Statistics and Probability Letters 183 (2022) 109353

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Searching for minimal optimal neural networks

Lam Si Tung Ho [a,*,1], Vu Dinh [b,1]

[a] Dalhousie University, Halifax, Nova Scotia, Canada
[b] Department of Mathematical Sciences, University of Delaware, DE, USA

ARTICLE INFO

Article history:
Received 20 May 2021
Received in revised form 11 October 2021
Accepted 26 December 2021
Available online 5 January 2022

MSC:
62J02
62J07
62M45
62F07

Keywords:
Neural networks
Model selection
Destructive technique
Adaptive Lasso
Consistency

ABSTRACT

Large neural network models have high predictive power but may suffer from overfitting if the training set is not large enough. Therefore, it is desirable to select an appropriate size for neural networks. The destructive approach, which starts with a large architecture and then reduces the size using a Lasso-type penalty, has been used extensively for this task. Despite its popularity, there is no theoretical guarantee for this technique. Based on the notion of minimal neural networks, we posit a rigorous mathematical framework for studying the asymptotic theory of the destructive technique. We prove that Adaptive group Lasso is consistent and can reconstruct the correct number of hidden nodes of one-hidden-layer feedforward networks with high probability. To the best of our knowledge, this is the first theoretical result establishing for the destructive technique.

1. Introduction

Artificial neural networks are highly expressive models that achieve excellent performance on many tasks. However, the performance of a neural network model depends heavily on its structure. In particular, training with oversized neural networks on small or moderate datasets can lead to overfitting. Moreover, training large neural networks requires a high memory and computation cost. Therefore, choosing the right size for neural networks is an important problem and has been studied intensively.

Two common approaches for this task are the constructive method (Bello, 1992) and the destructive technique (LeCun et al., 1990). The constructive method starts with a small neural network and gradually incorporates additional components until finding the best architecture. The destructive technique, on the other hand, starts with a large neural network and remove unimportant components. One drawback of the constructive method is that we have to train a new model each time we add a new component. On the contrary, we can utilize a Lasso-type penalty for the destructive technique to avoid this pitfall. Although this method has been used extensively, to the best of our knowledge, no theoretical result has been established, even for one-hidden-layer feedforward networks. The main challenge is that neural network models are non-linear and unidentifiable.

Choosing the architecture of neural networks can be considered as a model selection procedure. To study asymptotic properties of such procedures, we posit a rigorous mathematical framework using the notion of *minimal neural networks*.

* Corresponding author.
E-mail address: Lam.Ho@dal.ca (L.S.T. Ho).
[1] Authors contributed equally.

https://doi.org/10.1016/j.spl.2021.109353
0167-7152/

For simplicity, we focus on one-hidden-layer feedforward networks with hyperbolic tangent activation function, which will be called "networks" from now on. In particular, a function $f : \mathbb{R}^d \to \mathbb{R}$ is a network if $f(x) = v^\top \cdot \tanh(u \cdot x + b_1) + b_2$, where u is a $H \times d$ matrix; b_1, v are H-dimensional vectors; and b_2 is a real number. Here, $v^\top$ is the transpose of v; H is the number of nodes in the hidden layer, u, v are the weights; and b_1, b_2 are the biases. A network is called *minimal* if it does not have the same input–output map with a network with fewer hidden nodes. The best structure for the training network is the structure of an optimal network that is also minimal. It is obvious that there exists such a network if there exists an optimal network. Recall that a network is *optimal* if it minimizes the expected risk. Therefore, we need to search for a minimal network among optimal networks. We call it *a minimal optimal network*.

A popular penalty for automatically reducing the number of hidden nodes of neural networks is the Group Lasso (Murray and Chiang, 2015; Alvarez and Salzmann, 2016; Scardapane et al., 2017; Huang et al., 2018; Murray et al., 2019). The penalty groups the weights and the bias parameters of each hidden node together and shrinks them to zero simultaneously. Recent empirical studies in Dinh and Ho (2020a,b) suggest that the group Lasso penalty may not be as efficient as the Adaptive group Lasso for selecting between neural network models. In this paper, we propose an Adaptive group Lasso method for the destructive technique and prove that the proposed method is guaranteed to recover the architecture of minimal optimal networks with high probability. We use a simple simulation to illustrate that Adaptive group Lasso may be more advantageous than group Lasso in selecting the number of hidden nodes of networks.

Related work:. Rynkiewicz (2006) proposed an information criterion that can consistently select the number of hidden nodes of a network. Nevertheless, information criterion has little application in practice because it requires to be computed for all possible models. The performance of the destructive technique has been investigated extensively using both synthetic and real data (LeCun et al., 1990; Murray and Chiang, 2015; Alvarez and Salzmann, 2016; Scardapane et al., 2017; Huang et al., 2018). However, little work has been done to investigate theoretically the performance of these methods for neural network models although asymptotic properties of Lasso-type regularization methods have been studied extensively for linear model (Zou, 2006; Zhao and Yu, 2006; Wang and Leng, 2008; Meinshausen et al., 2009; Liu and Zhang, 2009). Notably, there are some recent theoretical works on asymptotic properties of Lasso-type regularization methods for feature selection under neural networks models (Dinh and Ho, 2020b; Feng and Simon, 2017; Farrell et al., 2018; Fallahgoul et al., 2019; Shen et al., 2019).

2. Mathematical framework

Let $\alpha = (u, v, b_1, b_2)$, we denote a network with weights u, v and biases b_1, b_2 by f_α, and the u, v, b_1, b_2 components of α by u_α, v_α, b_{1_α}, b_{2_α} respectively. Recall that a network is *minimal* if it does not have the same input–output map with a network with fewer hidden nodes. Sussmann (1992) provided the following necessary and sufficient conditions for a network f_α to be minimal:

Lemma 2.1. *A network f_α is minimal if and only if*

(i) $u_\alpha^{[:,i]} \neq 0$ *for all* i
(ii) $v_\alpha^{[i]} \neq 0$ *for all* i
(iii) $(u_\alpha^{[:,i]}, b_{1_\alpha}^{[i]}) \neq \pm(u_\alpha^{[:,j]}, b_{1_\alpha}^{[j]})$ *for all* $i \neq j$,

where $u_\alpha^{[:,i]}$ is the ith column of the matrix u_α and $v_\alpha^{[i]}$ is the ith component of the vector v_α.

Moreover, Sussmann (1992) showed that two minimal networks that have same input–output map can be transformed from one to another by a series of sign-flip and node-interchange transformations. A sign-flip transformation changes $(u_\alpha^{[:,i]}, b_{1_\alpha}^{[i]}) \to (-u_\alpha^{[:,i]}, -b_{1_\alpha}^{[i]})$ for a node i while a node-interchange transformation switches the labels of two nodes: $(i, j) \to (j, i)$. Therefore, a set of minimal networks that have the same input–output map is always finite.

We assume that the training data $(Y_k, X_k)_{k=1}^n$ are generated from the following model

$$Y_k = f_{\alpha^*}(X_k) + \epsilon_k \tag{1}$$

where $\{\epsilon_i\}_{k=1}^n$ are i.i.d. random variables that follow the normal distribution $\mathcal{N}(0, \sigma^2)$. Without loss of generality, we assume that the data-generating network f_{α^*} is a minimal network. Let H^* be the unknown number of hidden nodes of f_{α^*}. We consider the destructive techniques proposed by Alvarez and Salzmann (2016) for reconstructing f_{α^*}. Specifically, we fit a large network with H hidden nodes and remove redundant nodes using Adaptive group Lasso. Throughout this paper, we make the following assumption:

Assumption 2.2. We assume that $H^* \leq H$ and $\|\mathbf{vec}(\alpha^*)\|_\infty$ is bounded by a constant W. Here, $\|\cdot\|_\infty$ is the ℓ_∞ norm and $\mathbf{vec}(\cdot)$ is the vectorization operator.

Let $\mathcal{H}$ be the parameter space of all networks that have H hidden nodes such that $\|\mathbf{vec}(\alpha)\|_\infty \leq W$ for all $\alpha \in \mathcal{H}$, and $\mathcal{H}^*$ be the parameter space of all minimal networks that have the same input–output map with f_{α^*}. From now on, "a network in $\mathcal{H}$" means its parameter is in $\mathcal{H}$. A hidden node i is called a *zero node* of f_α if $(u_\alpha^{[:,i]}, v_\alpha^{[i]}, b_{1_\alpha}^{[i]}) = 0$, and it is

called a *non-significant node* of f_α if $u^{[:,i]} = 0$ or $v^{[i]} = 0$. Let $\overline{\mathcal{H}}^*$ be the parameter space of networks in $\mathcal{H}$ such that if we remove all zero nodes of a network in $\overline{\mathcal{H}}^*$, we obtain a network in $\mathcal{H}^*$. Similarly, let $\mathcal{K}$ be the parameter space of networks in $\mathcal{H}$ such that if we remove all non-significant nodes of a network $\mathcal{K}$, we obtain a network in $\mathcal{H}^*$. Finally, let $\mathcal{Q}$ be the parameter space of networks in $\mathcal{H}$ that have the same input–output map with f_{α^*}. It is obvious that $\overline{\mathcal{H}}^* \subset \mathcal{K} \subset \mathcal{Q} \subset \mathcal{H}$. We can think of $\overline{\mathcal{H}}^*$ as an embedding of $\mathcal{H}^*$ into $\mathcal{H}$. Since $\mathcal{H}^*$ is finite, $\overline{\mathcal{H}}^*$ is also finite.

For $\alpha, \beta \in \mathcal{H}$, we define a distance between α and β by $d(\alpha, \beta) = \|\mathbf{vec}(\alpha) - \mathbf{vec}(\beta)\|$, where $\|\cdot\|$ is the ℓ_2 norm. With this notation, we can rigorously define *consistency* and *model selection consistency*.

Definition 2.3. An estimator $\hat{\alpha}$ is consistent if $d(\hat{\alpha}, \overline{\mathcal{H}}^*) \to 0$ in probability.

Definition 2.4. An estimator $\hat{\alpha}$ is model selection consistent if

- $\hat{\alpha}$ is consistent
- for any $\delta > 0$, there exists N_δ such that if $n \geq N_\delta$, the probability that $f_{\hat{\alpha}}$ is a minimal network with H^* non-zero nodes is at least $1 - \delta$.

Next, we introduce the Adaptive group Lasso method for estimating the unknown parameter α. It is a two-step process:

- Step 1: obtain an initial estimator using Group Lasso

$$\hat{\alpha}_n^{GL} = \underset{\alpha \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{k=1}^{n} [Y_k - f_\alpha(X_k)]^2 + \zeta_n \sum_{i=1}^{H} \|w_i\|$$

  where $w_i = (u^{[:,i]}, v^{[i]}, b_1^{[i]})$ is the vector of all parameters that associated with the ith hidden node; $\zeta_n > 0$ is the regularizing parameter.
- Step 2: compute the Adaptive group Lasso estimator

$$\hat{\alpha}_n = \underset{\alpha \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{k=1}^{n} [Y_k - f_\alpha(X_k)]^2 + \lambda_n \sum_{i=1}^{H} \frac{\|w_i\|}{\|w_{i_{\hat{\alpha}_n^{GL}}}\|^\gamma}$$

  where γ is a positive number; $\lambda_n > 0$ is the regularizing parameter; and $w_{i_{\hat{\alpha}_n^{GL}}}$ is the w_i-component of $\hat{\alpha}_n^{GL}$. Here, we use the convention that $0/0 = 0$.

For convenience, we define

$$L(\alpha) = \sum_{i=1}^{H} \|w_i\|, \quad M_n(\alpha) = \sum_{i=1}^{H} \frac{\|w_i\|}{\|w_{i_{\hat{\alpha}_n^{GL}}}\|^\gamma}.$$

Next, we provide some basic lemmas that we need for our proofs. To ease the notations, we will use C_k, $k = 1, 2, \ldots$, for denoting generic constants.

Lemma 2.5. *We have* $|L(\alpha) - L(\beta)| \leq \sqrt{H} d(\alpha, \beta), \quad \forall \alpha \in \mathcal{H}$.

Lemma 2.6. *There exists* $C_2 > 0$ *such that for any* $\beta \in \mathcal{Q}$, *there exists* $\beta^* \in \overline{\mathcal{H}}^*$ *such that* $L(\beta) - L(\beta^*) \geq \min\left\{C_2, d(\beta, \overline{\mathcal{H}}^*)\right\}$.

Lemma 2.7. *There exists a constant* $C_1 > 0$ *such that for any* $\beta \in \mathcal{Q}$, *if the set* $\mathcal{U} = \{i : \|w_{i_\beta}\| < C_1\}$ *has* $H - H^*$ *nodes, then the network* β', *obtained by setting the weights and the biases of* β *at nodes in* $\mathcal{U}$ *to zero, belongs to* $\overline{\mathcal{H}}^*$.

The proofs of these lemmas can be found in the Supplementary.

3. Asymptotic properties of Adaptive group Lasso

Let $R_n(\alpha)$ and $R(\alpha)$ be the empirical risk and expected risk of f_α respectively. That is, $R_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - f_\alpha(X_i)]^2$ and $R(\alpha) = E([Y - f_\alpha(X)]^2)$. Note that $R(\alpha) = \sigma^2 = \min_{\beta \in \mathcal{H}} R(\beta)$ if and only if $\alpha \in \mathcal{Q}$.

3.1. Properties of risks

First, we state some properties of $R_n(\alpha)$ and $R(\alpha)$, which are proved in the Supplementary.

Lemma 3.1 (*Generalization Bound*)**.** *For any* $\delta > 0$, *there exist* $C_3(\delta) > 0$

$$|R_n(\alpha) - R(\alpha)| \leq C_3 \frac{\log n}{\sqrt{n}}, \quad \forall \alpha \in \mathcal{H}$$

with probability at least $1 - \delta$.

Lemma 3.2. *There exist $C_4, \nu > 0$ such that*

$$R(\alpha) - \sigma^2 \geq C_4 d(\alpha, \mathcal{Q})^\nu, \quad \forall \alpha \in \mathcal{H}.$$

Lemma 3.3 (*Lipschitzness*)**.** *There exists $C_5 > 0$ such that*

$$|R(\alpha) - R(\beta)| \leq C_5 d(\alpha, \beta), \quad \forall \alpha, \beta \in \mathcal{H}.$$

For any $\delta > 0$, there exists $C_6(\delta)$ such that

$$|R_n(\alpha) - R_n(\beta)| \leq C_6 d(\alpha, \beta), \quad \forall \alpha, \beta \in \mathcal{H}$$

with probability at least $1 - \delta$.

3.2. Model selection consistency of Adaptive group Lasso

Now, we are ready to prove that Adaptive group Lasso is structural consistent. The first step is deriving the convergence rate of group Lasso.

Theorem 3.4. *Assume that $\zeta_n \to 0$ and $\zeta_n\sqrt{n}/\log(n) \to \infty$. For any $\delta > 0$, when n is sufficiently large, we have*

$$d(\hat{\alpha}_n^{GL}, \overline{\mathcal{H}}^*) \leq 2C_3 \frac{\log n}{\zeta_n \sqrt{n}} + (1 + \sqrt{H}) \left(\frac{4C_3}{C_4} \frac{\log n}{\sqrt{n}} + \frac{2C_7}{C_4} \zeta_n^{\nu/(\nu-1)} \right)^{1/\nu}$$

with probability at least $1 - \delta$.

Proof. Please see the detailed proof in the Supplementary. □

A direct consequence of Theorem 3.4 is that the group Lasso estimator is consistent. In the second step, we prove that the Adaptive group Lasso estimate has at most H^* non-zero nodes.

Theorem 3.5. *Assume that $\zeta_n \to 0$, $\zeta_n\sqrt{n}/\log(n) \to \infty$, $\lambda_n \to 0$, and*

$$\frac{\lambda_n^{1/\gamma}}{2C_3 \frac{\log n}{\zeta_n \sqrt{n}} + \sqrt{H} \left(\frac{4C_3}{C_4} \frac{\log n}{\sqrt{n}} + \frac{2C_7}{C_4} \zeta_n^{\nu/(\nu-1)} \right)^{1/\nu}} \to \infty.$$

For any $\delta > 0$, when n is sufficiently large, $f_{\hat{\alpha}_n}$ has at most H^* non-zero nodes with probability at least $1 - \delta$.

Proof. Let $\alpha_n^* \in \overline{\mathcal{H}}^*$ such that $d(\hat{\alpha}_n^{GL}, \alpha_n^*) = d(\hat{\alpha}_n^{GL}, \overline{\mathcal{H}}^*)$. We denote the set of zero nodes of $f_{\alpha_n^*}$ by $\mathcal{I}_n$. Let $\pi_n : \mathcal{H} \to \mathcal{H}$ be a function that set all the weights and biases associated with nodes in $\mathcal{I}_n$ to 0. That is,

$$w_{i_{\pi_n(\alpha)}} = (u^{[:,i]}_{\pi_n(\alpha)}, v^{[i]}_{\pi_n(\alpha)}, b^{[i]}_{1_{\pi_n(\alpha)}}) = \begin{cases} 0 & \text{if } i \in \mathcal{I}_n \\ (u_\alpha^{[:,i]}, v_\alpha^{[i]}, b_{1_\alpha}^{[i]}) & \text{if } i \notin \mathcal{I}_n. \end{cases}$$

We have,

$$\begin{aligned} \sum_{i \in \mathcal{I}_n} \|w_{i_{\hat{\alpha}_n^{GL}}}\| &\leq \frac{R_n(\alpha_n^*) - R_n(\hat{\alpha}_n^{GL})}{\zeta_n} + \sum_{i \notin \mathcal{I}_n} \left(\|w_{i_{\alpha_n^*}}\| - \|w_{i_{\hat{\alpha}_n^{GL}}}\| \right) \\ &\leq 2C_3 \frac{\log n}{\zeta_n \sqrt{n}} + \frac{R(\alpha_n^*) - R(\hat{\alpha}_n^{GL})}{\zeta_n} + \sum_{i \notin \mathcal{I}_n} \left| \|w_{i_{\alpha_n^*}}\| - \|w_{i_{\hat{\alpha}_n^{GL}}}\| \right| \qquad (2) \\ &\leq C_3 \frac{\log n}{\zeta_n \sqrt{n}} + \sum_{i \notin \mathcal{I}_n} \left| \|w_{i_{\alpha_n^*}}\| - \|w_{i_{\hat{\alpha}_n^{GL}}}\| \right| \end{aligned}$$

with probability at least $1 - \delta$.

On the other hand, $d(\hat{\alpha}_n^{GL}, \pi_n(\hat{\alpha}_n^{GL})) \geq \|w_{i_{\hat{\alpha}_n^{GL}}}\|$ for all $i \in \mathcal{I}_n$. So,

$$\lambda_n \sum_{i \in \mathcal{I}_n} \frac{\|w_{i_{\hat{\alpha}_n}}\|}{d(\hat{\alpha}_n^{GL}, \pi_n(\hat{\alpha}_n^{GL}))^\gamma} \leq \lambda_n \sum_{i \in \mathcal{I}_n} \frac{\|w_{i_{\hat{\alpha}_n}}\|}{\|w_{i_{\hat{\alpha}_n^{GL}}}\|^\gamma} \leq R_n(\pi_n(\hat{\alpha}_n)) - R_n(\hat{\alpha}_n). \qquad (3)$$

By Lemma 3.3, we obtain

$$R_n(\pi_n(\hat{\alpha}_n)) - R_n(\hat{\alpha}_n) \leq C_6 d(\hat{\alpha}_n, \pi_n(\hat{\alpha}_n)) \leq C_6 \sum_{i \in \mathcal{I}_n} \|w_{i_{\hat{\alpha}_n}}\| \qquad (4)$$

with probability at least $1 - \delta$.

Assume that $\sum_{i\in\mathcal{I}_n}\|w_{i_{\hat{\alpha}_n}}\|>0$. From (2) and Theorem 3.4, we have

$$\begin{aligned} d(\hat{\alpha}_n^{GL},\pi_n(\hat{\alpha}_n^{GL})) &\le \sum_{i\in\mathcal{I}_n}\|w_{i_{\hat{\alpha}_n^{GL}}}\| \le 2C_3\frac{\log n}{\zeta_n\sqrt{n}} + \sum_{i\notin\mathcal{I}_n}\left|\|w_{i_{\alpha_n^*}}\| - \|w_{i_{\hat{\alpha}_n^{GL}}}\|\right| \\ &\le 2C_3\frac{\log n}{\zeta_n\sqrt{n}} + \sqrt{H}d(\hat{\alpha}_n^{GL},\alpha_n^*) = C_3\frac{\log n}{\zeta_n\sqrt{n}} + \sqrt{H}d(\hat{\alpha}_n^{GL},\overline{\mathcal{H}}^*) \\ &\le 2C_3(1+\sqrt{H})\frac{\log n}{\zeta_n\sqrt{n}} + \sqrt{H}(1+\sqrt{H})\left(\frac{4C_3}{C_4}\frac{\log n}{\sqrt{n}} + \frac{2C_7}{C_4}\zeta_n^{\nu/(\nu-1)}\right)^{1/\nu} \end{aligned} \tag{5}$$

with probability at least $1-\delta$.

Combining (3), (4), and (5), we get

$$\frac{\lambda_n^{1/\gamma}}{2C_3\frac{\log n}{\zeta_n\sqrt{n}} + \sqrt{H}\left(\frac{4C_3}{C_4}\frac{\log n}{\sqrt{n}} + \frac{2C_7}{C_4}\zeta_n^{\nu/(\nu-1)}\right)^{1/\nu}} \le (1+\sqrt{H})C_6^{1/\gamma}$$

which is a contradiction when n is large enough. Therefore, $\sum_{i\in\mathcal{I}_n}\|w_{i_{\hat{\alpha}_n}}\|=0$ with probability at least $1-\delta$. Hence, every node in $\mathcal{I}_n$ is a zero node of $\hat{\alpha}_n$. Recall that $\mathcal{I}_n$ is the set of zero nodes of $f_{\alpha_n^*}$ where $\alpha_n^*\in\overline{\mathcal{H}}^*$. So, $|\mathcal{I}_n| = H - H^*$. That is, $f_{\hat{\alpha}_n}$ has at most H^* non-zero nodes □

Finally, we derive the convergence rate of the Adaptive group Lasso.

Theorem 3.6. *Assume that* $\zeta_n\to 0$, $\zeta_n\sqrt{n}/\log(n)\to\infty$, $\lambda_n\to 0$, *and*

$$\frac{\lambda_n^{1/\gamma}}{2C_3\frac{\log n}{\zeta_n\sqrt{n}} + \sqrt{H}\left(\frac{4C_3}{C_4}\frac{\log n}{\sqrt{n}} + \frac{2C_7}{C_4}\zeta_n^{\nu/(\nu-1)}\right)^{1/\nu}} \to \infty.$$

For any $\delta>0$, *when* n *is sufficiently large, we have*

$$d(\hat{\alpha}_n,\overline{\mathcal{H}}^*) \le \left(\frac{2C_3}{C_4}\frac{\log n}{\sqrt{n}} + \frac{C_8}{C_4}\lambda_n\right)^{1/\nu}.$$

with probability at least $1-\delta$.

Proof. Let $\alpha_n^*\in\overline{\mathcal{H}}^*$ such that $d(\hat{\alpha}_n^{GL},\alpha_n^*) = d(\hat{\alpha}_n^{GL},\overline{\mathcal{H}}^*)$. By Lemma 3.1, we have, for $\delta>0$,

$$\begin{aligned} R(\hat{\alpha}_n) - R(\alpha_n^*) &\le R_n(\hat{\alpha}_n) - R_n(\alpha_n^*) + 2C_3\frac{\log n}{\sqrt{n}} \\ &\le 2C_3\frac{\log n}{\sqrt{n}} + \lambda_n[M_n(\alpha_n^*) - M_n(\hat{\alpha}_n)] \\ &\le 2C_3\frac{\log n}{\sqrt{n}} + \lambda_n M_n(\alpha_n^*). \end{aligned}$$

with probability at least $1-\delta$.

Since $\overline{\mathcal{H}}^*$ is finite, by Theorem 3.4, there exists C_8 such that when n is sufficiently large, $M_n(\alpha_n^*)\le C_8$ with probability at least $1-\delta$. By Lemma 3.2, when n is sufficiently large

$$C_4 d(\hat{\alpha}_n,\mathcal{Q})^\nu \le R(\hat{\alpha}_n) - R(\alpha_n^*) \le 2C_3\frac{\log n}{\sqrt{n}} + C_8\lambda_n$$

with probability at least $1-\delta$.

Let $\xi_n\in\mathcal{Q}$ such that $d(\hat{\alpha}_n,\xi_n) = d(\hat{\alpha}_n,\mathcal{Q})$. We now prove that $\xi_n\in\overline{\mathcal{H}}^*$ when n is sufficiently large. Let $\mathcal{I}_n$ be the set of zero node of $f_{\alpha_n^*}$. In the proof of Theorem 3.5, we prove that nodes in $\mathcal{I}_n$ are zero nodes of $f_{\hat{\alpha}_n}$ when n is sufficiently large. Therefore,

$$\|w_{i_{\xi_n}}\| \le d(\hat{\alpha}_n,\xi_n) = d(\hat{\alpha}_n,\mathcal{Q}) \le \left(\frac{2C_3}{C_4}\frac{\log n}{\sqrt{n}} + \frac{C_8}{C_4}\lambda_n\right)^{1/\nu} \tag{6}$$

for all $i\in\mathcal{I}_n$.

Consider $\xi_n'\in\mathcal{H}$ such that

$$w_{i_{\xi_n'}} = \begin{cases} w_{i_{\xi_n}} & \text{if } i\notin\mathcal{I}_n \\ 0 & \text{if } i\in\mathcal{I}_n. \end{cases}$$

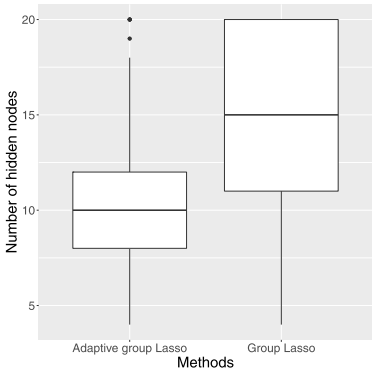

Fig. 1. Number of hidden nodes selected by our proposed Adaptive Group Lasso method and the Group Lasso method proposed in Murray and Chiang (2015). The number of hidden nodes of the data-generating network is 10.

By Lemma 2.7 and (6), when n is large enough, we have $\xi_n' \in \overline{\mathcal{H}}^*$ with probability at least $1-\delta$. Note that $w_{i_{\hat{\alpha}_n}} = 0$ for all $i \in \mathcal{I}_n$. So,

$$d(\hat{\alpha}_n, \mathcal{Q}) \leq d(\hat{\alpha}_n, \xi_n') \leq d(\hat{\alpha}_n, \xi_n) = d(\hat{\alpha}_n, \mathcal{Q}).$$

Thus, $w_{i_{\xi_n}} = 0$ for all $i \in \mathcal{I}_n$. That is, we have $\xi_n \in \overline{\mathcal{H}}^*$. Therefore,

$$d(\hat{\alpha}_n, \overline{\mathcal{H}}^*) = d(\hat{\alpha}_n, \mathcal{Q}) \leq \left(\frac{2C_3}{C_4}\frac{\log n}{\sqrt{n}} + \frac{C_8}{C_4}\lambda_n\right)^{1/\nu}. \quad \square$$

Combining Theorems 3.5 and 3.6 we can easily show that Adaptive group Lasso estimator is model selection consistent with appropriate choice of regularizer parameters ζ_n and λ_n. Specifically,

Theorem 3.7. *If* $\zeta_n \to 0$, $\zeta_n\sqrt{n}/\log(n) \to \infty$, $\lambda_n \to 0$, *and*

$$\frac{\lambda_n^{1/\gamma}}{2C_3\frac{\log n}{\zeta_n\sqrt{n}} + \sqrt{H}\left(\frac{4C_3}{C_4}\frac{\log n}{\sqrt{n}} + \frac{2C_7}{C_4}\zeta_n^{\nu/(\nu-1)}\right)^{1/\nu}} \to \infty,$$

then the Adaptive group Lasso estimator is model selection consistent.

Proof. Consistency comes directly from Theorem 3.6. By Theorem 3.5, the Adaptive group Lasso has at most H^* non-zero nodes. When n is sufficiently large, Lemma 2.1 guarantees that $f_{\hat{\alpha}_n}$ is a minimal network. So, it has exactly H^* non-zero nodes. $\square$

4. Illustration

The goal of this section is not to provide an extensive study on the performance of the destructive technique. This has been done thoroughly in LeCun et al. (1990), Murray and Chiang (2015), Alvarez and Salzmann (2016), Scardapane et al. (2017), Huang et al. (2018) using both synthetic and real data. Instead, we aim to illustrate our theoretical findings through simple experiments.

4.1. Simulation

In this experiment, we simulate 100 datasets of size $n = 5000$ according to the model (1) with $H^* = 10$, $\sigma^2 = 1$. The inputs (5 features) and parameters of the model are drawn independently from the standard normal distribution. For each dataset, we train a network with $H = 20$ over 10 000 epochs using our proposed Adaptive Group Lasso method and the Group Lasso method proposed in Murray and Chiang (2015), Murray et al. (2019). The regularizing constants of both methods are chosen from the set $\{0.001, 0.005, 0.01, 0.025, 0.05, 0.075, 0.1\}$ using the Akaike information criterion (AIC). Our optimization method is Proximal gradient method (Parikh and Boyd, 2014) (the learning rate is 0.01), which can identify the support of the estimates directly without the need of thresholding. For the Adaptive group Lasso, we choose $\gamma = 2$. The simulation is implemented in Python using Pytorch library.

We count the number of hidden nodes selected by each method. Fig. 1 summarizes the results of our simulation. The Adaptive Group Lasso performs better than the group Lasso in choosing the size of a network. Note that the best number of hidden nodes is 10.

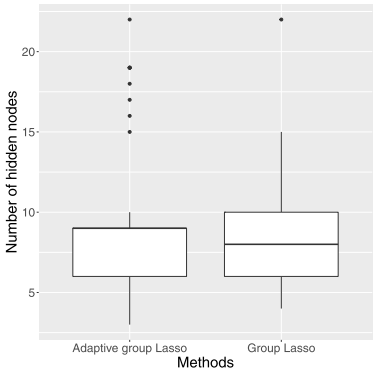

Fig. 2. Number of hidden nodes selected by our proposed Adaptive Group Lasso method and the group Lasso method proposed in Murray and Chiang (2015) (Boston housing dataset).

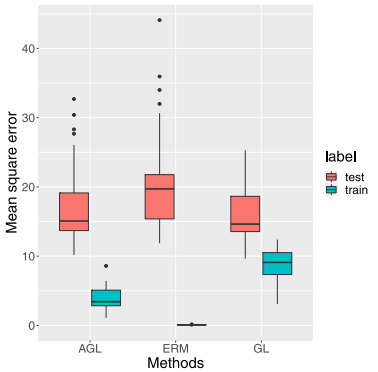

Fig. 3. The training and testing errors of the Adaptive Group Lasso methods and the group Lasso method proposed in Murray and Chiang (2015) (Boston housing dataset).

4.2. Boston housing dataset

We use our framework to study the Boston housing dataset.[2] This dataset consists of 506 observations of house prices and 13 predictors. We consider a network with 50 hidden nodes. The group Lasso and Adaptive group Lasso methods are then performed on this dataset using average test errors from 50 random train-test splits (with the size of the test sets being 25% of the original dataset) over 10 000 epochs. The regularizing constants of the algorithms are chosen from the set $\{0.1, 0.3, 0.5, 0.7, 1\}$ using the Akaike information criterion (AIC). As in the previous part, we use the proximal gradient method (with learning rate 0.01) for optimization and choose $\gamma = 2$ for the Adaptive group Lasso. We also consider the simple Empirical risk minimizer (ERM) in this experiment.

The number of hidden nodes selected by group Lasso and Adaptive group Lasso are presented in Fig. 2. Although the destructive methods choose much smaller networks (about one-fifth the size of the full networks), their prediction errors are slightly better than the ERM which uses the full network (see Fig. 3). The gap between training error and testing error of the destructive methods is also smaller compared to the ERM.

5. Discussion and conclusion

We prove that our proposed Adaptive group Lasso method is model selection consistent for the problem of selecting the number of hidden nodes of one-hidden-layer feedforward networks. To the best of our knowledge, this is the first theoretical result for the popular destructive technique. We also obtain the consistency of the group Lasso method as a byproduct of our proof. However, the question about the model selection consistency of the group Lasso estimator remains open. One interesting direction for future work is extending our results to deep neural networks. This requires further investigation on the properties of minimal deep neural networks. Another avenue for future direction is developing theory and methods for applying constructive and destructive approaches to select the number of layers of deep neural networks.

2 http://lib.stat.cmu.edu/datasets/boston.

In this paper, we assume that the true underlying function is a neural network model (Eq. (1)). The extension from model-based framework to the general cases with model mismatch is an intriguing question in learning with neural networks. In general, the projections of the true underlying function to the hypothesis space (in ℓ_2 distance for regression) might not be unique, and they might not be similar to each other in terms of the structure of interest. Understanding of these projections for neural networks is limited, and analyses of the general cases need to involve imposing certain strong conditions on them (Feng and Simon, 2017). For our problem of structure reconstruction, one possible set of conditions are: (1) the set of optimal projections (in function space) is finite, and (2) all optimal projections have the same number of hidden nodes. Our proofs can be adapted to this setting with minor adjustments.

Acknowledgments

LSTH was supported by startup funds from Dalhousie University, Canada, the Canada Research Chairs program, the NSERC Discovery Grant, Canada RGPIN-2018-05447, and the NSERC Discovery Launch Supplement, Canada DGECR-2018-00181. VD was supported by a startup fund from University of Delaware and National Science Foundation, USA grant DMS-1951474.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2021.109353.

References

Alvarez, J.M., Salzmann, M., 2016. Learning the number of neurons in deep networks. In: Advances in Neural Information Processing Systems. pp. 2270–2278.

Bello, M.G., 1992. Enhanced training algorithms, and integrated training/architecture selection for multilayer perceptron networks. IEEE Trans. Neural Netw. 3 (6), 864–875.

Dinh, V., Ho, L.S.T., 2020a. Consistent feature selection for neural networks via Adaptive Group Lasso. arXiv preprint arXiv:2006.00334.

Dinh, V., Ho, L.S.T., 2020b. Consistent feature selection for analytic deep neural networks. In: Advances in Neural Information Processing Systems.

Fallahgoul, H., Franstianto, V., Loeper, G., 2019. Towards explaining the ReLU feed-forward network. Available At SSRN.

Farrell, M.H., Liang, T., Misra, S., 2018. Deep neural networks for estimation and inference. arXiv preprint arXiv:1809.09953.

Feng, J., Simon, N., 2017. Sparse-input neural networks for high-dimensional nonparametric regression and classification. arXiv preprint arXiv:1711.07592.

Huang, G., Liu, S., Van der Maaten, L., Weinberger, K.Q., 2018. Condensenet: An efficient densenet using learned group convolutions, In: Proceedings of The IEEE Conference on Computer Vision And Pattern Recognition, pp. 2752–2761.

LeCun, Y., Denker, J.S., Solla, S.A., Howard, R.E., Jackel, L.D., 1990. Optimal brain damage. In: Advances In Neural Information Processing Systems, vol. 2, pp. 598–605.

Liu, H., Zhang, J., 2009. Estimation consistency of the group lasso and its applications. In: Artificial Intelligence and Statistics. pp. 376–383.

Meinshausen, N., Yu, B., et al., 2009. Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist. 37 (1), 246–270.

Murray, K., Chiang, D., 2015. Auto-Sizing Neural Networks: With Applications to n-gram Language Models. In: Proceedings Of The 2015 Conference On Empirical Methods In Natural Language Processing, pp. 908–916.

Murray, K., Kinnison, J., Nguyen, T.Q., Scheirer, W., Chiang, D., 2019. Auto-Sizing the Transformer Network: Improving Speed, Efficiency, and Performance for Low-Resource Machine Translation. In: Proceedings of The Third Workshop on Neural Generation and Translation.

Parikh, N., Boyd, S., 2014. Proximal algorithms. Found. Trends Optim. 1 (3), 127–239.

Rynkiewicz, J., 2006. Consistent estimation of the architecture of multilayer perceptrons. In: European Symposium on Artificial Neural Networks. pp. 149–154.

Scardapane, S., Comminiello, D., Hussain, A., Uncini, A., 2017. Group sparse regularization for deep neural networks. Neurocomputing 241, 81–89.

Shen, X., Jiang, C., Sakhanenko, L., Lu, Q., 2019. Asymptotic properties of neural network sieve estimators. arXiv preprint arXiv:1906.00875.

Sussmann, H.J., 1992. Uniqueness of the weights for minimal feedforward nets with a given input-output map. Neural Netw. 5 (4), 589–593.

Wang, H., Leng, C., 2008. A note on Adaptive Group Lasso. Comput. Statist. Data Anal. 52 (12), 5277–5286.

Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7 (Nov), 2541–2563.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101 (476), 1418–1429.