalue [17].

Many theoretical studies of neural networks have focused on the generalizability. For example, a universal approximation theorem was established that shows any continuous multivariate function can be represented precisely by a polynomial-sized two-layer network [18]. It was later shown that any continuous function could be approximated arbitrarily well by a two-layer perceptron with sigmoid activation functions [19], and an approximation error bound of using two-layer neural networks to fit arbitrary smooth functions has been established [20], [21]. Statistically, generalization error bounds for two-layer neural networks [21], [22] and multi-layer networks [23], [24], [25] have been developed. From an optimization perspective, the parameter estimation of neural networks was cast into a tensor decomposition problem where a provably global optimum can be obtained [26], [27], [28]. 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 [29]. It was proved that certain deep neural networks with few nonzero network parameters could achieve minimax rates of convergence [30]. A tight error bound free from the input dimension was developed by assuming that the data is generated from a generalized hierarchical interaction model [31].

Manuscript received 23 December 2021; revised 2 December 2022; accepted 26 April 2023. Date of publication 8 May 2023; date of current version 18 August 2023. This work was supported in part by the U.S. National Science Foundation under Grant DMS-2134148 and Grant CNS-2220286. *(Corresponding author: Jie Ding.)*

Gen Li is with the Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA.

Ganghua Wang and Jie Ding are with the School of Statistics, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: dingj@umn.edu).

Communicated by M. Davenport, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2023.3274152

This work theoretically studies the identifiability of neural networks and uses it for variable selection. Specifically, suppose data observations are generated from a neural network with only a few nonzero coefficients. The identifiability concerns the possibility of identifying those coefficients, which may be further used to identify a sparse set of input variables that are genuinely relevant to the response. In this direction, LASSO and its variant Group LASSO [32] have been advocated to regularize neural-network for variable selection in practice (see, e.g., [6], [8], [9], [33]).

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

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

Here, p and r denote the input dimension and the number of neurons, respectively. We will assume that data are generated from a regression function in $\mathcal{F}_r$ perturbed by a small term. We will study the following two questions.

First, if the underlying regression function f admits a parsimonious representation so 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 an $\mathcal{S}^\star$ 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 problem $\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \|\boldsymbol{W}\|_1$ under the constraint of

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 \le \sigma^2,$$

which is an alternative version of the ℓ_1-regularization. More notational details will be introduced in Subsection II-B.

Our theory gives positive answers to the above questions. 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. 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. While this does not address wide neural networks, we think it still corresponds to a nontrivial and interesting function class. For example, the class contains linear functions when input signals are bounded. 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. 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. 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 II introduces the main theoretical result and proposes an algorithm to perform variable selection. Section III uses simulated and real-world datasets to demonstrate the proposed theory and algorithm. Section IV concludes the paper.

II. Main Result

A. 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

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

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 $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ 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 $v \in \mathbb{R}$.

B. Formulation

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

$$y_i = \sum_{j=1}^{r} a_j^\star \cdot \mathrm{relu}(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i + b_j^\star) + \xi_i \tag{1}$$

with $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, 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

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

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 of Inequality 2 is for simplicity. If ξ_i's are unbounded random variables, the theoretical result to be introduced will still hold, with more explanations in the Appendix. 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 main problem to address is whether one 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 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 II-E.

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

$$\min_{\boldsymbol{W},\boldsymbol{a},\boldsymbol{b}} \|\boldsymbol{W}\|_1 \quad \text{subject to } \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{r} a_j \cdot \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 \le \sigma^2, \tag{3}$$

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 data-generating process.

C. Main Result

We make the following technical assumption.

Assumption 1: For some constant $B \ge 1$,

$$1 \le \|\boldsymbol{w}_j^\star\|_2 \le B \quad \text{and} \quad |b_j^\star| \le B \quad \forall 1 \le j \le r. \tag{4}$$

In addition, for some constant $\omega > 0$,

$$\max_{j,k=1,\ldots,r, j\neq k} \frac{\left|\langle \boldsymbol{w}_j^\star, \boldsymbol{w}_k^\star\rangle\right|}{\|\boldsymbol{w}_j^\star\|_2 \|\boldsymbol{w}_k^\star\|_2} \le \frac{1}{r^\omega}. \tag{5}$$

Remark 1: (Discussion of Assumption 1) The condition in 4 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 5 requires that the angle of any two different coefficient vectors is not too small. This condition is analogous to a bounded-eigenvalue condition often assumed for linear regression problems, where each $w_j^\star$ is understood as a column in the design matrix. This condition is by no means mild or easy to verify in practice. Nevertheless, as our focused regime is large p, n but small r, we think the condition in 5 is still reasonable. For example, when $r = 2$, this condition simply requires $w_1^\star \neq w_2^\star$.

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). A skeptical reader may ask how the constants exactly depend on the sparsity and r. We will provide a more elaborated result in Subsection II-D and introduce the proof there.

Theorem 1: Under Assumption 1, there exist some 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 $1-\delta$,

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star \text{ and } \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^\star\boldsymbol{\Pi}^\top\|_{\mathrm{F}} + \|\widehat{\boldsymbol{b}} - \boldsymbol{\Pi}\boldsymbol{b}^\star\|_2 \le c_1\sigma \tag{6}$$

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

$$n > c_2 \log^4 \frac{p}{\delta} \qquad \text{and} \qquad \sigma < c_3. \tag{7}$$

Remark 2: (Interpretation 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 states 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 2.

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 6 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 III.

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,

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

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

Corollary 1 (Variable Selection): Let $\widehat{\mathcal{S}}_{c_1\sigma} \subseteq \{1,\ldots,p\}$ denote the sets of i's such that $\|\widehat{\boldsymbol{w}}_{i,\cdot}\|_2 > c_1\sigma$. Under the same assumption as in Theorem 1, and $\min_{1\le i\le r}\left\|\boldsymbol{w}_{i,\cdot}^\star\right\|_2 > 2c_1\sigma$, for any $\delta > 0$, one has

$$\mathbb{P}(\mathcal{S}^\star = \widehat{\mathcal{S}}_{c_1\sigma}) \ge 1 - \delta,$$

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 same assumption as in Theorem 1, there exists a constant $c_2 > 0$ depending only on the sparsity of $\boldsymbol{W}^\star$ such that for any $\delta > 0$, one has with probability at least $1-\delta$,

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

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

Corollary 2 states that among all the possible representations $\boldsymbol{W}$ in 1 (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 I.

It is worth mentioning that since the weight matrix $\boldsymbol{W}$ of the neural network is row-sparse, Group-LASSO is a suitable alternative to LASSO. We leave the analysis of Group-LASSO for future study.

D. Elaboration on the Main Result

Suppose that $\boldsymbol{W}^\star$ has at most s nonzero entries. The following theorem is a more elaborated version of Theorem 1.

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

$$\widehat{\boldsymbol{a}} = \boldsymbol{\Pi}\boldsymbol{a}^\star \text{ and } \|\widehat{\boldsymbol{W}} - \boldsymbol{W}^\star\boldsymbol{\Pi}^\top\|_{\mathrm{F}} + \|\widehat{\boldsymbol{b}} - \boldsymbol{\Pi}\boldsymbol{b}^\star\|_2 \le c_1\sigma \quad (8)$$

for some permutation $\boldsymbol{\Pi} \in \{0,1\}^{r\times r}$, provided that Assumption 1 holds and

$$n > c_2 s^3 r^{13} \log^4 \frac{p}{\delta} \qquad \text{and} \qquad \sigma < \frac{c_3}{r}. \quad (9)$$

Remark 3: (Sketch Proof of Theorem 1) The proof of Theorem 1 is nontrivial and is included in the Appendix. 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 allowing the invariance under any permutation of neurons (Remark 2). Then, we let

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

where $(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})$ is the solution of the problem in 3, and develop the following upper and lower bounds of it:

$$\Delta_n \le c_6\left(\frac{r}{S} + \frac{r\log^3\frac{p}{n\delta}}{n}\right) D_1^2 + c_6\sigma^2 \quad \text{and}$$
$$\Delta_n \ge c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \quad (10)$$

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) [15] 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 10 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.

Remark 4 (Alternative Assumption and Result): We provide an alternative to Theorem 2. Consider the following Assumption 1' as an alternative to Assumption 1.

Assumption 1': For some constant $B > 0$,

$$\|\boldsymbol{w}_j^\star\|_2 \le B \quad \text{and} \quad |b_j^\star| \le B \quad \text{for all } 1 \le j \le r.$$

In addition,

$$\mathbb{E}\Big[\langle \boldsymbol{a}, \mathrm{relu}(\boldsymbol{W}^\top\boldsymbol{x} + \boldsymbol{b})\rangle - \langle \boldsymbol{a}^\star, \mathrm{relu}(\boldsymbol{W}^{\star\top}\boldsymbol{x} + \boldsymbol{b}^\star)\rangle\Big]^2 \ge \psi D_2\left[(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}), (\boldsymbol{W}^\star, \boldsymbol{a}^\star, \boldsymbol{b}^\star)\right]^2, \quad (11)$$

and

$$n > \frac{c_2}{\psi} s^3 r^3 \log^4\frac{p}{\delta}. \quad (12)$$

With Assumption 1' instead of Assumption 1, one can still derive the same result as Theorem 2. The proof of the above result is similar to that of Theorem 2, except that we insert Inequality (11) instead of Inequality (22) into (21) in Appendix A.

E. Variable Selection

To solve the optimization problem 3 in practice, we consider the following alternative problem,

$$\min_{\boldsymbol{W}\in\mathbb{R}^{p\times r}, \boldsymbol{a}\in\mathbb{R}^r, \boldsymbol{b}\in\mathbb{R}^r} \left\{\frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{j=1}^r a_j \cdot \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j)\right)^2 + \lambda\|\boldsymbol{W}\|_1\right\}. \quad (13)$$

It has been empirically shown that algorithms such as the stochastic gradient descent can find a good approximate solution to the above optimization problem [34], [35]. Next, we will 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 hyper-parameters including the penalty term λ, the number of neurons r, learning rate, and the number of epochs. We suggest the usual approach that splits the available 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.

Variable importance. Inspired by Corollary 1, we interpret the ℓ_2-norm of $\widehat{\boldsymbol{w}}_{i,\cdot}$ as the importance of the i-th variable, for $i = 1, \ldots, p$. Corollary 1 indicates that 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. 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.

Extension to deep neural networks. Inspired by 13, we can intuitively generalize the proposed method to deep neural networks by penalizing the ℓ_1-norm of the weight matrix in the input layer. Though we do not have a theoretical analysis for this broader setting, numerical studies show that it is still effective.

III. 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 LASSO ('LASSO'), orthogonal matching pursuit ('OMP'), random forest ('RF'), gradient boosting ('GB'), neural networks with group LASSO ('GLASSO') [5], group sparse regularization ('GSR') [8], and LNET ('LNET') [9]. The 'NN' hyperparameters to search over are the penalty term $\lambda \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$, the number of neurons $r \in \{20, 50, 100\}$, the learning rate in $\{0.01, 0.005, 0.001\}$, and the number of epochs in $\{100, 200, 500\}$. Moreover, we extend 'NN' to a neural network that contains an additional hidden layer of ten neurons. We distinguish the proposed method with two-layer and three-layer neural networks by 'NN-2' and 'NN-3', respectively. Further experimental details are included in Appendix E.

A. Synthetic Datasets

1) NN-Generated Dataset: The first experiment uses the data generated from Equation (1) with $p = 100$ variables, $r = 16$ 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 is zero-mean Gaussian with standard deviation σ set to be 0, 0.5, 1, and 5. For each σ, we evaluate its mean squared error on the test dataset and three quantities for variable selection: the number of correctly selected variables ('TP', the larger the better), wrongly selected variables ('FP', the smaller the better), and area-under-curve score ('AUC', the larger the better). Here, 'AUC' is evaluated based on the variable importance given by each method, which is detailed in Appendix E. The procedure is independently replicated 20 times.

The results are reported in Table I and Table II, which suggest that 'NN' has the best overall performance for both selection and prediction. In particular, 'NN-2' and 'NN-3' have almost the same performance among all situations, which empirically demonstrates that the proposed method also works for deeper neural networks. It is interesting to compare 'NN' with 'LNET': 'NN' has slightly higher test error than 'LNET' when the noise level is small, but a much smaller false positive rate and higher AUC score than 'LNET'. It indicates that 'NN' is more accurate for variable selection, while 'LNET' tends to over-select variables for better prediction accuracy. Also, all the methods perform worse as the noise level σ increases.

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.

TABLE I
Performance Comparison on the NN-Generated Data, in Terms of the Number of Correctly ('TP'), Wrongly ('FP') Selected Features, and the AUC Score for Different σ. The Standard Errors Are Within the Parentheses

Method	Metric	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
LASSO	TP	8.30 (2.57)	9.10 (1.22)	9.20 (1.03)	6.70 (3.68)
	FP	10.40 (6.99)	10.70 (6.09)	13.35 (6.17)	8.60 (8.13)
	AUC	0.96 (0.06)	0.96 (0.05)	0.96 (0.05)	0.87 (0.14)
OMP	TP	8.45 (1.53)	8.30 (1.35)	8.00 (2.17)	6.00 (2.12)
	FP	0.10 (0.30)	0.15 (0.36)	0.25 (0.70)	0.65 (1.06)
	AUC	0.92 (0.08)	0.91 (0.07)	0.91 (0.09)	0.80 (0.11)
RF	TP	6.95 (3.12)	5.55 (3.29)	5.75 (3.27)	4.20 (2.60)
	FP	0.45 (0.67)	0.40 (0.66)	0.35 (0.57)	1.20 (2.18)
	AUC	0.99 (0.02)	0.97 (0.03)	0.95 (0.04)	0.86 (0.12)
GB	TP	6.85 (2.97)	7.15 (3.09)	5.75 (3.69)	5.65 (3.20)
	FP	1.35 (1.46)	1.60 (1.91)	2.10 (2.07)	5.20 (6.69)
	AUC	0.98 (0.02)	0.97 (0.03)	0.97 (0.03)	0.88 (0.11)
GLASSO	TP	9.35 (1.42)	9.80 (0.51)	9.45 (0.59)	6.35 (2.22)
	FP	0.10 (0.44)	0.65 (0.96)	1.00 (1.38)	8.50 (11.45)
	AUC	1.00 (0.00)	1.00 (0.00)	0.99 (0.02)	0.84 (0.14)
GSR	TP	9.55 (1.96)	10.00 (0.00)	9.90 (0.30)	7.90 (2.81)
	FP	1.25 (2.05)	0.70 (1.19)	1.35 (2.26)	22.45 (29.67)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	0.84 (0.18)
LNET	TP	10.00 (0.00)	10.00 (0.00)	9.95 (0.22)	5.10 (2.91)
	FP	66.40 (11.15)	59.00 (14.64)	41.60 (21.62)	8.05 (14.41)
	AUC	0.63 (0.06)	0.67 (0.08)	0.77 (0.12)	0.83 (0.12)
NN-2	TP	10.00 (0.00)	9.85 (0.36)	9.80 (0.51)	7.80 (1.96)
	FP	0.75 (1.13)	1.25 (2.09)	2.55 (5.80)	12.30 (10.57)
	AUC	**1.00 (0.00)**	**1.00 (0.01)**	**1.00 (0.01)**	**0.88 (0.13)**
NN-3	TP	9.75 (0.16)	9.85 (0.08)	9.10 (0.45)	6.95 (0.44)
	FP	0.75 (0.34)	0.65 (0.26)	0.45 (0.18)	2.50 (0.73)
	AUC	1.00 (0.00)	1.00 (0.00)	0.99 (0.00)	0.85 (0.03)

TABLE II
Performance Comparison on the NN-Generated Data, in Terms of the Average Mean Squared Error for Different σ

Method	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
LASSO	5.15 (0.64)	6.49 (1.08)	4.25 (0.49)	7.42 (0.79)
OMP	5.08 (0.72)	6.35 (1.06)	5.48 (1.14)	6.41 (0.63)
RF	12.62 (2.57)	7.07 (1.34)	9.89 (2.19)	15.94 (3.60)
GB	8.19 (2.20)	4.06 (0.70)	7.08 (1.17)	10.60 (1.56)
GLASSO	1.09 (0.06)	1.19 (0.07)	1.79 (0.08)	9.14 (0.32)
GSR	0.57 (0.03)	0.64 (0.04)	0.95 (0.04)	5.43 (0.37)
LNET	**0.51 (0.02)**	**0.63 (0.02)**	0.99 (0.03)	5.10 (0.34)
NN-2	0.67 (0.04)	0.74 (0.04)	1.06 (0.05)	**3.91 (0.22)**
NN-3	0.62 (0.03)	0.77 (0.04)	**0.87 (0.04)**	4.05 (0.25)

The other settings are the same as Subsubsection III-A.1. The results are presented in Tables III and IV.

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 'NN' approach is almost as good as the linear methods. Note that 'NN-3' outperforms 'NN-2' in this case. One possible explanation is that deeper neural networks have much larger expressivity than two-layer networks. On the other hand, the tree-based methods 'RF' and 'GB' perform significantly worse. This is possibly because the sample size $n = 60$ is relatively small, so the tree-based methods have a large variance. Meanwhile, the 'NN' uses the 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.

3) Friedman Dataset: This experiment uses the Friedman dataset with the following nonlinear data-generating process, $y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10\ x_4 + 5\ x_5 + \xi$.

TABLE III

PERFORMANCE COMPARISON ON THE LINEAR DATA, IN TERMS OF THE NUMBER OF CORRECTLY ('TP'), WRONGLY ('FP') SELECTED FEATURES, AND THE AUC SCORE FOR DIFFERENT σ

Method		$\sigma = 0$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$
LASSO	TP	3.00 (0.00)	3.00 (0.00)	2.85 (0.11)	2.05 (0.18)
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.50 (0.19)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.90 (0.03)
OMP	TP	3.00 (0.00)	2.90 (0.10)	2.95 (0.05)	2.15 (0.18)
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.30 (0.12)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.88 (0.03)
RF	TP	1.15 (0.08)	1.30 (0.12)	1.25 (0.12)	1.60 (0.16)
	FP	0.00 (0.00)	0.00 (0.00)	0.05 (0.05)	0.20 (0.11)
	AUC	1.00 (0.00)	0.99 (0.01)	0.99 (0.00)	0.83 (0.03)
GB	TP	1.35 (0.16)	1.35 (0.16)	1.30 (0.14)	1.90 (0.20)
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.30 (0.17)
	AUC	1.00 (0.00)	0.99 (0.01)	0.99 (0.01)	0.91 (0.02)
GLASSO	TP	2.80 (0.13)	2.70 (0.16)	2.40 (0.19)	1.95 (0.18)
	FP	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.70 (0.20)
	AUC	1.00 (0.00)	1.00 (0.00)	0.99 (0.00)	0.80 (0.04)
GSR	TP	2.90 (0.10)	2.90 (0.07)	2.80 (0.13)	1.90 (0.16)
	FP	0.00 (0.00)	0.10 (0.10)	0.00 (0.00)	0.55 (0.15)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.84 (0.03)
LNET	TP	3.00 (0.00)	3.00 (0.00)	2.85 (0.15)	1.70 (0.19)
	FP	0.00 (0.00)	0.20 (0.09)	0.95 (0.26)	0.55 (0.23)
	AUC	1.00 (0.00)	0.98 (0.01)	0.88 (0.03)	0.80 (0.04)
NN-2	TP	2.50 (0.17)	2.40 (0.17)	2.55 (0.14)	2.25 (0.17)
	FP	0.05 (0.05)	0.20 (0.15)	0.25 (0.12)	0.75 (0.18)
	AUC	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.89 (0.2)
NN-3	TP	3.00 (0.00)	2.65 (0.15)	2.90 (0.07)	2.10 (0.20)
	FP	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.35 (0.15)
	AUC	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**0.92 (0.02)**

TABLE IV

PERFORMANCE COMPARISON ON THE LINEAR DATA, IN TERMS OF THE NUMBER OF AVERAGE MEAN SQUARED ERROR FOR DIFFERENT σ

Method	$\sigma = 0$	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$
LASSO	0.00 (0.00)	0.04 (0.01)	0.17 (0.03)	4.09 (0.45)
OMP	**0.00 (0.00)**	**0.02 (0.00)**	**0.09 (0.01)**	4.19 (0.44)
RF	3.57 (0.22)	3.58 (0.21)	3.52 (0.16)	7.89 (0.55)
GB	2.45 (0.15)	3.04 (0.17)	3.01 (0.17)	11.59 (0.81)
GLASSO	0.10 (0.01)	0.21 (0.02)	0.34 (0.03)	4.95 (0.32)
GSR	0.09 (0.01)	0.18 (0.02)	0.30 (0.04)	4.16 (0.34)
LNET	0.18 (0.02)	0.17 (0.02)	0.34 (0.04)	3.70 (0.59)
NN-2	0.09 (0.01)	0.19 (0.02)	0.37 (0.04)	3.96 (1.37)
NN-3	0.03 (0.00)	0.10 (0.02)	0.17 (0.02)	**3.17 (0.45)**

TABLE V

PERFORMANCE COMPARISON ON THE FRIEDMAN DATA, IN TERMS OF THE NUMBER OF CORRECTLY ('TP'), WRONGLY ('FP') SELECTED FEATURES, AND THE AUC SCORE FOR DIFFERENT σ

Method		$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
LASSO	TP	4.05 (0.05)	4.00 (0.00)	4.10 (0.07)	4.00 (0.10)
	FP	1.45 (0.47)	1.85 (0.49)	2.00 (0.42)	3.10 (0.61)
	AUC	0.91 (0.01)	0.91 (0.01)	0.91 (0.01)	0.90 (0.01)
OMP	TP	4.00 (0.00)	4.00 (0.00)	4.00 (0.00)	3.80 (0.15)
	FP	0.10 (0.07)	0.10 (0.07)	0.10 (0.07)	0.05 (0.05)
	AUC	0.90 (0.00)	0.90 (0.00)	0.90 (0.00)	0.89 (0.01)
RF	TP	4.60 (0.27)	4.60 (0.27)	4.80 (0.19)	4.10 (0.29)
	FP	0.10 (0.07)	0.00 (0.00)	0.05 (0.05)	0.25 (0.12)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
GB	TP	4.70 (0.21)	4.90 (0.10)	4.80 (0.19)	4.30 (0.25)
	FP	0.00 (0.00)	0.00 (0.00)	0.05 (0.05)	0.40 (0.18)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	**1.00 (0.00)**
GLASSO	TP	4.80 (0.15)	4.80 (0.09)	4.35 (0.20)	3.80 (0.09)
	FP	0.05 (0.05)	0.25 (0.12)	0.10 (0.10)	0.95 (0.37)
	AUC	1.00 (0.00)	0.99 (0.01)	0.99 (0.00)	0.88 (0.01)
GSR	TP	4.20 (0.22)	4.60 (0.18)	4.70 (0.16)	4.00 (0.07)
	FP	0.15 (0.11)	0.35 (0.11)	0.25 (0.10)	2.30 (0.52)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.91 (0.01)
LNET	TP	4.75 (0.24)	4.75 (0.24)	5.00 (0.00)	3.45 (0.17)
	FP	21.20 (2.00)	25.90 (1.97)	31.35 (1.55)	2.15 (1.24)
	AUC	0.74 (0.02)	0.69 (0.02)	0.65 (0.02)	0.91 (0.02)
NN-2	TP	4.80 (0.13)	4.40 (0.20)	4.80 (0.13)	4.20 (0.13)
	FP	0.50 (0.22)	0.50 (0.26)	0.60 (0.29)	1.25 (0.32)
	AUC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.95 (0.01)
NN-3	TP	4.85 (0.08)	4.90 (0.07)	4.85 (0.08)	4.45 (0.11)
	FP	0.35 (0.16)	0.25 (0.12)	0.55 (0.22)	3.85 (0.70)
	AUC	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	0.96 (0.01)

TABLE VI

PERFORMANCE COMPARISON ON THE FRIEDMAN DATA, IN TERMS OF THE AVERAGE MEAN SQUARED ERROR FOR DIFFERENT σ

Method	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$
LASSO	6.16 (0.05)	6.17 (0.06)	6.07 (0.05)	6.97 (0.12)
OMP	5.95 (0.07)	6.01 (0.07)	6.04 (0.06)	6.41 (0.16)
RF	5.21 (0.05)	5.19 (0.09)	5.30 (0.07)	7.81 (0.15)
GB	2.70 (0.05)	2.74 (0.04)	2.82 (0.06)	6.33 (0.15)
GLASSO	4.49 (0.12)	4.64 (0.17)	5.43 (0.14)	9.89 (0.23)
GSR	1.57 (0.05)	1.73 (0.07)	2.31 (0.11)	7.49 (0.19)
LNET	**1.05 (0.19)**	**1.29 (0.26)**	**1.52 (0.12)**	9.08 (0.47)
NN-2	1.58 (0.04)	1.71 (0.06)	2.14 (0.06)	**5.86 (0.11)**
NN-3	1.44 (0.04)	1.62 (0.04)	1.92 (0.04)	6.02 (0.16)

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 V and VI. For this nonlinear dataset, 'NN' and 'GB' accurately find the significant variables and exclude redundant ones, while the linear methods fail to select the quadratic factor x_3. As for the prediction performance, neural network-based methods outperform other methods. In particular, 'NN' is better than 'GLASSO' and 'GSR', while 'LNET' exhibits better prediction and worse selection performance as seen in previous experiments.

B. BGSBoy Dataset

The BGSBoy dataset involves 66 boys from the Berkeley guidance study (BGS) of children born in 1928-29 in Berkeley, CA [36]. The dataset includes the height ('HT'), weight ('WT'), 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.

$$\text{BMI18} = \text{WT18}/(\text{HT18}/100)^2, \tag{14}$$

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 VII.

From the results, both linear and NN-based methods can identify 'WT18' and 'HT18' in all the replications. Meanwhile, tree-based methods may miss 'HT18' but select 'LG18' instead, which is only correlated with the response. Interestingly, we find that the linear methods still predict well in this experiment. A possible reason is that Equation (14) can be well-approximated by a first-order Taylor expansion on

TABLE VII

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 20 TIMES)

Method	LASSO	OMP	RF	GB	GLASSO	GSR	LNET	NN-2	NN-3
RMSE	0.05 (0.00)	0.04 (0.00)	3.10 (0.37)	2.32 (0.30)	0.15 (0.06)	0.09 (0.03)	**0.03 (0.00)**	0.06 (0.02)	0.12 (0.02)
Top 3 frequently selected features	WT18 (20) HT18 (20) HT9 (16)	WT18 (20) HT18 (20) ST18 (7)	WT18 (20) LG18 (20) HT18 (10)	WT18 (20) LG18 (20) HT18 (19)	WT18 (20) HT18 (20) LG18 (8)	WT18 (20) HT18 (20) LG18 (12)	WT18 (20) HT18 (20) HT9 (9)	WT18 (20) HT18 (20) HT9 (12)	WT18 (20) HT18 (20) LG18 (17)

TABLE VIII

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 20 TIMES)

Method	LASSO	OMP	RF	GB	
RMSE	14.20 (0.08)	16.75 (0.12)	9.60 (0.12)	11.02 (0.09)	
Top 10 frequently selected features	BUILDINGID_2 (20) BUILDINGID_1 (20) USERID_9 (20) USERID_16 (20) USERID_10 (18) WAP099 (18) USERID_7 (14) WAP121 (10) WAP118 (8) WAP039 (7)	BUILDINGID_2 (20) BUILDINGID_1 (20) WAP099 (17) USERID_10 (17) USERID_16 (14) USERID_7 (13) USERID_9 (10) WAP120 (8) WAP124 (8) WAP101 (8)	BUILDINGID_1 (20) BUILDINGID_2 (20) WAP141 (16) WAP120 (16) WAP117 (14) WAP035 (13) WAP173 (13) WAP167 (10) WAP118 (8) WAP113 (8)	BUILDINGID_2 (20) WAP120 (20) BUILDINGID_1 (20) WAP141 (17) WAP099 (16) PHONEID_14 (14) WAP113 (13) WAP114 (12) WAP117 (12) WAP140 (9)	
Method	**GLASSO**	**GSR**	**LNET**	**NN-2**	**NN-3**
RMSE	11.25 (0.14)	11.44 (0.18)	11.19 (0.09)	**8.86 (0.09)**	**8.86 (0.09)**
Top 10 frequently selected features	BUILDINGID_1 (20) BUILDINGID_2 (20) PHONEID_14 (18) SPACEID_202 (17) USERID_8 (16) USERID_16 (16) FLOOR_3 (12) USERID_9 (9) WAP478 (8) WAP099 (7)	BUILDINGID_1 (20) BUILDINGID_2 (20) USERID_16 (17) SPACEID_202 (15) PHONEID_6 (14) SPACEID_203 (14) USERID_7 (13) SPACEID_201 (12) WAP141 (6) WAP176 (6)	BUILDINGID_1 (20) BUILDINGID_2 (20) PHONEID_22 (20) PHONEID_6 (20) PHONEID_8 (20) SPACEID_103 (20) SPACEID_136 (20) SPACEID_201 (20) SPACEID_202 (20) SPACEID_203 (20)	SPACEID_202 (20) BUILDINGID_1 (20) BUILDINGID_2 (20) USERID_16 (19) SPACEID_203 (16) SPACEID_201 (14) WAP140 (7) WAP121 (5) WAP030 (5) SPACEID_224 (5)	SPACEID_202 (20) BUILDINGID_1 (20) BUILDINGID_2 (20) USERID_16 (17) SPACEID_203 (13) SPACEID_201 (13) USERID_9 (8) WAP121 (8) WAP030 (7) SPACEID_224 (5)

'HT18' at the value of around 180, and the range of 'HT18' is within a small interval around 180.

C. 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 [37]. Specifically, we have 520 Wireless Access Points (WAPs) signals (which are continuous variables) and 'FLOOR', 'BUILDING', '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 VIII.

Based on the results, the 'NN' achieves the best prediction performance 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 with some overlapping, e.g., 'PHONEID_14' selected by 'GLASSO' and 'GB', 'USERID_16' selected by 'NN' and 'LASSO', which indicate the potentially important variables. 'LNET' again selects more variables than other methods. There are nearly 60 variables selected by 'LNET' in every replication. 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 very high missing rate of WAPs signals (97% on average, as reported in [37]), the interpretation of 'NN' seems reasonable.

D. 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. Using group LASSO or a mixed type of penalty has a similar performance as 'NN', while 'LNET' tends to over-select importance variables. Fourth, the selection and prediction performances are often positively associated for 'NN', but may not be the case for baseline methods.

IV. 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, a limitation of the work is that we considered only a small r. An interesting future problem is to study r that may grow fast with p and n. Second, our experiments show that the algorithm can be extended to deeper neural networks. It will be exciting to generalize the main theorem to the multi-layer cases.

The Appendix includes proofs and experimental details.

Appendix A
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 $\widehat{\boldsymbol{W}}$, where S will be specified momentarily. 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

$$d_1(\boldsymbol{w}_1, a_1, b_1, \boldsymbol{w}_2, a_2, b_2) = \begin{cases} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 + |b_1 - b_2| & \text{if } a_1 = a_2; \\ \|\boldsymbol{w}_1\|_1 + \|\boldsymbol{w}_2\|_1 + |b_1| + |b_2| & \text{if } a_1 \neq a_2, \end{cases} \tag{15}$$

and

$$d_2(\boldsymbol{w}_1, a_1, b_1, \boldsymbol{w}_2, a_2, b_2) = \begin{cases} \sqrt{\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + |b_1 - b_2|^2} & \text{if } a_1 = a_2; \\ 1 & \text{if } a_1 \neq a_2. \end{cases} \tag{16}$$

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

$$D_1 \coloneqq \min_{\pi} \sum_{j=1}^{r} d_1(\boldsymbol{w}_{\pi(j)}, a_{\pi(j)}, b_{\pi(j)}, \boldsymbol{w}_j^\star, a_j^\star, b_j^\star), \tag{17a}$$

$$D_2 \coloneqq \min_{\pi} \sqrt{\sum_{j=1}^{r} d_2(\boldsymbol{w}_{\pi(j)}, a_{\pi(j)}, b_{\pi(j)}, \boldsymbol{w}_j^\star, a_j^\star, b_j^\star)^2} \tag{17b}$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \geq c_4 \min\left\{\frac{1}{r}, D_2^2\right\} \tag{18}$$

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

$$n \geq c_5 S^3 r^4 \log^4 \frac{p}{\delta}. \tag{19}$$

Lemma 2: Then, there exists a universal constant $c_6 > 0$ such that

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

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

By comparing the bounds given in Lemma 1 and 2, 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

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

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 (17). Observing the fact that for $S \geq 2s$, one has $\mathcal{S}^\star \subset \widehat{\mathcal{S}}^\star \subset \mathcal{S}$, we have

$$\|\boldsymbol{w}_j - \boldsymbol{w}_j^\star\|_2 \geq \|\boldsymbol{w}_{j,\widehat{\mathcal{S}}^\star} - \boldsymbol{w}_j^\star\|_2 = \|\widehat{\boldsymbol{w}}_j - \boldsymbol{w}_j^\star\|_2,$$

after some permutation, and then

$$D_2^\star \leq D_2.$$

In addition, after some permutation, we have $D_1^\star \geq \left\|\widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star} - \boldsymbol{W}^\star\right\|_1 \geq \|\boldsymbol{W}^\star\|_1 - \left\|\widehat{\boldsymbol{W}}_{\mathcal{S}^\star}\right\|_1$ and $\|\boldsymbol{W}\|_1 \leq \left\|\widehat{\boldsymbol{W}}\right\|_1 \leq \|\boldsymbol{W}^\star\|_1$. Then,

$$D_1 \leq D_1^\star + \left\|\widehat{\boldsymbol{W}}_{\mathcal{S}} - \widehat{\boldsymbol{W}}_{\widehat{\mathcal{S}}^\star}\right\|_1 \leq D_1^\star + \left\|\widehat{\boldsymbol{W}}\right\|_1 - \left\|\widehat{\boldsymbol{W}}_{\mathcal{S}^\star}\right\|_1 \leq 2D_1^\star.$$

Combined with Lemma 3 in Appendix D, the above results give

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

provided that for some constant $c_7 > 0$

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

such that

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

Then, we can conclude the proof since after appropriate permutation,

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

APPENDIX B
PROOF OF LEMMA 1 (LOWER BOUND)

This can be seen from the following three properties.

- Consider the case that

$$D_1 \le \epsilon = \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

With probability at least $1-\delta$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) \\ &\qquad - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\Bigg]^2 \\ &= \frac{D_1^2}{\epsilon^2}\frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{r} a_j \mathrm{relu}(\widetilde{\boldsymbol{w}}_j^\top \boldsymbol{x}_i + \widetilde{b}_j) \\ &\qquad - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\Bigg]^2, \end{aligned} \tag{20}$$

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

- For any $\epsilon > 0$ and

$$D_1 \ge \frac{\epsilon}{\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}},$$

there exists some universal constant $C_1 > 0$, such that with probability at least $1-\delta$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) \\ &\qquad - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\Bigg]^2 \\ &\ge \mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star)\right]^2 \\ &\quad - C_1 D_1^2 \log\frac{pn}{\delta}\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}}. \end{aligned} \tag{21}$$

- For some universal constant $C_2 > 0$

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x} + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x} + b_j^\star)\right]^2 \\ &\ge C_2 \min\left\{\frac{1}{r}, D_2^2\right\}. \end{aligned} \tag{22}$$

Putting the above 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 (22) into (21) gives that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star)\right]^2 \\ &\ge C_2 \min\left\{\frac{1}{r}, D_2^2\right\} - C_1 D_1^2 \log\frac{pn}{\delta}\sqrt{\frac{S}{n}\log\frac{pr}{S}\log\frac{BS}{\epsilon\delta}} \\ &\ge \frac{C_2}{2}\min\left\{\frac{1}{r}, D_2^2\right\}, \end{aligned} \tag{23}$$

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

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

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

$$D_1 < \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}},$$

it follows from (20) and (23) that

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

Summing up, we conclude the proof by verifying (20), (21), and (22) below.

A. Proof of (20)

Since $D_1 \le \epsilon = \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}$, without loss of generality, we assume that $a_j = a_j^\star$ for $1 \le j \le r$, and

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

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

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

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

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

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

$$\|\boldsymbol{x}_i\|_\infty \le \sqrt{2\log\frac{4pn}{\delta}} \qquad \text{for all } 1\le i\le n. \tag{24}$$

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

$$u(\boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j) = u(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}_i + b_j^\star), \tag{25}$$

provided that

$$\epsilon \le \frac{\delta}{4nr}\sqrt{\frac{\pi}{\log\frac{4pn}{\delta}}}.$$

Note that $u(x)=1$ if $x>0$, and $u(x)=0$ if $x\le 0$. Then combining with the definition of $\widetilde{\boldsymbol{w}}_j$ and $\widetilde{b}_j$, the above property yields

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

and the claim is proved. Here, the last equality holds due to (25) and $a_j=a_j^\star$ for $j=1,\dots,r$.

B. Proof of (21)

Notice that

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

which leads to

$$\begin{aligned}
&\left|\sum_{j=1}^r a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^r a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right|\\
&\le D_1\max\left\{\|\boldsymbol{x}\|_\infty,1\right\}.
\end{aligned}\tag{26}$$

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

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

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\left(z_i^2-\mathbb{E}\left[z_i^2\right]\right) &(27)\\
&=\frac{1}{n}\sum_{i=1}^n\left\{z_i^2\mathbb{1}(\mathcal{E})-\mathbb{E}\left[z_i^2\mathbb{1}(\mathcal{E})\right]-\mathbb{E}\left[z_i^2\mathbb{1}(\overline{\mathcal{E}})\right]\right\}\\
&\ge -4D_1^2\log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2}{\delta}}-D_1^2\frac{\delta}{n}\\
&\ge -5D_1^2\log\frac{4pn}{\delta}\sqrt{\frac{1}{n}\log\frac{2}{\delta}}. &(28)
\end{aligned}$$

Here, the first line holds due to (24); 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

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

where $[n]\coloneqq\{1,2,\dots,n-1\}$. 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 (26), we can derive that

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

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 $1-\delta$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\left(z_i^2-\mathbb{E}\left[z_i^2\right]\right)-\frac{1}{n}\sum_{i=1}^n\left(\widetilde{z}_i^2-\mathbb{E}\left[\widetilde{z}_i^2\right]\right)\\
&\ge -4\epsilon\left(\epsilon+D_1\right)\log\frac{4pn}{\delta}.
\end{aligned}\tag{29}$$

In addition, it can be verified that

$$\log|\mathcal{N}_\epsilon|\le C_5 S\log\frac{pr}{S}\log\frac{BS}{\epsilon}, \tag{30}$$

for some universal constant $C_5 > 0$. Combining (28), (29), and (30) leads to

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

It follows that (21) holds.

C. Proof of (22)

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

$$\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\right]^2 \geq C_6 \min\left\{\frac{1}{r}, D_2^2\right\}. \tag{31}$$

Next, we will assume that

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

Otherwise, Inequality 31 already holds. According to Lemma 4, one has for any constant $k \geq 0$, there exists some constant $\alpha_k > 0$ such that

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=1}^{r} a_j \mathrm{relu}(\boldsymbol{w}_j^\top \boldsymbol{x}) - \sum_{j=1}^{r} a_j^\star \mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x})\right]^2 \\ &\geq \alpha_k \left\| \sum_{j=1}^{r} a_j \|\boldsymbol{w}_j\|_2 \left(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\right)^{\otimes 2k} - \sum_{j=1}^{r} a_j^\star \|\boldsymbol{w}_j^\star\|_2 \left(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\right)^{\otimes 2k} \right\|_F^2.\end{aligned} \tag{32}$$

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

$$\left|\langle \boldsymbol{v}_{j_1}^\star, \boldsymbol{v}_{j_2}^\star\rangle\right| \leq \frac{1}{r^2}. \tag{33}$$

where

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

and

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

Then, (32) 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

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

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 (33).

Moreover, by means of AM-GM inequality and (33), one can see that

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

Then combining with (31), the above result and the counterpart for $\beta_j^\star > 0$ lead to

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

which gives

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

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

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

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

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

and

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

Then for any i,

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

which, together with (31), 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

$$\|\sqrt{\beta_j}\boldsymbol{v}_j-\sqrt{\beta_j^\star}\boldsymbol{v}_j^\star\|_2^2\leq\frac{1}{r},\qquad\text{for all } 1\leq j\leq r.\tag{34}$$

Then

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

Here, the first line comes from (32); the second line holds through the following claim

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

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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right]^2\\
&\geq\sum_{k\geq\frac{12}{\omega}}^{\infty}\Bigg\|\sum_{j=1}^{r}a_jf_k\big(\frac{b_j}{\|\boldsymbol{w}_j\|_2}\big)\|\boldsymbol{w}_j\|_2\big(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\big)^{\otimes k}\\
&\qquad-\sum_{j=1}^{r}a_j^\star f_k\big(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)\|\boldsymbol{w}_j^\star\|_2\big(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)^{\otimes k}\Bigg\|_F^2\\
&\gtrsim\sum_{j=1}^{r}\sum_{k\geq\frac{12}{\omega}}^{\infty}\left\|a_jf_k\big(\frac{b_j}{\|\boldsymbol{w}_j\|_2}\big)\boldsymbol{w}_j-a_j^\star f_k\big(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\big)\boldsymbol{w}_j^\star\right\|_F^2\\
&\gtrsim\sum_{j=1}^{r}\inf_{R_l(\boldsymbol{x})}\mathbb{E}\Big[a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)\\
&\qquad-a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)-R_l(\boldsymbol{x})\Big]^2\\
&\gtrsim\sum_{j=1}^{r}\left(\|\boldsymbol{w}_j-\boldsymbol{w}_j^\star\|_2^2+|b_j-b_j^\star|^2\right).
\end{aligned}\tag{35}$$

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.

Appendix C
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 $\left(\widehat{\boldsymbol{W}},\widehat{\boldsymbol{a}},\widehat{\boldsymbol{b}}\right)$ with respect to the problem in 3. According to the triangle inequality, one has

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}z_i^2}\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{z}_i^2}+2\sigma.\tag{36}$$

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

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

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 $C_7>0$,

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

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

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

and

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

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 $\|\widehat{\boldsymbol{W}}\|_1\le\|\boldsymbol{W}^\star\|_1$ and

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

Putting everything together gives the desired result.

APPENDIX D
TECHNICAL LEMMAS

Lemma 3: For any $(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b})$ with $\|\boldsymbol{W}\|_0+\|\boldsymbol{b}\|_0+\|\boldsymbol{W}^\star\|_0+\|\boldsymbol{b}^\star\|_0\le S$. Assume that $\|\boldsymbol{W}\|_1+\|\boldsymbol{b}\|_1\le\|\boldsymbol{W}^\star\|_1+\|\boldsymbol{b}^\star\|_1$ and $\|\boldsymbol{w}_j^\star\|_2^2+|b_j^\star|^2\le1$. Then one has

$$D_1\le2\sqrt{S}D_2,\tag{37}$$

where D_1,D_2 are defined in (17).

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

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

Then, according to the AM-GM inequality, one has

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

which implies that

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

Thus we conclude the proof. □

Lemma 4 (Theorem 2.1 [27]): For any constant $k\ge0$, there exists some universal function $f_k:\mathbb{R}\to\mathbb{R}$ such that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{r}a_j\mathrm{relu}(\boldsymbol{w}_j^\top\boldsymbol{x}+b_j)-\sum_{j=1}^{r}a_j^\star\mathrm{relu}(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}+b_j^\star)\right]^2\\
&=\sum_{k=0}^{\infty}\left\|\sum_{j=1}^{r}a_jf_k\left(\frac{b_j}{\|\boldsymbol{w}_j\|_2}\right)\|\boldsymbol{w}_j\|_2\left(\frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2}\right)^{\otimes k}\right.\\
&\qquad\left.-\sum_{j=1}^{r}a_j^\star f_k\left(\frac{b_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\right)\|\boldsymbol{w}_j^\star\|_2\left(\frac{\boldsymbol{w}_j^\star}{\|\boldsymbol{w}_j^\star\|_2}\right)^{\otimes k}\right\|_F^2,
\end{aligned}\tag{38}$$

with

$$\alpha_k\coloneqq f_{2k}(0)>0,\qquad\text{for all } k>0.\tag{39}$$

In addition, we have

$$\begin{aligned}
&\inf_{R_l}\mathbb{E}\left[a\mathrm{relu}(\boldsymbol{w}^\top\boldsymbol{x}+b)-\sum_{j=1}^{r}a^\star\mathrm{relu}(\boldsymbol{w}^{\star\top}\boldsymbol{x}+b^\star)-R_l(\boldsymbol{x})\right]^2\\
&=\sum_{k>l}^{\infty}\left\|af_k\left(\frac{b}{\|\boldsymbol{w}\|_2}\right)\|\boldsymbol{w}\|_2\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\right)^{\otimes k}\right.\\
&\qquad\left.-a^\star f_k\left(\frac{b^\star}{\|\boldsymbol{w}^\star\|_2}\right)\|\boldsymbol{w}^\star\|_2\left(\frac{\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2}\right)^{\otimes k}\right\|_F^2,
\end{aligned}\tag{40}$$

where R_l denote a polynomial function of $\boldsymbol{x}$ with degree less than l.

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

$$\frac{1}{n}\sum_{i=1}^{n}\left|\boldsymbol{w}^\top\boldsymbol{x}_i\right|^2\le c\|\boldsymbol{w}\|_2^2+c\frac{\log^3\frac{p}{n\delta}}{n}\|\boldsymbol{w}\|_1^2,\tag{41}$$

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

TABLE IX
HYPER-PARAMETERS USED IN OUR EXPERIMENTS

Dataset	Linear	NN-generated	Friedman	BGSBoy	UJIIndoorLoc
Epochs	{100,200,500}				
Batch size	32	32	32	8	128
Optimizer	ADAM				
Learning rate	{0.001, 0.005, 0.01}				
Scheduler	N/A				

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 $1 - \delta$,

$$\left\|\boldsymbol{X}^\top \boldsymbol{w}\right\|_2^2 \leq 2\|\boldsymbol{w}\|_2^2 \tag{42}$$

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 (42), 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

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

We conclude the proof by combining the above inequalities. □

APPENDIX E
FURTHER EXPERIMENTS DETAILS

The hyper-parameters used in Section III are summarized in Table IX.

We briefly explain the variable selection procedure. We first obtain a vector of the variables' importance. For 'LASSO' and 'OMP', we use the absolute value of the estimated coefficient as the variable importance; for 'NN', 'GLASSO', and 'GSR', we obtain the importance by applying row-wise ℓ_2-norm to the weight matrix in the input layer of the neural network; for 'RF', 'GB', and 'LNET', we use the importance produced by those methods. Once we have the importance vector, we can obtain the receiver operating characteristic (ROC) curve for synthetic datasets by varying the cut-off thresholds and calculate the AUC score. As for variable selection, we apply GMM of two mixtures to the importance vector for the synthetic datasets. The variables in the cluster with higher importance are considered significant. Then, we calculate the correctly or wrongly selected variables accordingly. For BGSBoy and UJIIndoorLoc datasets, the variables with the three- and ten-largest importance are selected, respectively.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the Associate Editor for their constructive review comments.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[2] S. A. van de Geer, "High-dimensional generalized linear models and the lasso," *Ann. Statist.*, vol. 36, no. 2, pp. 614–645, Apr. 2008.
[3] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statist. Med.*, vol. 16, no. 4, pp. 385–395, Feb. 1997.
[4] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1, no. 2. Cambridge, MA, USA: MIT Press, 2016.
[5] J. Wang, Q. Cai, Q. Chang, and J. M. Zurada, "Convergence analyses on sparse feedforward neural networks via group lasso regularization," *Inf. Sci.*, vol. 381, pp. 250–269, Mar. 2017.
[6] W. Yang, G. Wang, J. Ding, and Y. Yang, "A theoretical understanding of neural network compression from sparse linear approximation," 2022, *arXiv:2206.05604*.
[7] E. Diao, G. Wang, J. Zhan, Y. Yang, J. Ding, and V. Tarokh, "Pruning deep neural networks from a sparsity perspective," in *Proc. ICLR*, 2023, pp. 1–28.
[8] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, "Group sparse regularization for deep neural networks," *Neurocomputing*, vol. 241, pp. 81–89, Jun. 2017.
[9] I. Lemhadri, F. Ruan, and R. Tibshirani, "LassoNet: Neural networks with feature sparsity," in *Proc. AISTATS*, 2021, pp. 10–18.
[10] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. USENIX*, 2016, pp. 265–283.
[11] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, vol. 32, 2019, pp. 1–12.
[12] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, Nov. 2018.
[13] J. Zhang, Y. Yang, and J. Ding, "Information criteria for model selection," *WIREs Comput. Statist.*, Feb. 2023, Art. no. e1607.
[14] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Jul. 2001.
[15] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
[16] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Nov. 2006.
[17] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, Aug. 2009.

[18] A. N. Kolmogorov, "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition," *Doklady Akademii Nauk*, vol. 114, no. 5, pp. 953–956, 1957.
[19] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signals Syst.*, vol. 2, pp. 183–192, Dec. 1989.
[20] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.
[21] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.*, vol. 14, no. 1, pp. 115–133, 1994.
[22] G. Li, Y. Gu, and J. Ding, "L_1 regularization in two-layer neural networks," *IEEE Signal Process. Lett.*, vol. 29, pp. 135–139, 2021.
[23] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Proc. COLT*, 2015, pp. 1376–1401.
[24] N. Golowich, A. Rakhlin, and O. Shamir, "Size-independent sample complexity of neural networks," in *Proc. COLT*, 2018, pp. 297–299.
[25] G. Li and J. Ding, "Towards understanding variation-constrained deep neural networks," *IEEE Trans. Signal Process.*, vol. 71, pp. 631–640, 2023.
[26] M. Janzamin, H. Sedghi, and A. Anandkumar, "Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods," 2015, *arXiv:1506.08473*.
[27] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," in *Proc. ICLR*, 2018, pp. 1–45.
[28] M. Mondelli and A. Montanari, "On the connection between learning two-layer neural networks and tensor decomposition," in *Proc. AISTATS*, 2019, pp. 1051–1060.
[29] A. R. Barron and J. M. Klusowski, "Complexity, statistical risk, and metric entropy of deep nets using total path variation," *Stat*, vol. 1050, p. 6, 2019.
[30] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Ann. Statist.*, vol. 48, no. 4, pp. 1875–1897, Aug. 2020.
[31] B. Bauer and M. Kohler, "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Ann. Statist.*, vol. 47, no. 4, pp. 2261–2285, Aug. 2019.
[32] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, Feb. 2006.
[33] V. C. Dinh and L. S. Ho, "Consistent feature selection for analytic deep neural networks," in *Proc. NeurIPS*, 2020, pp. 2420–2431.
[34] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*. Berlin, Germany: Springer, 2010, pp. 177–186.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[36] R. D. Tuddenham, "Physical growth of California boys and girls from birth to eighteen years," *Univ. Calif. Publ. Child Develop.*, vol. 1, no. 2, pp. 183–364, 1954.
[37] J. Torres-Sospedra et al., "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proc. IPIN*, Oct. 2014, pp. 261–270.

Gen Li received the B.S. degree in electronic engineering and mathematics from Tsinghua University in 2016 and the Ph.D. degree in electronic engineering from Tsinghua University in 2021. He is currently a Post-Doctoral Researcher with the Department of Statistics and Data Science, The Wharton School, University of Pennsylvania. His research interests include reinforcement learning, high-dimensional statistics, machine learning, signal processing, and mathematical optimization. He has received the Excellent Graduate Award and the Excellent Thesis Award from Tsinghua University.

Ganghua Wang received the B.S. degree from Peking University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Statistics, University of Minnesota, Twin Cities, MN, USA. His research interests include the foundations of machine learning theory and machine learning safety.

Jie Ding (Member, IEEE) received the B.S. degree from Tsinghua University, Beijing, in 2012, and the Ph.D. degree in engineering science from Harvard University, Cambridge, in 2017. He joined as a Faculty Member of the University of Minnesota, Twin Cities, in 2018, where he has been an Assistant Professor with the School of Statistics, with a graduate faculty appointment with the Department of Electrical and Computer Engineering. His research interests include the intersection of machine learning, statistics, signal processing, and information theory.