The Canadian Journal of Statistics
Vol. 50, No. 1, 2022, Pages 59–85
La revue canadienne de statistique

A nonlinear sparse neural ordinary differential equation model for multiple functional processes

Yijia LIU[1], Lexin LI[2], and Xiao WANG[1]*

[1]*Department of Statistics, Purdue University, West Lafayette, IN, USA*
[2]*Department of Biostatistics and Epidemiology, University of California at Berkeley, Berkeley, CA, USA*

Key words and phrases: Deep neural networks; multivariate functions; nonconvex optimization; ordinary differential equation; ℓ_0-penalty.

MSC 2020: Primary 62R10; secondary 62G05.

Abstract: In this article, we propose a new sparse neural ordinary differential equation (ODE) model to characterize flexible relations among multiple functional processes. We characterize the latent states of the functions via a set of ODEs. We then model the dynamic changes of the latent states using a deep neural network (DNN) with a specially designed architecture and a sparsity-inducing regularization. The new model is able to capture both nonlinear and sparse-dependent relations among multivariate functions. We develop an efficient optimization algorithm to estimate the unknown weights for the DNN under the sparsity constraint. We establish both the algorithmic convergence and selection consistency, which constitute the theoretical guarantees of the proposed method. We illustrate the efficacy of the method through simulations and a gene regulatory network example. *The Canadian Journal of Statistics* 50: 59–85; 2022 © 2021 Statistical Society of Canada

Résumé: Afin de caractériser des relations flexibles entre plusieurs processus fonctionnels, les auteurs de cet article proposent un nouveau modèle d'équations différentielles ordinaires (EDO) neuronales éparses. Dans un premier temps, ils commencent par caractériser les états latents des fonctions via un ensemble d'équations différentielles ordinaires, pour ensuite modéliser les changements dynamiques d'états latents en utilisant un réseau neuronal profond (RNP) avec une architecture spécialement conçue et une régularisation induisant l'éparpillement. Le nouveau modèle est capable de capturer à la fois des relations non linéaires et des relations de dépendance éparse entre des fonctions multivariées. Un algorithme d'optimisation efficace pour estimer les poids inconnus des RNP sous contraintes d'éparpillement est également proposé. Les auteurs établissent les convergences algorithmique et de la sélection qui témoignent du bon comportement théorique de la méthode proposée. Enfin, l'efficacité de la méthode est illustrée à l'aide de simulations numériques et un exemple de réseaux de régulation génétique. *La revue canadienne de statistique* 50: 59–85; 2022 © 2021 Société statistique du Canada

1. INTRODUCTION

Ordinary differential equations (ODEs) have been widely used to model dynamical systems in science and engineering applications. Examples include neuroscience (Izhikevich, 2006), genomics (Chou & Voit, 2009), chemical engineering (Biegler, Damiano & Blau, 1986), and infectious diseases (Wu, 2005), among many others. An ODE model involves a system of

* *Corresponding author: wangxiao@purdue.edu*

differential equations that link the functions and their derivatives. Moreover, the system is usually observed on a set of discrete time points with measurement error.

There have been a number of pioneering works studying statistical modelling of ODEs. However, most existing solutions constrain the forms of functional relations in the system. Particularly, Lu et al. (2011) studied a system of linear ODEs. Zhang et al. (2015) extended the linear ODEs by including the two-way interactions. Dattner & Klaassen (2015) studied a generalized linear form of ODEs, but without interactions. Ramsay et al. (2007), Henderson & Michailidis (2014), Wu et al. (2014), and Chen, Shojaie & Witten (2017) all considered an additive ODE model and used spline basis expansion to incorporate possible nonlinear effects. Dai & Li (2021) recently developed a reproducing kernel version of a nonlinear ODE model through smoothing spline analysis of variance.

In the past decade, deep neural networks (DNNs) have demonstrated outstanding performance in numerous challenging tasks such as image recognition and natural language processing (Goodfellow et al., 2016). In this article, we employ DNNs to develop a highly flexible nonlinear ODE model. Specifically, we propose a new sparse neural ODE (SNODE) model to estimate and uncover the possibly nonlinear structure of an ODE system from noisy observations. We approximate the unknown and possibly nonlinear effects in ODEs by a DNN with a specially designed architecture. Moreover, we adopt the strategy of Chen et al. (2020) to incorporate a sparsity regularization to the DNN, which helps to produce a sparse estimate of the ODE system and improves the interpretability. The sparsity structure is scientifically motivated, and has been commonly adopted in numerous applications. For instance, Gardner et al. (2003) and Cai, Bazerque & Giannakis (2013) have advocated that gene regulatory networks and various other biochemical networks are sparse, and Zhang et al. (2015) have demonstrated that connectivity networks in the brain are also sparse. Such tendency towards sparsity may be due to the fact that connections consume energy, and biological units tend to minimize energy-consuming activities (Bullmore & Sporns, 2009). We then develop an efficient optimization algorithm that integrates a DNN-based ODE solver (Chen et al., 2018) with sparsity estimation. We assess the theoretical performance of the proposed method by studying the algorithmic convergence and by establishing that selection consistency is achieved when the objective function satisfies certain regularity conditions. Finally, we investigate empirical performance, numerically compare with a number of alternative ODE solutions, and review the strengths and limitations of our method.

The rest of the article is organized as follows. Section 2 introduces the ODE system and the sparse DNN architecture. Section 3 presents the optimization algorithm. Section 4 establishes the theoretical guarantees. Section 5 presents the simulations, and Section 6 gives an analysis of a gene network dataset. Section 7 concludes the paper. The Appendix collects all technical proofs.

2. MODEL

We first present a general ODE system. We then introduce a sparse DNN architecture to model the ODE system.

2.1. The ODE System

Let $\boldsymbol{x}(t) = \big(x_1(t), \ldots, x_p(t)\big)A^T : \mathcal{D} \to \mathbb{R}^p$ denote p smooth functional processes over a compact domain $\mathcal{D} \subset \mathbb{R}$. For instance, $\boldsymbol{x}(t)$ can denote the expression levels of p genes, or the neuronal states of p brain regions, at time t. Let $\boldsymbol{u}(t) = \big(u_1(t), \ldots, u_q(t)\big)^T : \mathcal{D} \to \mathbb{R}^q$ denote q experimental input functions, e.g., some stimulus functions. The ODE model characterizes the dynamic changes of $\boldsymbol{x}(t)$ starting from some initial state $\boldsymbol{x}(0) = \boldsymbol{x}_0$ under the influence of experimental inputs $\boldsymbol{u}(t)$ by

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{f}\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right), \tag{1}$$

 DOI: 10.1002/cjs.11666

where $\boldsymbol{f} = \left(f_1, \ldots, f_p\right)^T : \mathbb{R}^{p+q} \to \mathbb{R}^p$ is a set of unknown, possibly nonlinear functions, and f_i characterizes the influences exerted by the present states $x_1(t), \ldots, x_p(t)$ and the experimental input $u_1(t), \ldots, u_q(t)$ on the ith process $x_i(t)$, $i = 1, \ldots, p$. Model (1) is deterministic, as we treat $\boldsymbol{f}$ as fixed functions.

Moreover, we postulate that f_i is a nonlinear sparse function that only depends on some of the x_j, or equivalently, $x_i(t)$ is affected by only a subset of the $x_j(t)$, $i, j = 1, \ldots, p$. That is, we assume $f_i \in \mathcal{F}_{s_i^*}$, where

$$\mathcal{F}_{s_i^*} = \left\{ f : \mathbb{R}^{p+q} \to \mathbb{R} : \exists\, \bar{f}\left(\boldsymbol{x}_{[S_i]}, \boldsymbol{u}\right) = f(\boldsymbol{x}, \boldsymbol{u}), \boldsymbol{x}_{[S_i]} \in \mathbb{R}^{s_i^*}, \forall\, \boldsymbol{x} \in \mathbb{R}^p, \bar{f} \text{ is Lipschitz} \right\},$$

$\boldsymbol{x}_{[S_i]}$ represents the subvector of $\boldsymbol{x}$ only including those entries with indices in $S_i \subseteq \{1, 2, \ldots, p\}$, and the cardinality $|S_i| = s_i^*$, which reflects the true sparsity.

Typically, the ODE system (1) is observed on a finite set of discrete time points $\{t_1, \ldots, t_n\}$ with additional measurement error

$$\boldsymbol{y}(t_k) = \boldsymbol{x}(t_k) + \boldsymbol{\varepsilon}(t_k), \quad k = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\varepsilon}(t) = \left(\varepsilon_1(t), \ldots, \varepsilon_p(t)\right)^T : \mathcal{D} \to \mathbb{R}^p$ denotes the p-dimensional zero-mean measurement error process. For simplicity, we assume $\boldsymbol{\varepsilon}(t)$ is a white noise process, but it is possible to consider auto-correlated errors as well.

2.2. Sparse DNN Architecture

We propose to approximate the functional $\boldsymbol{f}$ by a DNN with a special architecture, as displayed in Figure 1. The network consists of two key parts: a selection layer, and a sequence of approximation layers.

Specifically, for each function f_i, $i = 1, \ldots, p$, the selection layer involves two sets of parameters: the weights $\boldsymbol{w}_i = \left(w_{1i}, \ldots, w_{pi}\right)^T \in \mathbb{R}^p$ that reflect the influences of the functional process $\boldsymbol{x}(t)$, and the weights $\boldsymbol{\theta}_i^u = \left(\theta_{1i}^u, \ldots, \theta_{qi}^u\right)^T \in \mathbb{R}^q$ that reflect the influences of the stimulus function $\boldsymbol{u}(t)$. This selection layer transforms the input $(\boldsymbol{x}, \boldsymbol{u}) = (x_1, \ldots, x_p, \boldsymbol{u})^T \in \mathbb{R}^{p+q}$ into

$$\left(\boldsymbol{w}_i, \boldsymbol{\theta}_i^u\right) \odot (\boldsymbol{x}, \boldsymbol{u}) = \left(w_{1i}x_1, \ldots, w_{pi}x_p, \langle \boldsymbol{\theta}_i^u, \boldsymbol{u} \rangle\right)^T.$$

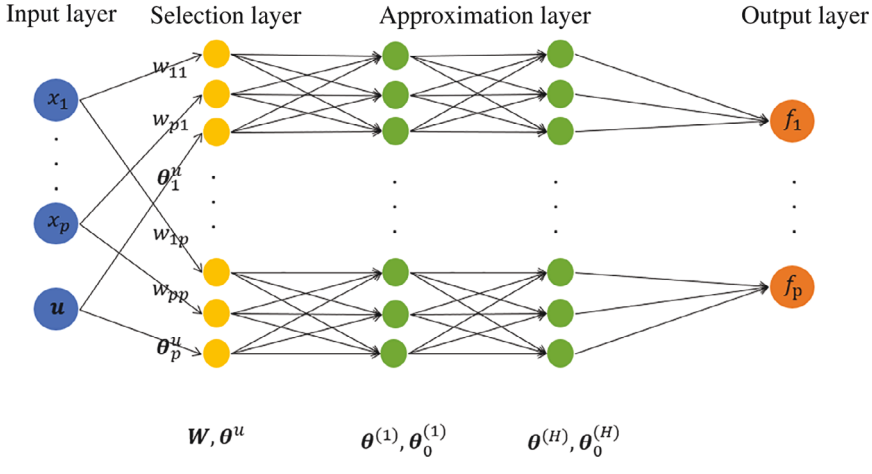

FIGURE 1: Architecture of the proposed deep neural networks for the sparse neural ODE.

We then impose a sparsity constraint on the weight $\boldsymbol{w}_i$, such that $\|\boldsymbol{w}_i\|_0 = s_i$, where $\|\cdot\|_0$ denotes the ℓ_0 norm and equals the number of nonzero elements in this vector, and s_i is the working sparsity parameter, $i = 1, \ldots, p$. Such a constraint allows us to select the individual functions x_j that influence x_i.

Next, for each function f_i, $i = 1, \ldots, p$, the approximation layer contains a number of fully connected neural network layers. Let H be the number of approximation layers, $\boldsymbol{z}_{h-1} \in \mathbb{R}^{n_{h-1}}$ be the weighted input to the neurons in the hth approximation layer, and n_h be the number of neurons in the hth approximation layer, $h = 1, \ldots, H$. We approximate f_i by

$$g_i(\boldsymbol{x}) = T_H \circ \sigma \circ T_{H-1} \circ \sigma \circ \ldots \sigma \circ T_1 \circ \left(\left(\boldsymbol{w}_i, \boldsymbol{\theta}_i^u \right) \odot (\boldsymbol{x}, \boldsymbol{u}) \right),$$

where $T_h\left(\boldsymbol{z}_{h-1}\right) = \boldsymbol{\theta}^{(h)} \boldsymbol{z}_{h-1} + \boldsymbol{\theta}_0^{(h)}$ is an affine transformation with the parameters $\boldsymbol{\theta}^{(h)} \in \mathbb{R}^{n_h \times n_{h-1}}$ and $\boldsymbol{\theta}_0^{(h)} \in \mathbb{R}^{n_h}$, $h = 1, \ldots, H$, $n_0 = p^2 + pq$, and $\sigma(\cdot)$ is the activation function. Some common choices of the activation function include the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$, the tanh function $\sigma(x) = \tanh(x)$, and the ReLU function $\sigma(x) = \max(0, x)$.

Let $\mathrm{vec}(\cdot)$ denote the operator that vectorizes a matrix. Write

$$\begin{aligned} \boldsymbol{W} &= \left(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p \right), \\ \boldsymbol{\Theta} &= \left(\left(\boldsymbol{\theta}_1^u\right)^T, \ldots, \left(\boldsymbol{\theta}_p^u\right)^T, \mathrm{vec}\left(\boldsymbol{\theta}^{(1)}\right)^T, \left(\boldsymbol{\theta}_0^{(1)}\right)^T, \ldots, \mathrm{vec}\left(\boldsymbol{\theta}^{(H)}\right)^T, \left(\boldsymbol{\theta}_0^{(H)}\right)^T \right)^T, \end{aligned} \tag{3}$$

such that $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ collects all the parameters in the selection layer, and $\boldsymbol{\Theta} \in \mathbb{R}^d$, with $d = pq + \sum_{h=1}^{H} n_h n_{h-1} + n_h$, collects all the parameters for the stimulus function $\boldsymbol{u}$ and in the set of approximation layers. Let $\boldsymbol{s} = (s_1, \ldots, s_p)^T$ collect all the working sparsity parameters, and $\mathcal{H}_{s,p}$ denote the collection of all such neural networks.

Given the observed data $\{\boldsymbol{y}(t_k)\}_{k=1}^n$ under models (1) and (2), our goal is to find a member in $\mathcal{H}_{s,p}$ to approximate the high-dimensional and possibly nonlinear functional $\boldsymbol{f}$ that encodes the regulatory relations among $\boldsymbol{x}(t)$, and also identify the underlying sparsity structure among all pairs of $\boldsymbol{x}$.

3. ALGORITHM

We develop an algorithm to estimate the unknown parameters in our sparse neural ODE model. We solve the following optimization problem:

$$\begin{aligned} &\text{minimize } \mathcal{L}(\boldsymbol{W}, \boldsymbol{\Theta}) = \ell(\boldsymbol{W}, \boldsymbol{\Theta}) + \lambda_1 \|\boldsymbol{W}\|_2^2 + \lambda_2 \|\boldsymbol{\Theta}\|_2^2, \\ &\text{subject to } \left\|\boldsymbol{w}_i\right\|_0 \le s_i, \quad i = 1, \ldots, p, \end{aligned} \tag{4}$$

where the loss function is of the form

$$\ell(\boldsymbol{f}) = \frac{1}{n} \sum_{k=1}^{n} \left\| \boldsymbol{y}(t_k) - \hat{\boldsymbol{x}}(t_k; \boldsymbol{f}) \right\|_2^2, \tag{5}$$

$\boldsymbol{f}$ is a function of the unknown parameters $\boldsymbol{W}$ and $\boldsymbol{\Theta}$, and $\hat{\boldsymbol{x}}(t)$ is an estimate of $\boldsymbol{x}(t)$. We adopt an ODE solver similar to that in Chen et al. (2018) to obtain $\hat{\boldsymbol{x}}(t)$. The ℓ_2 ridge regularization terms $\|\boldsymbol{W}\|_2^2$ and $\|\boldsymbol{\Theta}\|_2^2$ are introduced to prevent overfitting of the DNN. The ℓ_0 sparsity regularization is placed on $\boldsymbol{W}$ to achieve sparsity recovery and variable selection. Other choices of sparsity regularization include the ℓ_1 penalty (Allen, 2013), and the group lasso penalty (Yuan & Lin, 2006). Nevertheless, we choose the ℓ_0 penalty thanks to its nice theoretical properties (Zhang &

Zhang, 2012). There are two major challenges in solving (4). One is to compute the gradient of $\ell(\boldsymbol{f})$ with respect to $\boldsymbol{W}$ and $\boldsymbol{\Theta}$, respectively. The other is to incorporate the sparsity constraint. We next discuss how to address these two issues in some detail.

First, to compute the gradient of $\ell(\boldsymbol{f})$ with respect to $\boldsymbol{W}$ and $\boldsymbol{\Theta}$, we adopt the adjoint sensitivity method (Pontryagin, 2018). The key is to compute the adjoint, $\boldsymbol{a}(t) = \partial \ell(\boldsymbol{f})/\partial \boldsymbol{x}(t)$, whose dynamics are given by another ODE

$$\frac{d\boldsymbol{a}(t)}{dt} = -\boldsymbol{a}(t)^T \frac{\partial \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t))}{\partial \boldsymbol{x}} \equiv -\boldsymbol{a}(t)^T \boldsymbol{b}(t),$$

with the initial value $\boldsymbol{a}(t_1) = \boldsymbol{0}$. Note that $\boldsymbol{b}(t)$ can be efficiently evaluated by some numerical derivative evaluation during back-propagation (Baydin et al., 2018). The gradient with respect to $\boldsymbol{\Theta}$ is then

$$\frac{\partial \ell(\boldsymbol{f})}{\partial \boldsymbol{\Theta}} = -\int_{t_n}^{t_1} \boldsymbol{a}(t)^T \frac{\partial \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t))}{\partial \boldsymbol{\Theta}} dt.$$

The gradient with respect to $\boldsymbol{W}$ is computed similarly. Moreover, the computations of $\hat{\boldsymbol{x}}(t)$, $\boldsymbol{a}(t)$, and the two derivatives can all be done in a single call to the ODE solver. We write the derivative as

$$\left(\boldsymbol{v}_{\boldsymbol{w}_1}, \ldots, \boldsymbol{v}_{\boldsymbol{w}_p}, \boldsymbol{v}_{\boldsymbol{\Theta}}\right) = \nabla \ell(\boldsymbol{W}, \boldsymbol{\Theta}) = \left(\frac{\partial^T \ell(\boldsymbol{f})}{\partial \boldsymbol{W}}, \frac{\partial^T \ell(\boldsymbol{f})}{\partial \boldsymbol{\Theta}}\right)^T, \tag{6}$$

where $\boldsymbol{v}_{\boldsymbol{w}_i} \in \mathbb{R}^p$, $i = 1, \ldots, p$, and $\boldsymbol{v}_{\boldsymbol{\Theta}} \in \mathbb{R}^d$, where d is given after (3).

Next, to incorporate the sparsity constraint, we adopt and extend the algorithm of Bahmani, Raj & Boufounos (2013), which iteratively updates the sparse $\hat{\boldsymbol{W}}$ and the nonsparse $\hat{\boldsymbol{\Theta}}$. Specifically, we first initialize all the parameters in the selection layer by a uniform distribution, i.e., $w_{ij}^{(0)} \sim \text{Uniform}\left(-1/\sqrt{p+q}, 1/\sqrt{p+q}\right)$, and $\theta_{i'j}^{u(0)} \sim \text{Uniform}\left(-1/\sqrt{p+q}, 1/\sqrt{p+q}\right)$, for $i, j = 1, \ldots, p$, $i' = 1, \ldots, q$. We initialize all the parameters in the approximation layers by a normal distribution, i.e., $\boldsymbol{\Theta}^{(h)(0)} \sim \text{Normal}\left(\boldsymbol{0}, 0.1^2 \times \boldsymbol{I}_{n_h n_{h-1}}\right)$ for $h = 1, \ldots, H$, where $\boldsymbol{I}_{n_h n_{h-1}}$ is an identity matrix of dimension $n_h \times n_{h-1}$. We next compute the gradient of $\ell(\boldsymbol{f})$ with respect to $\boldsymbol{W}$ and $\boldsymbol{\Theta}$, and obtain the gradient vector at the current iteration r as $\left(\boldsymbol{v}_{\boldsymbol{w}_1}^{(r)}, \ldots, \boldsymbol{v}_{\boldsymbol{w}_p}^{(r)}, \boldsymbol{v}_{\boldsymbol{\Theta}}^{(r)}\right)$ following Equation (6). We then record the indices of the largest $2s_i$ entries of $\boldsymbol{v}_{\boldsymbol{w}_i}^{(r)}$ and the indices of the nonzero entries of the current estimate $\boldsymbol{w}_i^{(r)}$, and merge them to form the index set $\mathcal{T}_i^{(r)} \subseteq \{1, \ldots, p\}$, $i = 1, \ldots, p$. By construction, $\mathcal{T}_i^{(r)}$ has at most $3s_i$ indices. Finally, we minimize $\mathcal{L}(\boldsymbol{W}, \boldsymbol{\Theta})$ as in (4), but force all the entries of $\boldsymbol{w}_i$ whose corresponding indices are not in $\mathcal{T}_i^{(r)}$ to zero. We take the s_i largest absolute values of the minimizer of this constrained minimization as the updated estimate for $\boldsymbol{w}_i^{(r+1)}$, $\boldsymbol{W}^{(r+1)} = \left(\boldsymbol{w}_i^{(r+1)}, \ldots, \boldsymbol{w}_p^{(r+1)}\right)$, and $\boldsymbol{\Theta}^{(r+1)}$. We stop the iterations when the support $\mathcal{T}_i^{(r)}$ does not change from the previous step, or when it has reached the maximum number of iterations. We discuss the selection of the tuning parameters λ_1, λ_2, and $\boldsymbol{s} = (s_1, \ldots, s_p)$ later in Section 5.1.

We summarize the complete estimation procedure in Algorithm 1.

4. CONVERGENCE ANALYSIS

We next establish the theoretical guarantees of our proposed SNODE estimator. We note that the optimization problem in (4) is nonconvex and is NP-hard. We show that our estimator converges to the true parameters under some reasonable conditions on the objective function and that we can recover the true sparsity structure under some mild regularity conditions.

Algorithm 1. SNODE estimation procedure.

Input: $\boldsymbol{y}(t_1), \ldots, \boldsymbol{y}(t_n)$ and $\boldsymbol{s} = (s_1, \ldots, s_p)$

1: Initialization: $\boldsymbol{W}^{(0)}, \boldsymbol{\Theta}^{(0)}$.
2: **repeat**
3: Compute the gradient vector $\left(\boldsymbol{v}_{\boldsymbol{w}_1}^{(r)}, \ldots, \boldsymbol{v}_{\boldsymbol{w}_p}^{(r)}, \boldsymbol{v}_{\boldsymbol{\Theta}}^{(r)}\right) = \nabla \ell(\boldsymbol{W}, \boldsymbol{\Theta})$.
4: Form the support $\mathcal{T}_i^{(r)}$ by merging the indices of the largest $2s_i$ entries of $\boldsymbol{v}_{\boldsymbol{w}_i}^{(r)}$ and the indices of the nonzero entries of $\boldsymbol{w}_i^{(r)}$, $i = 1, \ldots, p$.
5: Minimize $\mathcal{L}(\boldsymbol{W}, \boldsymbol{\Theta})$ while constraining to zero the entries of $\boldsymbol{w}_i$ whose corresponding indices are not in $\mathcal{T}_i^{(r)}$.
6: Take the s_i largest absolute values of the minimizer to update the estimates $\boldsymbol{W}^{(r+1)}$ and $\boldsymbol{\Theta}^{(r+1)}$.
7: **until** the stopping criterion is met.

Output: $\hat{\boldsymbol{W}} = \boldsymbol{W}^{(r+1)}, \hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(r+1)}$

We first introduce the group generalized stable restricted Hessian (G2SRH) condition for the objective function, which generalizes a similar condition from Chen et al. (2020). We refer to a vector with s nonzero entries as an s-sparse vector.

Definition 1 (Group generalized stable restricted Hessian condition). Suppose the objective function $\mathcal{L}$ is twice continuously differentiable. Let $\boldsymbol{H}_{\mathcal{L}}(\cdot)$ be the Hessian of $\mathcal{L}$. Let $\tilde{\boldsymbol{w}}_i \in \mathbb{R}^p$ be any s_i-sparse vector, and $\tilde{\boldsymbol{\Theta}} \in \mathbb{R}^d$ be a vector of the same dimension as $\boldsymbol{\Theta}$, $i = 1, \ldots, p$. Let $\tilde{\boldsymbol{\Delta}} = \left(\tilde{\boldsymbol{w}}_1^T, \ldots, \tilde{\boldsymbol{w}}_p^T, \tilde{\boldsymbol{\Theta}}^T\right)^T$, and $\boldsymbol{s} = (s_1, \ldots, s_p)^T$. Furthermore, define

$$A_s(\boldsymbol{W}, \boldsymbol{\Theta}) = \sup\Big\{ \tilde{\boldsymbol{\Delta}}^{\mathrm{T}} \boldsymbol{H}_{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{\Theta}) \tilde{\boldsymbol{\Delta}} \mid \operatorname{supp}(\boldsymbol{w}_i) \cup \operatorname{supp}(\tilde{\boldsymbol{w}}_i) \le s_i, \quad \forall i = 1, \ldots, p, \|\tilde{\boldsymbol{\Delta}}\|_2 = 1 \Big\}, \tag{7}$$

$$B_s(\boldsymbol{W}, \boldsymbol{\Theta}) = \inf\Big\{ \tilde{\boldsymbol{\Delta}}^{\mathrm{T}} \boldsymbol{H}_{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{\Theta}) \tilde{\boldsymbol{\Delta}} \mid \operatorname{supp}(\boldsymbol{w}_i) \cup \operatorname{supp}(\tilde{\boldsymbol{w}}_i) \le s_i, \quad \forall i = 1, \ldots, p, \|\tilde{\boldsymbol{\Delta}}\|_2 = 1 \Big\}.$$

Then $\mathcal{L}$ is said to satisfy the G2SRH condition with constant μ_s if

$$1 \le \frac{A_s(\boldsymbol{W}, \boldsymbol{\Theta})}{B_s(\boldsymbol{W}, \boldsymbol{\Theta})} \le \mu_s.$$

We make some remarks on this condition. First, the G2SRH condition relaxes the convexity requirement for $\mathcal{L}$, but instead requires that the curvature of the objective function over some special sparse space be bounded locally. Second, it requires that the condition number of the submatrix of the Hessian $\boldsymbol{H}_{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{\Theta})$ whose rows correspond to the nonzero rows of $\boldsymbol{W}$ be not greater than μ_s. See also the discussion in Chen et al. (2020). Third, this G2SRH condition can be viewed as a generalization of the stable restricted Hessian condition of Bahmani, Raj & Boufounos (2013) used in feature selection, and the restricted isometry property of Candes, Romberg & Tao (2006) used in compressive sensing.

We next provide a simple example to further illustrate the G2SRH condition.

Example 1. Consider the objective function $\mathcal{L}(\boldsymbol{W}, \boldsymbol{\Theta}) = \left({\boldsymbol{w}_1}^T \boldsymbol{Q}_1 \boldsymbol{w}_1 + {\boldsymbol{w}_2}^T \boldsymbol{Q}_2 \boldsymbol{w}_2 + \lambda_1 \|\boldsymbol{w}_1\|_2^2 + \lambda_1 \|\boldsymbol{w}_2\|_2^2 + \lambda_2 \|\boldsymbol{\Theta}\|_2^2\right)/2$, where $\boldsymbol{Q}_1 = 2 \times \mathbf{1}\mathbf{1}^T - \boldsymbol{I}_p$, $\boldsymbol{Q}_2 = \boldsymbol{I}_p$. This objective function's Hessian is of the form

$$\boldsymbol{H}_{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{\Theta}) = \begin{bmatrix} \boldsymbol{Q}_1 + \lambda_1 \boldsymbol{I}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_2 + \lambda_1 \boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \lambda_2 \boldsymbol{I}_d \end{bmatrix}.$$

Note that all diagonal entries of $\boldsymbol{Q}_1$ are 1, and all off-diagonal entries are 2. Therefore, for any 1-sparse vector $\tilde{\boldsymbol{w}}_1$ and $\tilde{\boldsymbol{w}}_2$, we have $\tilde{\boldsymbol{w}}_1^T \boldsymbol{Q}_1 \tilde{\boldsymbol{w}}_1 = \|\tilde{\boldsymbol{w}}_1\|_2^2$, and $\tilde{\boldsymbol{w}}_2^T \boldsymbol{Q}_2 \tilde{\boldsymbol{w}}_2 = \|\tilde{\boldsymbol{w}}_2\|_2^2$. Then for any $\tilde{\boldsymbol{\Delta}} = \left(\tilde{\boldsymbol{w}}_1^T, \tilde{\boldsymbol{w}}_2^T, \tilde{\boldsymbol{\Theta}}^T\right)^T$ that satisfies $\|\tilde{\boldsymbol{\Delta}}\|_2^2 = 1$, we have $\tilde{\boldsymbol{\Delta}}^\top \boldsymbol{H}_{\mathcal{L}} \tilde{\boldsymbol{\Delta}} = (1 + \lambda_1)\left(\|\tilde{\boldsymbol{w}}_1\|_2^2 + \|\tilde{\boldsymbol{w}}_2\|_2^2\right) + \lambda_2 \|\tilde{\boldsymbol{\Theta}}\|_2^2$. If $1 + \lambda_1 > \lambda_2$, then $A_1(\boldsymbol{W}, \boldsymbol{\Theta}) = 1 + \lambda_1$, and $B_1(\boldsymbol{W}, \boldsymbol{\Theta}) = \lambda_2$. Therefore, $\mathcal{L}$ satisfies the G2SRH condition with $\mu_1 = (1 + \lambda_1)/\lambda_2$. On the other hand, we note that the Hessian of $\mathcal{L}$ is not positive semi-definite, and thus $\mathcal{L}$ is not convex, because $\check{\boldsymbol{w}}_1^T \left(\boldsymbol{Q}_1 + \lambda_1 \boldsymbol{I}_p\right) \check{\boldsymbol{w}}_1 = 2\lambda_1 - 2 < 0$ when $\check{\boldsymbol{w}}_1 = (1, -1, 0, \ldots, 0)^T \in \mathbb{R}^p$.

The next theorem shows that our SNODE estimator from Algorithm 1 converges to the population minimizer. Let $\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$ denote the population minimizer of (4), and let $\left(\boldsymbol{W}^{(r)}, \boldsymbol{\Theta}^{(r)}\right)$ denote the estimator from the rth iteration of Algorithm 1. Write $\boldsymbol{\Delta}^* = \left(\mathrm{vec}(\boldsymbol{W}^*)^T, (\boldsymbol{\Theta}^*)^T\right)^T \in \mathbb{R}^{p^2+d}$, and $\boldsymbol{\Delta}^{(r)} = \left(\mathrm{vec}\left(\boldsymbol{W}^{(r)}\right)^T, \left(\boldsymbol{\Theta}^{(r)}\right)^T\right)^T \in \mathbb{R}^{p^2+d}$. Then the derivative $\nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$ is a vector of dimension $p^2 + d$. Let $\mathcal{I}_i$ denote the set of positions of the $3s_i$ largest entries of $\nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$ in absolute value corresponding to each $\boldsymbol{w}_i^*$, $i = 1, \ldots, p$, and $\mathcal{S}_{\boldsymbol{\Theta}}$ denote the set of positions of $\boldsymbol{\Theta}^*$ in $\nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$. Let $\mathcal{S}_{\boldsymbol{\Delta}^*} = \left(\bigcup_{i=1}^p \mathcal{I}_i\right) \cup \mathcal{S}_{\boldsymbol{\Theta}^*}$.

Theorem 1. *Suppose that $\mathcal{L}$ satisfies the G2SRH condition with $\mu_{4s} \leq \left(1 + \sqrt{3}\right)/2$. Furthermore, suppose that $B_{4s}((\boldsymbol{W}, \boldsymbol{\Theta})) \geq \epsilon$ for some $\epsilon > 0$ and all $4s_i$-sparse $\boldsymbol{w}_i$. Then, the estimate $\left(\boldsymbol{W}^{(r)}, \boldsymbol{\Theta}^{(r)}\right)$ of the rth iteration of Algorithm 1 satisfies that*

$$\begin{aligned} \left\|\boldsymbol{\Delta}^{(r)} - \boldsymbol{\Delta}^*\right\|_2 \leq \; & \frac{1}{2^r} \left\|\boldsymbol{\Delta}^{(0)} - \boldsymbol{\Delta}^*\right\|_2 \\ & + \frac{6 + 2\sqrt{3}}{\epsilon} \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[\mathcal{S}_{\boldsymbol{\Delta}^*}]} \right\|_2, \end{aligned} \tag{8}$$

where $\nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[\mathcal{S}_{\boldsymbol{\Delta}^*}]}$ denotes the subvector of $\nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$ only including those entries with indices in $\mathcal{S}_{\boldsymbol{\Delta}^*}$.

We note that the bound in (8) has two terms. The first term converges to zero as the iteration number r goes to infinity. The second term, $\nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$, determines how accurate the estimator can be. From (4), $\nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$ can be decomposed into two parts:

$$\nabla \ell\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right) + 2\left(\lambda_1 \|\boldsymbol{W}^*\|_1 + \lambda_2 \|\boldsymbol{\Theta}^*\|_1\right). \tag{9}$$

When $\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$ is sufficiently close to an unconstrained minimum of ℓ, then the first term of (9) is negligible, since $\nabla \ell\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)$ is small at the minimum. Meanwhile, the magnitude of the second term in (9) can be controlled by adjusting the ratio of λ_1 and λ_2.

The next theorem shows that our estimator can recover the true sparsity structure. Let $\mathcal{S}_i \subseteq \{1, \ldots, p\}$ denote the set of indices corresponding to the nonzero entries of $\boldsymbol{w}_i^*$, and $\hat{\mathcal{S}}_i^{(r)}$ the

set of indices corresponding to the nonzero entries of $\boldsymbol{w}_i^{(r)}$ at the rth iteration for $i = 1, \ldots, p$ and $r = 0, 1, \ldots$. Let $m_i = \min\{|w_{ij}^*|, j \in \mathcal{S}_i\}$ be the minimum nonzero entry in absolute value in $\boldsymbol{w}_i^*$, and $M_i = \max\{\max(|\boldsymbol{w}_{ij}^{(0)}|, |\boldsymbol{w}_{ij}^*|), j \in \hat{\mathcal{S}}_i^{(0)} \cup \mathcal{S}_i\}$.

Theorem 2. *Suppose the same conditions as in Theorem 1 hold.*

(a) The set $\hat{\mathcal{S}}_i^{(r)}$, $i = 1, \ldots, p$, *satisfies that*

$$\left\| \left(\boldsymbol{w}_{1\left[\mathcal{S}_1 \setminus \hat{\mathcal{S}}_1^{(r)}\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[\mathcal{S}_p \setminus \hat{\mathcal{S}}_p^{(r)}\right]}^{*T} \right)^T \right\|_2 \leq \frac{1}{2^r} \left\| \boldsymbol{\Delta}^{(0)} - \boldsymbol{\Delta}^* \right\|_2 + \frac{6 + 2\sqrt{3}}{\epsilon} \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[\mathcal{S}_{\boldsymbol{\Delta}^*}]} \right\|_2,$$

where $\mathcal{S}_i \setminus \hat{\mathcal{S}}_i^{(r)}$ *denotes the set of all indices that are in* $\mathcal{S}_i$ *but are not in* $\hat{\mathcal{S}}_i^{(r)}$.
(b) Furthermore, if, for some $0 < \xi < 1$

$$\frac{6 + 2\sqrt{3}}{\epsilon \xi} \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[\mathcal{S}_{\boldsymbol{\Delta}^*}]} \right\|_2 \leq \sum_{i=1}^{p} m_i,$$

and for some $c > 0$, *max* $\left(\|\boldsymbol{\Theta}^*\|_\infty, \|\boldsymbol{\Theta}^{(0)}\|_\infty\right) \leq c/2$, *then we have*

$$\hat{\mathcal{S}}_i^{(r)} = \mathcal{S}_i \ \text{ when } \ r \geq \log_2 \frac{2\sum_{i=1}^{p} \sqrt{s_i} M_i + c\sqrt{d}}{(1 - \xi)\sum_{i=1}^{p} m_i}. \tag{10}$$

Following the remark after Theorem 1, it is reasonable to assume $\left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)|_{[\mathcal{S}_{\boldsymbol{\Delta}^*}]} \right\|_2$ is bounded. In addition, the conditions on $\|\boldsymbol{\Theta}^*\|_\infty$ and $\|\boldsymbol{\Theta}^{(0)}\|_\infty$ are mild, because we can apply weight normalization to all approximation layers. Then (10) shows that our algorithm can recover the exact support after a finite number of iterations.

5. SIMULATION STUDIES

5.1. Setup and Implementation Details

We consider two simulation examples: a linear ODE model, and a nonlinear ODE model. The observed data $\boldsymbol{y}(t)$ are drawn at equally spaced time points $\{t_1, \ldots, t_n\}$ in $[0, 1]$, with measurement errors following a normal distribution with mean 0 and standard deviation 0.1. We consider a single experimental input $u(t)$ that equals 0 for the first half of time interval and 1 for the second half.

For implementation, we employ the ODE solver of Chen et al. (2018). To obtain the initial values, we adopt a simple moving average technique to smooth the observed time series and remove short-term fluctuations, by

$$y_{MA,i}(t_k) = \frac{1}{4} y_i(t_k) + \frac{1}{2} y_i(t_{k+1}) + \frac{1}{4} y_i(t_{k+2}), \quad i = 1, \ldots, p, \quad k = 1, \ldots, n - 2.$$

In addition, considering that the zero-one signal of $u(t)$ is nondifferentiable at the change point $t = (t_1 + t_n)/2$, we consider a smooth approximation

$$\hat{u}(t, \gamma) = \frac{1}{1 + \exp\left[-8\left(t - \frac{t_1 + t_n}{2}\right)\right]},$$

and feed this approximated version of $u(t)$ into the DNNs. We use one approximation layer with 128 neurons for the linear ODE model, and two approximation layers, each with 128 neurons, for the nonlinear ODE model. We use the ReLU activation function in both cases. For simplicity, we set all the sparsity parameters equal such that $s_i = s$, $i = 1, \ldots, p$, and tune s using a BIC-type criterion following Chen & Chen (2008) and Gao & Song (2010):

$$\text{BIC} = n\log\,\hat{\sigma}^2 + p\,s \cdot \log n,$$

where $\hat{\sigma}^2$ is the mean squared error of the estimated time series $\hat{\boldsymbol{x}}(t)$. We choose s that minimizes BIC. We tune λ_1 and λ_2 following Chen et al. (2020) by first fixing their ratio and then choosing the ratio that minimizes the loss function of Equation (5). Furthermore, after we obtain the final estimate of the selection layer, we further update the approximation layers, which is similar in spirit to refitting the model after variable selection as in Chen et al. (2020).

We compare our method with three alternative ODE solutions: the linear ODE of Zhang et al. (2015), the additive ODE of Chen, Shojaie & Witten (2017), and the kernel ODE of Dai & Li (2021). All three alternative methods can be implemented within a unified framework with different choices of the kernel functions: a linear kernel for the linear ODE, an additive Matérn kernel for the additive ODE, and a general Matérn kernel for the kernel ODE. We evaluate the empirical performance using the estimation error, $\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2$, and the false selection rate (FSR)

$$\text{FSR} = \frac{\sum_{i=1}^{p} \left|\hat{\mathcal{S}}_i \backslash \mathcal{S}_{i,m}\right|}{\sum_{i=1}^{p} \left|\hat{\mathcal{S}}_i\right|},$$

where $\mathcal{S}_i$ and $\hat{\mathcal{S}}_i$ denote the support of the true $\boldsymbol{w}_i^*$ and of its estimate, respectively. We repeat the replications 50 times, and report the average results.

5.2. A Linear ODE Example

We first consider a linear ODE example, following a setup similar to that in Zhang et al. (2015). Specifically, we generate an ODE system with $p = 8$ as

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{A}\boldsymbol{x}(t) + u(t)\boldsymbol{B}\boldsymbol{x}(t) + u(t)\boldsymbol{I}_8,$$

where $\boldsymbol{x}(t) \in \mathbb{R}^{8\times 1}$, and $\boldsymbol{A} = \boldsymbol{B} \in \mathbb{R}^{8\times 8}$. Figure 2a shows the matrix $\boldsymbol{A}$, which takes a block diagonal structure with two blocks. Figure 2b shows the corresponding network structure among the eight nodes of $\boldsymbol{x}(t)$. We first set the sample size at $n = 1000$, and later consider different values of n for the SNODE method only.

Table 1 reports the average FSR and estimation error for various ODE methods for this linear ODE example, showing that our SNODE method achieves the smallest selection error as well as the smallest estimation error. On the other hand, our method has the largest standard error for the estimation error. This reflects the classical bias−variance trade-off: our method is the most flexible and achieves the smallest bias, but pays the price of having a larger variance in terms of the estimation. Figure 3 shows the true signal trajectories $\boldsymbol{x}(t)$ and the estimated trajectories for one data replication. It appears that our method estimates the signal trajectories well.

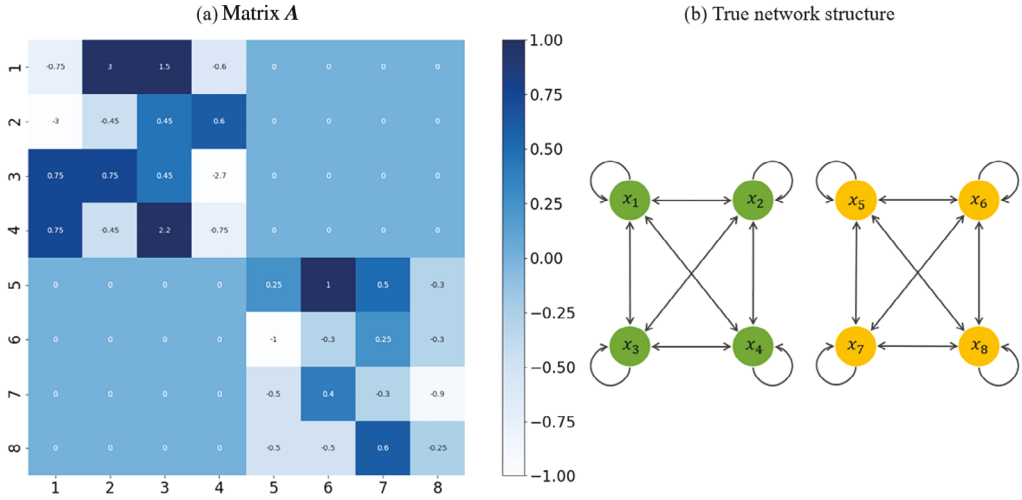

FIGURE 2: Linear ODE example: (a) the coefficient matrix $\boldsymbol{A} = \boldsymbol{B}$; and (b) the true network structure.

TABLE 1: Comparison of various ODE methods in terms of false selection rate (FSR) and estimation error (error; standard deviations in parentheses)

	Linear model		Nonlinear model	
Method	FSR	Error	FSR	Error
SNODE	0.205	2.530 (1.218)	0.251	3.286 (1.546)
Kernel ODE	0.250	3.157 (0.289)	0.392	3.908 (0.318)
Additive ODE	0.350	3.690 (0.344)	0.446	4.460 (0.374)
Linear ODE	0.242	3.287 (0.296)	0.539	4.984 (0.327)

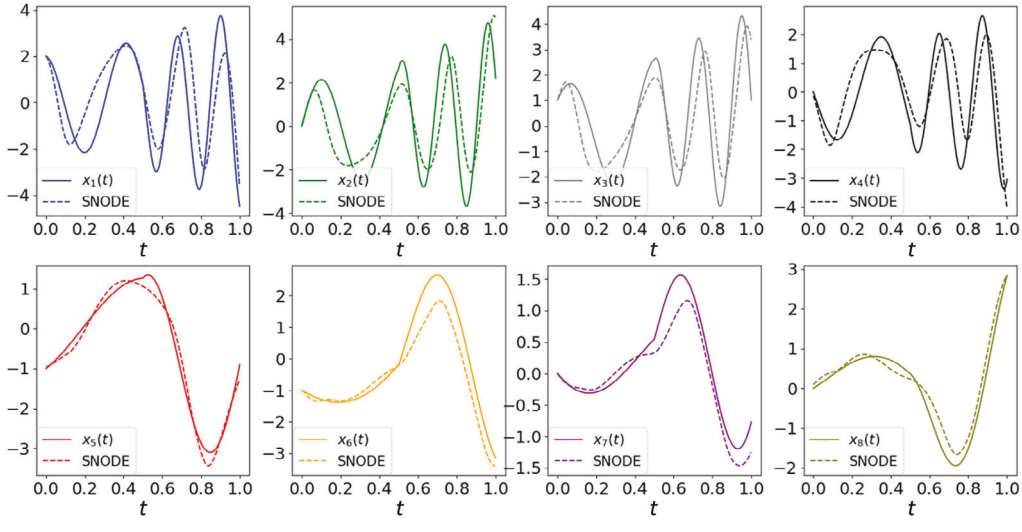

FIGURE 3: Linear ODE example: the true and estimated trajectories.

 DOI: 10.1002/cjs.11666

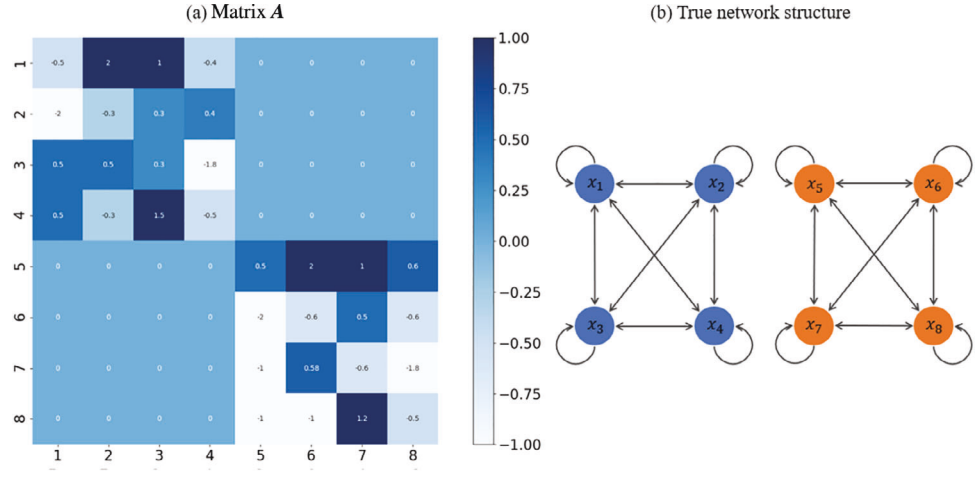

FIGURE 4: Nonlinear ODE example: (a) the coefficient matrix $\boldsymbol{A}$; and (b) the true network structure.

5.3. A Nonlinear ODE Example

We next consider a nonlinear ODE example. Specifically, we generate an ODE system with $p = 8$ as

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{x}(t) \odot [\boldsymbol{B}\boldsymbol{x}(t)] + 0.1u(t)\boldsymbol{x}(t) + u(t)\boldsymbol{I}_8,$$

where $\boldsymbol{x}(t) \in \mathbb{R}^{8\times 1}$, $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{8\times 8}$, and $\odot$ is the Hadamard product. Figure 4a shows the coefficient matrix $\boldsymbol{A}$, and Figure 4b shows the corresponding network structure according to $\boldsymbol{A}$. The matrix $\boldsymbol{B} = (B_{i,j})_{8\times 8}$ has $B_{5,6} = -0.5$, $B_{8,7} = -0.5$, and the rest equal to 0. This model thus includes two interaction terms, which correspond to interactions between $x_5(t)$ and $x_6(t)$, and between $x_7(t)$ and $x_8(t)$. We again fix the sample size at $n = 1000$.

Table 1 reports the average FSR and estimation error for various ODE methods for this nonlinear ODE example. Once again, our SNODE method achieves the smallest selection error and estimation error, but the largest estimation variation. Figure 5 shows the true signal trajectories $\boldsymbol{x}(t)$ and the average estimated trajectories for one data replication. Again, it is seen that our method works well for this nonlinear example.

5.4. Sample Size and Nonlinear Function Estimation

We have considered the case with $n = 1000$, which corresponds to the situation with dense signal observations. Next, we investigate the performance of our method under a smaller sample size by employing the same linear and nonlinear ODE models, but setting either $n = 50$ or $n = 100$. Recognizing that DNN methods usually require a relatively large training sample size, we employ linear interpolation to interpolate the observed data, and increase the training sample size to $n' = 1000$ for each signal trajectory. We also apply the moving average technique to smooth the trajectories. Table 2 reports the results based on the interpolated samples and compares them with the previous results with $n = 1000$. It can be seen that performance degrades a little with a smaller sample size, but remains reasonably close.

Next, we investigate the performance of our method in terms of estimating the dynamic function $\boldsymbol{f}$ in model (1). Table 3 reports the estimation error, $\|\hat{f}_i - f_i\|_2^2$, $i = 1, \ldots, p$. It can be seen that our method estimates $\boldsymbol{f}$ reasonably well.

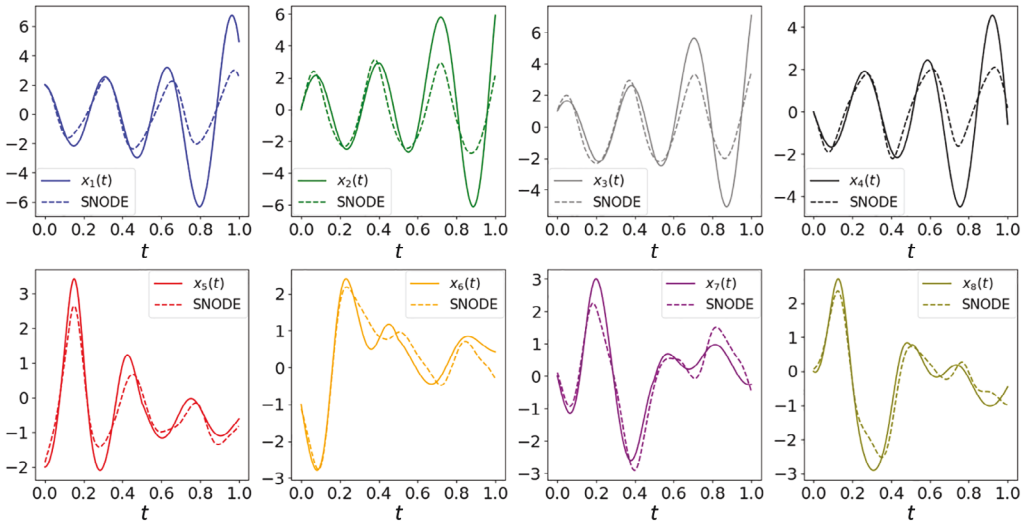

FIGURE 5: Nonlinear ODE example: the true and estimated trajectories.

TABLE 2: Comparison of false selection rate (FSR) and estimation error (error; standard deviations in parentheses) for different sample sizes, averaged over 50 data replications.

	Linear model		Nonlinear model	
Sample size	FSR	Error	FSR	Error
50	0.217	2.706 (1.061)	0.276	3.511 (1.541)
100	0.214	2.655 (1.122)	0.269	3.447 (1.535)
1000	0.205	2.530 (1.218)	0.251	3.286 (1.547)

TABLE 3: Evaluation of the estimation accuracy of the dynamic function $\boldsymbol{f}$ (standard deviations in parentheses).

	$\left\|\hat{f}_i - f_i\right\|_2^2$			
Model	$i = 1$	$i = 2$	$i = 3$	$i = 4$
Linear model	2.324 (0.298)	2.028 (0.348)	1.946 (0.292)	1.812 (0.345)
Nonlinear model	1.031 (0.436)	2.285 (2.252)	1.964 (1.394)	1.002 (1.667)
	$i = 5$	$i = 6$	$i = 7$	$i = 8$
Linear model	0.417 (0.035)	0.843 (0.104)	1.843 (0.135)	0.950 (0.151)
Nonlinear model	0.816 (0.245)	0.548 (0.097)	0.440 (0.088)	0.412 (0.086)

 DOI: 10.1002/cjs.11666

6. APPLICATION TO GENE NETWORK ANALYSIS

We illustrate our method with an in silico benchmark gene network data provided by GeneNetWeaver (GNW). The data has been generated for accessing the performance of reverse engineering of gene regulatory networks from yeast or *Escherichia coli* (Schaffter, Marbach & Floreano, 2011), and was used in the third DREAM challenges (Marbach et al., 2009). The regulatory networks are determined by a system of ODEs with external perturbations.

We investigate five networks from GNW with $p = 10$ nodes, which have been previously studied by Henderson & Michailidis (2014), Chen, Shojaie & Witten (2017), and Dai & Li (2021). For each network, GNW provides a set of noiseless gene expressions where the trajectories are normalized to the range $[0, 1]$ and are measured at 21 evenly spaced time points. Similar to the previous analyses, we add independent normal measurement error with mean 0 and standard deviation 0.25. We employ linear interpolation to interpolate the observed data, and increase the training sample size to $n' = 1000$ for each signal trajectory. We then apply the moving average technique to smooth the trajectory for the initial values:

$$y_{MA,i}(t_k) = \frac{1}{15}\sum_{m=0}^{14} y_i(t_{k+m}), \quad i = 1, \ldots, 10, \quad k = 1, \ldots, 986.$$

Figure 6 shows the noiseless trajectories, the trajectories with added error, and the smoothed trajectories of the 10 genes from one experiment on *E. coli*.

We apply our method with two approximation layers, each with 64 neurons. The rest of the model setup is the same as the one used in Section 5. For each network, we run 100 data replications. Table 4 reports the area under the receiver operating characteristic curve (AUC), our sparse neural ODE, the linear ODE of Zhang et al. (2015), the additive ODE of Chen, Shojaie & Witten (2017), and the kernel ODE of Dai & Li (2021). It can be seen that our method performs the best, achieving the largest AUC in all cases.

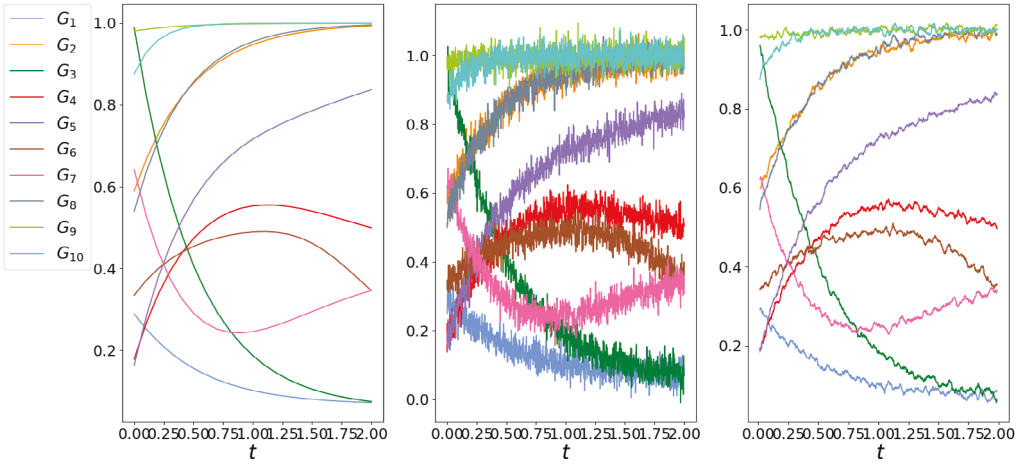

FIGURE 6: Left: The noiseless trajectories of 10 genes using linear interpolation. Middle: The trajectories of 10 genes after adding independent $\mathcal{N}(0, 0.025^2)$ noise. Right: The smoothed noisy trajectories.

TABLE 4: Comparison table of SNODE, kernel ODE, additive ODE, and linear ODE methods based on AUC (standard deviations in parentheses) over 100 simulations; boldface indicates largest AUC.

Dataset	SNODE	Kernel ODE	Additive ODE	Linear ODE
Ecoli1	**0.705** (0.003)	0.582 (0.003)	0.541 (0.003)	0.460 (0.004)
Ecoli2	**0.742** (0.003)	0.662 (0.002)	0.632 (0.003)	0.562 (0.003)
Yeast1	**0.723** (0.003)	0.603 (0.002)	0.541 (0.003)	0.436 (0.003)
Yeast2	**0.622** (0.002)	0.599 (0.002)	0.562 (0.004)	0.536 (0.003)
Yeast3	**0.682** (0.002)	0.612 (0.002)	0.569 (0.002)	0.487 (0.003)

7. DISCUSSION

We have proposed a new sparse neural ODE model to characterize flexible relations among multiple functional processes. The key is to model the dynamic changes of the latent states through a set of ODEs, and a DNN with a specially designed architecture. We have developed an estimation algorithm, established the theoretical guarantees, and demonstrated the efficacy of the proposed method.

The main advantages of our method include its ability to capture both linear and nonlinear relations among multivariate functions, thanks to the flexibility of the DNN model, and the interpretability of the final model thanks to the special regularization structure. Limitations include the intensive computations and the large sample size required for the DNN model fitting requires, but given the rapid advancement of computing power and availability of larger imaging datasets, we expect these problems to be alleviated.

ACKNOWLEDGEMENTS

Li's research was partially supported by National Institutes of Health (NIH) grant R01AG061303 and National Science Foundation (NSF) grant CIF 2102227. Wang's research was partially supported by National Science Foundation (NSF) grant 1613060. The authors thank the Editor, the Associate Editor, and two referees for their constructive comments.

REFERENCES

Allen, G. I. (2013). Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics*, 22, 284–299.

Arora, R., Basu, A., Mianjy, P., & Mukherjee, A. (2018). Understanding deep neural networks with rectified linear units. *International Conference on Learning Representations.* Vancouver Convention Center.

Bahmani, S., Raj, B., & Boufounos, P. T. (2013). Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14, 807–841.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018). Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18, 13–14.

Bertsekas, D. P. (2014). *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York.

Biegler, L., Damiano, J., & Blau, G. (1986). Nonlinear parameter estimation: A case study comparison. *AIChE Journal*, 32, 29–45.

Bullmore, E. & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10, 186–198.

Cai, X., Bazerque, J., & Giannakis, G. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology*, 9, e1003068.

Candes, E. J., Romberg, J. K., & Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59, 1207–1223.

Chen, J. & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 759–771.

Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018),* Montréal, Canada, 6571–6583.

Chen, S., Shojaie, A., & Witten, D. M. (2017). Network reconstruction from high dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112, 1697–1707.

Chen, Y., Gao, Q., Liang, F., & Wang, X. (2020). Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2020.1814305

Chou, I.-C. & Voit, E. O. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences*, 219, 57–83.

Comminges, L. & Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40, 2667–2696.

Dai, X. & Li, L. (2021). Kernel ordinary differential equations. *Journal of the American Statistical Association*, 1–35. https://doi.org/10.1080/01621459.2021.1882466

Dattner, I. & Klaassen, C. A. J. (2015). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9, 1939–1973.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Feng, J. & Simon, N. (2017). *Sparse-input neural networks for high-dimensional nonparametric regression and classification*, https://arxiv.org/abs/1711.07592.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19, 1273–1302.

Gao, X. & Song, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105, 1531–1540.

Gardner, T. S., Di Bernardo, D., Lorenz, D., & Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301, 102–105.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep Learning*, Vol. 1. MIT Press, Cambridge.

Hairer, E., Nørsett, S., & Wanner, G. (2000). *Solving Ordinary Differential Equations. I. Nonstiff Problems*, 2nd ed., Springer, Berlin.

Hasan, M. K. & Pal, C. (2019). A new smooth approximation to the zero one loss with a probabilistic interpretation. *ACM Transactions on Knowledge Discovery from Data*, 14, 1:1–1:28.

Henderson, J. & Michailidis, G. (2014). Network reconstruction using nonparametric additive ODE models. *PLoS ONE*, 9, e94003.

Izhikevich, E. M. (2006). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press, Cambridge.

Kutta, W. (1901). Beitrag zur näherungsweisen integration totaler differentialgleichungen. *Zeitschrift für angewandte Mathematik und Physik*, 46, 435–453.

Li, Y., Chen, C.-Y., & Wasserman, W. W. (2015). Deep feature selection: Theory and application to identify enhancers and promoters. *International Conference on Research in Computational Molecular Biology,* Springer, Berlin, 205–217.

Liang, F., Li, Q., & Zhou, L. (2018). Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113, 955–972.

Lu, T., Liang, H., Li, H., & Wu, H. (2011). High-dimensional odes coupled with mixed effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, 106, 1242–1258.

Marbach, D., Schaffter, T., Mattiussi, C., & Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16, 229–239.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., et al. (2017). *Automatic differentiation in pytorch*. https://openreview.net/pdf?id=BJJsrmfCZ

Pontryagin, L. S. (2018). *Mathematical Theory of Optimal Processes*, 1st ed., Routledge, London.

Ramsay, J., Hooker, G., Campbell, D., & Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of Royal Statistical Society B*, 69, 741–796.

Schaffter, T., Marbach, D., & Floreano, D. (2011). Genenetweaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27, 2263–2270.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Wu, H. (2005). Statistical methods for HIV dynamic studies in AIDS clinical trials. *Statistical Methods in Medical Research*, 14, 171–192.

Wu, H., Lu, T., Xue, H., & Liang, H. (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109, 700–716.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zhang, C.-H. & Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27, 576–593.

Zhang, T., Wu, J., Li, F., Caffo, B., & Boatman-Reich, D. (2015). A dynamic directional model for effective brain connectivity using electrocorticographic (ECOG) time series. *Journal of the American Statistical Association*, 110, 93–106.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

APPENDIX A

Let $\boldsymbol{v}_{[\mathcal{I}]}$ denote restriction of the vector $\boldsymbol{v}$ to the indices in set $\mathcal{I}$, or a vector that equals $\boldsymbol{v}$ except for coordinates in $\mathcal{I}^c$ where it is zero, depending on the context. Let $\boldsymbol{P}_{\mathcal{I}}$ denote the restriction of the identity matrix to the columns indicated by $\mathcal{I}$. For convenience, we also denote $\alpha_k(\boldsymbol{p},\boldsymbol{q}) = \int_0^1 A_k(\boldsymbol{p}+(1-t)\boldsymbol{q})dt$, $\beta_k(\boldsymbol{p},\boldsymbol{q}) = \int_0^1 B_k(\boldsymbol{p}+(1-t)\boldsymbol{q})dt$, and $\gamma_k(\boldsymbol{p},\boldsymbol{q}) = \alpha_k(\boldsymbol{p},\boldsymbol{q}) - \beta_k(\boldsymbol{p},\boldsymbol{q})$, where $A_k(.)$ and $B_k(.)$ are defined by (7) respectively, and $\boldsymbol{p},\boldsymbol{q}$ are any vectors of the same dimension.

In Theorems 1 and 2, we establish the convergence of our estimates and the selection consistency. To prove Theorems 1 and 2, we first establish a series of results on how the algorithm operates on its current estimate, leading to the convergence of the estimates over iterations. Specifically, in order to search for population minimizer $\boldsymbol{\Delta}^* = \left(\left(\boldsymbol{w}_1^*\right)^T, \ldots, \left(\boldsymbol{w}_p^*\right)^T, (\boldsymbol{\Theta}^*)^T\right)^T$, we obtain the pruned estimates $\boldsymbol{\Delta}^{(r-1)} = \left(\left(\boldsymbol{w}_1^{(r-1)}\right)^T, \ldots, \left(\boldsymbol{w}_p^{(r-1)}\right)^T, \left(\boldsymbol{\Theta}^{(r-1)}\right)^T\right)^T$ at the $(r-1)$th iteration of the SNODE algorithm and the intermediate estimates $\boldsymbol{\Upsilon}^{(r)} = \left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r)}\right)^T, \ldots, \left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r)}\right)^T, \left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r)}\right)^T\right)^T$, which are explicitly defined in Lemma A2 at the rth iteration, as well as the pruned estimates $\boldsymbol{\Delta}^{(r)} = \left(\left(\boldsymbol{w}_1^{(r)}\right)^T, \ldots, \left(\boldsymbol{w}_p^{(r)}\right)^T, \left(\boldsymbol{\Theta}^{(r)}\right)^T\right)^T$ at the rth iteration.

To get the results of Theorems 1 and 2, we follow the flow of proof below:

1. We construct the upper bound for $\left\| \left(\left(\boldsymbol{w}_1^{(r-1)} - \boldsymbol{w}^*\right)^T_{\left[(\mathcal{V}_1^{(r-1)})^c\right]}, \ldots, \left(\boldsymbol{w}_p^{(r-1)} - \boldsymbol{w}^*\right)^T_{\left[(\mathcal{V}_p^{(r-1)})^c\right]} \right)^T \right\|_2$ (see proof of Lemma A1), which helps establishing the upper bound for an iteration-invariant $\left\| \left(\boldsymbol{w}^{*T}_{\left[\mathcal{T}_1^c\right]}, \ldots, \boldsymbol{w}^{*T}_{\left[\mathcal{T}_p^c\right]} \right)^T \right\|_2$. Here $\mathcal{V}_i^{(r-1)}$ is the $2s_i$ coordinates of $\boldsymbol{v}_{\boldsymbol{w}_i}^{(r-1)}$ with the largest magnitude for $i = 1, \ldots, p$.
2. Based on some properties of *G2SRH* (see Definition 1, Propositions A1 and A2), we construct the upper bound for the distance between the intermediate estimates in

the $(r-1)$th iteration and iteration-invariant $\left\| \left(\boldsymbol{w}^{*T}_{1\left[S_1 \setminus \hat{S}_1^{(r)}\right]}, \ldots, \boldsymbol{w}^{*T}_{p\left[S_p \setminus \hat{S}_p^{(r)}\right]}, \left(\boldsymbol{\Theta}^*\right)^T \right)^T - \left(\left(\boldsymbol{\Upsilon}^{(r)}_{\boldsymbol{w}_1}\right)^T, \ldots, \left(\boldsymbol{\Upsilon}^{(r)}_{\boldsymbol{w}_p}\right)^T, \left(\boldsymbol{\Upsilon}^{(r)}_{\boldsymbol{\Theta}}\right)^T \right)^T \right\|_2$ (see proof of Lemma A2).

3. Based on the results from step 1 and step 2, we obtain an inequality related to the estimation error in the $(r-1)$th iteration $\left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2$ and the estimation error in the rth iteration $\left\|\boldsymbol{\Delta}^{(r)} - \boldsymbol{\Delta}^*\right\|_2$ (see proof of Lemma A3).
4. Based on the result from step 3, we can obtain the inequalities for $\left\|\boldsymbol{\Delta}^{(r)} - \boldsymbol{\Delta}^*\right\|_2$ and $\left\| \left(\boldsymbol{w}^{*T}_{1\left[S_1 \setminus \hat{S}_1^{(r)}\right]}, \ldots, \boldsymbol{w}^{*T}_{p\left[S_p \setminus \hat{S}_p^{(r)}\right]} \right)^T \right\|_2$ in a recursive way (see proof of Theorems 1 and 2).

Now we can proceed to prove the main results of Theorems 1 and 2.

Proof of Theorem 1. Let $\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}$ denote the set $\operatorname{supp}\left(\boldsymbol{w}_i^{(r-1)} - \boldsymbol{w}^*\right)$, i.e., the support set which includes the indices of the nonzero entries of $\left(\boldsymbol{w}_i^{(r-1)} - \boldsymbol{w}^*\right)$, $i = 1, \ldots, p$. For consistency with the notation of $\boldsymbol{s}$, let $\boldsymbol{s}'$ denote $\left(s_1', \ldots, s_p'\right)^T$.

Following Definition 1, it is easy to verify that for $s_i \le s_i'$ and any vector $\boldsymbol{v}$, we have $A_{\boldsymbol{s}}(\boldsymbol{v}) \le A_{\boldsymbol{s}'}(\boldsymbol{v})$ and $B_{\boldsymbol{s}}(\boldsymbol{v}) \ge B_{\boldsymbol{s}'}(\boldsymbol{v})$. Henceforth, for $s_i \le s_i'$ and any pair of vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, we have $\alpha_{\boldsymbol{s}}(\boldsymbol{p},\boldsymbol{q}) \le \alpha_{\boldsymbol{s}'}(\boldsymbol{p},\boldsymbol{q})$, $\beta_{\boldsymbol{s}}(\boldsymbol{p},\boldsymbol{q}) \ge \beta_{\boldsymbol{s}'}(\boldsymbol{p},\boldsymbol{q})$, and $\mu_{\boldsymbol{s}} \le \mu_{\boldsymbol{s}'}$. Consequently, for any function that satisfies $\mu_{\boldsymbol{s}}$-GSRH,

$$\frac{\alpha_{\boldsymbol{s}}(\boldsymbol{p},\boldsymbol{q})}{\beta_{\boldsymbol{s}}(\boldsymbol{p},\boldsymbol{q})} = \frac{\int_0^1 A_{\boldsymbol{s}}(t\boldsymbol{p} + (1+t)\boldsymbol{q})dt}{\int_0^1 B_{\boldsymbol{s}}(t\boldsymbol{p} + (1+t)\boldsymbol{q})dt} \le \frac{\int_0^1 \mu_{\boldsymbol{s}} B_{\boldsymbol{s}}(t\boldsymbol{p} + (1+t)\boldsymbol{q})dt}{\int_0^1 B_{\boldsymbol{s}}(t\boldsymbol{p} + (1+t)\boldsymbol{q})dt} = \mu_{\boldsymbol{s}},$$

holds and therefore $\frac{\gamma_{\boldsymbol{s}}(\boldsymbol{p},\boldsymbol{q})}{\beta_{\boldsymbol{s}}(\boldsymbol{p},\boldsymbol{q})} \le \mu_{\boldsymbol{s}} - 1$. Thus, applying Lemma A3 to the estimation in the rth iteration, we obtain

$$\begin{aligned}
&\left\|\boldsymbol{\Delta}^{(r)} - \boldsymbol{\Delta}^*\right\|_2 \\
&\quad \le \left(\mu_{4\boldsymbol{s}} - 1\right)\mu_{4\boldsymbol{s}}\left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 + \frac{2\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}\right]}\right\|_2}{\beta_{4\boldsymbol{s}}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \\
&\qquad + \mu_{4\boldsymbol{s}}\frac{\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^p \left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]}\right\|_2 + \left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]}\right\|_2}{\beta_{2\boldsymbol{s}}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \\
&\quad \le \left(\mu_{4\boldsymbol{s}} - 1\right)\mu_{4\boldsymbol{s}}\left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 + \frac{2\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\mathcal{S}_{\boldsymbol{\Delta}^*}\right]}\right\|_2}{\beta_{4\boldsymbol{s}}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \\
&\qquad + 2\mu_{4\boldsymbol{s}}\frac{\left\|\nabla\mathcal{L}(\boldsymbol{W}^*, \boldsymbol{\Theta}^*_{\left[\mathcal{S}_{\boldsymbol{\Delta}^*}\right]}\right\|_2}{\beta_{2\boldsymbol{s}}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}.
\end{aligned}$$

Applying the assumption $\mu_{4s} \leq \frac{1+\sqrt{3}}{2}$ and $\epsilon \leq B_{4s}(\boldsymbol{W}, \boldsymbol{\Theta})$ for some $\epsilon > 0$ for all $4s_i$-sparse $\boldsymbol{w}_i$, $i = 1, \ldots, p$, we have

$$\|\Delta^{(r)} - \boldsymbol{\Delta}^*\|_2 \leq \frac{1}{2}\left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 + \frac{3+\sqrt{3}}{\epsilon}\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[S_{\boldsymbol{\Delta}^*}]}\right\|_2. \tag{A1}$$

Theorem 1 follows using (A1) recursively. ■

Proof of Theorem 2. Since $\left\|\left(\boldsymbol{w}_{1\left[S_1\setminus\hat{S}_1^{(r)}\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[S_p\setminus\hat{S}_p^{(r)}\right]}^{*T}\right)^T\right\|_2 = \left\|\left(\left(\boldsymbol{w}_1^{(r)} - \boldsymbol{w}_1^*\right)_{\left[S_1\setminus\hat{S}_1^{(r)}\right]}^T, \ldots, \left(\boldsymbol{w}_p^{(r)} - \boldsymbol{w}_p^*\right)_{\left[S_p\setminus\hat{S}_p^{(r)}\right]}^T\right)^T\right\|_2 \leq \left\|\boldsymbol{\Delta}^{(r)} - \boldsymbol{\Delta}^*\right\|_2$, based on Theorem 1, we have

$$\left\|\left(\boldsymbol{w}_{1\left[S_1\setminus\hat{S}_1^{(r)}\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[S_p\setminus\hat{S}_p^{(r)}\right]}^{*T}\right)^T\right\|_2 \leq \frac{1}{2}\left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 + \frac{3+\sqrt{3}}{\epsilon}\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[S_{\boldsymbol{\Delta}^*}]}\right\|_2. \tag{A2}$$

Part (1) of Theorem 2 follows using (A2) recursively.

Furthermore

$$\begin{aligned}
&\left\|\left(\boldsymbol{w}_{1\left[S_1\setminus\hat{S}_1^{(r)}\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[S_p\setminus\hat{S}_p^{(r)}\right]}^{*T}\right)^T\right\|_2 \\
&\leq 2^{-r}\left\|\boldsymbol{\Delta}^{(0)} - \boldsymbol{\Delta}^*\right\|_2 + \frac{6+2\sqrt{3}}{\epsilon}\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[S_{\boldsymbol{\Delta}^*}]}\right\|_2 \\
&\leq 2^{-r}\left(\sum_{i=1}^{p}\left\|\boldsymbol{w}_i^{(0)} - \boldsymbol{w}_i^*\right\|_2 + \left\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\right\|_2\right) + \frac{6+2\sqrt{3}}{\epsilon}\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[S_{\boldsymbol{\Delta}^*}]}\right\|_2 \\
&\leq 2^{-r}\left(2\sum_{i=1}^{p}\sqrt{s_i}M_i + c\sqrt{D}\right) + \xi\sum_{i=1}^{p}m_i \\
&\leq \sum_{i=1}^{p}m_i \text{ if } r \geq \frac{(2\sum_{i=1}^{p}\sqrt{s_i}M_i + c\sqrt{d}}{(1-\xi)\sum_{i=1}^{p}m_i}.
\end{aligned}$$

The second inequality follows from the triangle inequality. With $\max\left(\left\|\boldsymbol{\Theta}^*\right\|_\infty, \left\|\boldsymbol{\Theta}^{(0)}\right\|_\infty\right) \leq c/2$, $\left\|\boldsymbol{\Theta}^{(0)} - \boldsymbol{\Theta}^*\right\|_2 \leq c\sqrt{d}$ holds, where c is a constant. Combining this fact and the assumption $\sum_{i=1}^{p}m_i \geq \frac{6+2\sqrt{3}}{\epsilon\xi}\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{[S_{\boldsymbol{\Delta}^*}]}\right\|_2$ with $0 < \xi < 1$, the third inequality can be obtained. The last inequality follows after some algebra. This implies $\hat{S}_i^{(r)} = S_i$, if $r \geq \log_2\frac{2\sum_{i=1}^{p}\sqrt{s_i}M_i + c\sqrt{d}}{(1-\xi)\sum_{i=1}^{p}m_i}$. ■

Proposition A1. *Let $\boldsymbol{Q}(t)$ be a matrix-valued function such that for all $t \in [0,1]$, $\boldsymbol{Q}(t)$ is symmetric and its eigenvalues lie in interval $[B(t), A(t)]$ with $B(t) > 0$. Then for any vector $\boldsymbol{v}$, we have*

$$\left(\int_0^1 B(t)dt\right)\|\boldsymbol{v}\|_2 \le \left\|\left(\int_0^1 \boldsymbol{Q}(t)dt\right)\boldsymbol{v}\right\|_2 \le \left(\int_0^1 A(t)dt\right)\|\boldsymbol{v}\|_2.$$

Proposition A2. *Let $\boldsymbol{Q}(t)$ be a matrix-valued function such that for all $t \in [0,1]$, $\boldsymbol{Q}(t)$ is symmetric and its eigenvalues lie in the interval $[B(t), A(t)]$ with $B(t) > 0$. If Γ is a subset of row/column indices of $\boldsymbol{Q}(t)$, then for any vector $\boldsymbol{v}$, we have*

$$\left\|\left(\int_0^1 \boldsymbol{P}_\Gamma^T \boldsymbol{Q}(t)\boldsymbol{P}_{\Gamma^c} dt\right)\boldsymbol{v}\right\|_2 \le \left(\int_0^1 \frac{A(t)-B(t)}{2}dt\right)\|\boldsymbol{v}\|_2.$$

Proof of Propositions A1 and A2. The detailed proof of Propositions A1 and A2 can be found in Bahmani, Raj & Boufounos (2013), and thus is omitted here. ■

Lemma A1. *The estimate at the $(r-1)$th iteration $\left(\left(\boldsymbol{w}_1^{(r-1)}\right)^T, \ldots, \left(\boldsymbol{w}_p^{(r-1)}\right)^T, \left(\boldsymbol{\Theta}^{(r-1)}\right)^T\right)^T$ satisfies*

$$\begin{aligned}
&\left\|\left(\left(\boldsymbol{w}_1^{(r-1)} - \boldsymbol{w}^*\right)^T_{\left[\left(\mathcal{V}_1^{(r-1)}\right)^c\right]}, \ldots, \left(\boldsymbol{w}_p^{(r-1)} - \boldsymbol{w}^*\right)^T_{\left[\left(\mathcal{V}_p^{(r-1)}\right)^c\right]}\right)^T\right\|_2 \\
&\quad \le \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 \times \frac{\gamma_{4s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right) + \gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \\
&\qquad + \frac{\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^p \left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]}\right\|_2 + \left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]}\right\|_2}{\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}.
\end{aligned}$$

Proof of Lemma A1. For simplicity, we rewrite $\boldsymbol{v}^{(r-1)} = \left(\left(\boldsymbol{v}_{\boldsymbol{w}_1}^{(r-1)}\right)^T, \ldots, \left(\boldsymbol{v}_{\boldsymbol{w}_p}^{(r-1)}\right)^T, \left(\boldsymbol{v}_{\boldsymbol{\Theta}}^{(r-1)}\right)^T\right)^T$ in the following proof.

Since $\mathcal{V}_i^{(r-1)}$ is the $2s_i$ coordinates of $\boldsymbol{v}_{\boldsymbol{w}_i}^{(r-1)}$ with the largest magnitude and $|\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}| \le 2s_i$, we have $\left\|\boldsymbol{v}^{(r-1)}_{\left[\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right]}\right\|_2 \le \left\|\boldsymbol{v}^{(r-1)}_{\left[\mathcal{V}_i^{(r-1)}\right]}\right\|_2$ for $i = 1, \ldots, p$. Therefore

$$\left\|\boldsymbol{v}^{(r-1)}_{\left[\bigcup_{i=1}^p \left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]}\right\|_2 \le \left\|\boldsymbol{v}^{(r-1)}_{\left[\bigcup_{i=1}^p \left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]}\right\|_2. \tag{A3}$$

According to the algorithm, $\boldsymbol{v}^{(r-1)} = \nabla\mathcal{L}(\boldsymbol{W}^{(r-1)}, \boldsymbol{\Theta}^{(r-1)})$, and thus

$$\begin{aligned}
&\left\|\boldsymbol{v}^{(r-1)}_{\left[\bigcup_{i=1}^p \left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]}\right\|_2 \\
&\quad \ge \left\|\nabla\mathcal{L}\left(\boldsymbol{W}^{(r-1)}, \boldsymbol{\Theta}^{(r-1)}\right)_{\left[\bigcup_{i=1}^p \left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} - \nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^p \left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]}\right\|_2
\end{aligned}$$

$$
\begin{aligned}
&- \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 \\
&= \left\| \left(\int_0^1 \boldsymbol{P}^T_{\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{(r-1)} + (1-t)\boldsymbol{\Delta}^*\right) dt \right) \left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right) \right\|_2 \\
&\quad - \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 \\
&\geq \mathrm{I} + \mathrm{II} - \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2,
\end{aligned}
$$

where we have

$$
\begin{aligned}
\mathrm{I} = &\left\| \left(\int_0^1 \boldsymbol{P}^T_{\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{(r-1)} + (1-t)\boldsymbol{\Delta}^*\right) \boldsymbol{P}_{\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \cap \mathcal{V}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}} dt \right) \right. \\
&\left. \times \left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2, \\
\mathrm{II} = &\left\| \left(\int_0^1 \boldsymbol{P}^T_{\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{(r-1)} + (1-t)\boldsymbol{\Delta}^*\right) \boldsymbol{P}_{\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \cap \mathcal{V}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}} dt \right) \right. \\
&\left. \times \left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \cap \mathcal{V}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}\right]} \right\|_2,
\end{aligned}
$$

and the last inequality is derived from the triangle inequality by splitting $\left(\bigcup_{i=1}^{p} \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}$ into two sets $\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)$ and $\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \cap \mathcal{V}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}$.

Applying Propositions A1 and A2, we have

$$
\begin{aligned}
&\left\| \boldsymbol{v}^{(r-1)}|_{\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)} \right\|_2 \\
&\geq \beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right) \left\| \left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 \\
&\quad - \frac{\gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2} \left\| \left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \cap \mathcal{V}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}\right]} \right\|_2
\end{aligned}
$$

$$
\begin{aligned}
&- \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 \\
&\geq \beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right) \left\| \left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 \\
&\quad - \frac{\gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2} \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 - \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 . \qquad \text{(A4)}
\end{aligned}
$$

Similarly, by Propositions A1 and A2, we get

$$
\begin{aligned}
&\left\| \boldsymbol{v}^{(r-1)}_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2 \\
&\leq \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^{(r-1)}, \boldsymbol{\Theta}^{(r-1)}\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} - \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2 \\
&\quad + \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2 \\
&= \mathrm{III} + \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2 \\
&\leq \frac{\gamma_{4s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2} \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 + \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2 , \qquad \text{(A5)}
\end{aligned}
$$

where we have

$$
\begin{aligned}
\mathrm{III} = \Bigg\| \Bigg(\int_0^1 \boldsymbol{P}^T_{\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{(r-1)} + (1-t)\boldsymbol{\Delta}^*\right) \boldsymbol{P}_{\left(\bigcup_{i=1}^{p} \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right) \cup S_{\boldsymbol{\Theta}}} dt \Bigg) \\
\times \left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right)_{\left[\left(\bigcup_{i=1}^{p} \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right) \cup S_{\boldsymbol{\Theta}}\right]} \Bigg\| .
\end{aligned}
$$

Since $\left\|\left(\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right)_{\left[\cup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]}\right\|_2 = \left\| \left(\left(\boldsymbol{w}_1^{(r-1)} - \boldsymbol{w}_1^*\right)^T_{\left[\left(\mathcal{V}_1^{(r-1)}\right)^c\right]}, \ldots, \left(\boldsymbol{w}_p^{(r-1)} - \boldsymbol{w}_p^*\right)^T_{\left[\left(\mathcal{V}_p^{(r-1)}\right)^c\right]} \right)^T \right\|$, by combining (A3)–(A5) we obtain

$$
\frac{\gamma_{4s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2} \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 + \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2
$$

$$
\begin{aligned}
&\geq \left\| \boldsymbol{v}^{(r-1)}_{\left[\bigcup_{i=1}^{p}\left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2 \\
&\geq \left\| \boldsymbol{v}^{(r-1)}_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 \\
&\geq \beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right) \left\| \left(\left(\boldsymbol{w}_1^{(r-1)} - \boldsymbol{w}_1^*\right)^T_{\left[\left(\mathcal{V}_1^{(r-1)}\right)^c\right]}, \ldots, \left(\boldsymbol{w}_p^{(r-1)} - \boldsymbol{w}_p^*\right)^T_{\left[\left(\mathcal{V}_p^{(r-1)}\right)^c\right]} \right) \right\|_2 \\
&\quad - \frac{\gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2} \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 - \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 .
\end{aligned}
$$

Therefore

$$
\begin{aligned}
&\left\| \left(\left(\boldsymbol{w}_1^{(r-1)} - \boldsymbol{w}^*\right)^T_{\left[\left(\mathcal{V}_1^{(r-1)}\right)^c\right]}, \ldots, \left(\boldsymbol{w}_p^{(r-1)} - \boldsymbol{w}^*\right)^T_{\left[\left(\mathcal{V}_p^{(r-1)}\right)^c\right]} \right)^T \right\|_2 \\
&\leq \frac{\gamma_{4s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right) + \gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 \\
&\quad + \frac{\left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 + \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^{p} \mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2}{\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}.
\end{aligned}
$$

■

Lemma A2. *The vector given by*

$$
\begin{aligned}
&\left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r-1)}\right)^T, \ldots, \left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r-1)}\right)^T, \left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)^T \right)^T = \arg\min_{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p, \boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p, \boldsymbol{\Theta}) \\
&s.t.\ \ \boldsymbol{w}_{i\left[\left(\mathcal{T}_i^{(r-1)}\right)^c\right]} = 0,
\end{aligned} \tag{A6}
$$

for $i = 1, \ldots, p$, *satisfies*

$$
\begin{aligned}
&\left\| \left(\boldsymbol{w}^{*T}_{1\left[\mathcal{T}_1^{(r-1)}\right]}, \ldots, \boldsymbol{w}^{*T}_{p\left[\mathcal{T}_p^{(r-1)}\right]}, (\boldsymbol{\Theta}^*)^T \right) - \left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r-1)}\right)^T, \ldots, \left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r-1)}\right)^T, \left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)^T \right)^T \right\|_2 \\
&\leq \frac{\left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^{p} \mathcal{T}_i^{(r-1)}\right) \cup S_{\boldsymbol{\Theta}}\right]} \right\|_2}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \\
&\quad + \frac{\gamma_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \left\| \left(\boldsymbol{w}^{*T}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}, \ldots, \boldsymbol{w}^{*T}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]} \right)^T \right\|_2 .
\end{aligned}
$$

Proof of Lemma A2. For simplicity, we rewrite $\left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{W}}^{(r-1)}\right)^T, \left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)^T\right)^T = \left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r-1)}\right)^T, \ldots, \left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r-1)}\right)^T, \left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)^T\right)^T$ in the following proof.

By definition

$$\begin{aligned}&\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right) - \nabla\mathcal{L}\left(\boldsymbol{\Upsilon}_{\boldsymbol{W}}^{(r-1)}, \boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)\\ &\quad= \int_0^1 \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) dt \left(\boldsymbol{\Delta}^* - \boldsymbol{\Upsilon}^{(r-1)}\right).\end{aligned}$$

Since $\left(\boldsymbol{\Upsilon}_{\boldsymbol{W}}^{(r-1)}, \boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)$ is the solution of Equation (16), we have $\nabla\mathcal{L}\left(\boldsymbol{\Upsilon}_{\boldsymbol{W}}^{(r-1)}, \boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{R}_{\boldsymbol{\Theta}}\right]} = 0$.

Therefore

$$\begin{aligned}\nabla\mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}\right]} &= \int_0^1 \boldsymbol{P}^T_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) dt \left(\boldsymbol{\Delta}^* - \boldsymbol{\Upsilon}^{(r-1)}\right)\\ &= \int_0^1 \boldsymbol{P}^T_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) \boldsymbol{P}_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} dt\\ &\quad\times \left(\boldsymbol{\Delta}^* - \boldsymbol{\Upsilon}^{(r-1)}\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}\right]}\\ &\quad+ \int_0^1 \boldsymbol{P}^T_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) \boldsymbol{P}_{\bigcup_{i=1}^p \left(\mathcal{T}_i^{(r-1)}\right)^c} dt\\ &\quad\times \left(\boldsymbol{\Delta}^* - \boldsymbol{\Upsilon}^{(r-1)}\right)_{\left[\bigcup_{i=1}^p \left(\mathcal{T}_i^{(r-1)}\right)^c\right]}. \qquad \text{(A7)}\end{aligned}$$

Since $\mathcal{L}$ has μ_{4s}-GSRH and $\mathcal{T}_i^{(r-1)} \cup \operatorname{supp}\left(t\boldsymbol{w}_i^{(r-1)} + (1-t)\boldsymbol{\Upsilon}_{\boldsymbol{w}_i}^{(r-1)}\right) \leq 2s_i$ for all $i = 1, \ldots, p$ and $t \in [0, 1]$, functions $A_{4s}(.)$ and $B_{4s}(.)$, which are defined by Equation (7), exist such that

$$B_{4s}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) \leq \lambda_{\min}\left(\boldsymbol{P}^T_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) \boldsymbol{P}_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}}\right),$$

and

$$A_{4s}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) \geq \lambda_{\max}\left(\boldsymbol{P}^T_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) \boldsymbol{P}_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}}\right).$$

Thus, we can apply Proposition A1 here and obtain

$$\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right) \leq \lambda_{\min}\left(\int_0^1 \boldsymbol{P}^T_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} \boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^* + (1-t)\boldsymbol{\Upsilon}^{(r-1)}\right) \boldsymbol{P}_{\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}} dt\right),$$

and

$$\alpha_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)\geq\lambda_{\max}\left(\int_{0}^{1}\boldsymbol{P}^{T}_{\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}}\boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{*}+(1-t)\boldsymbol{\Upsilon}^{(r-1)}\right)\boldsymbol{P}_{\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}}dt\right).$$

This indicates that the matrix $\int_0^1 \boldsymbol{P}^T_{\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}}\boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{*}+(1-t)\boldsymbol{\Upsilon}^{(r-1)}\right)\boldsymbol{P}_{\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}}dt$, denoted by $\boldsymbol{G}$, is positive-definite. Hence, it is invertible and

$$\frac{1}{\alpha_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}\leq\lambda_{\min}\left(\boldsymbol{G}^{-1}\right)\leq\lambda_{\max}(\boldsymbol{G}^{-1})\leq\frac{1}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}. \tag{A8}$$

Multiplying both sides of Equation (A7) by $\boldsymbol{G}^{-1}$, we get

$$\begin{aligned}
&\boldsymbol{G}^{-1}\nabla\mathcal{L}\left(\boldsymbol{W}^{*},\boldsymbol{\Theta}^{*}\right)_{\left[\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}\right]}\\
&\quad=\left(\boldsymbol{\Delta}^{*}-\boldsymbol{\Upsilon}^{(r-1)}\right)_{\left[\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}\right]}+\boldsymbol{G}^{-1}\\
&\qquad\times\left(\int_{0}^{1}\boldsymbol{P}^{T}_{\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup\mathcal{R}_{\boldsymbol{\Theta}}}\boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{*}+(1-t)\boldsymbol{\Upsilon}^{(r-1)}\right)\boldsymbol{P}_{\bigcup_{i=1}^{p}\left(\mathcal{T}_{i}^{(r-1)}\right)^{c}}dt\right)\\
&\qquad\times\left(\boldsymbol{w}^{*T}_{1\left[\left(\mathcal{T}_{1}^{(r-1)}\right)^{c}\right]},\ldots,\boldsymbol{w}^{*T}_{p\left[\left(\mathcal{T}_{p}^{(r-1)}\right)^{c}\right]}\right)^{T},
\end{aligned}$$

where we derive the last equality from the fact that $\left(\boldsymbol{\Delta}^{*}-\boldsymbol{\Upsilon}^{(r-1)}\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{T}_{i}^{(r-1)}\right)^{c}\right]}=\operatorname{vec}\left(\boldsymbol{W}^{*}-\boldsymbol{\Upsilon}_{\boldsymbol{W}}^{(r-1)}\right)^{T}_{\left[\bigcup_{i=1}^{p}\left(\mathcal{T}_{i}^{(r-1)}\right)^{c}\right]}=\left(\boldsymbol{w}^{*T}_{1\left[\left(\mathcal{T}_{1}^{(r-1)}\right)^{c}\right]},\ldots,\boldsymbol{w}^{*T}_{p\left[\left(\mathcal{T}_{p}^{(r-1)}\right)^{c}\right]}\right)^{T}$, because $\left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r-1)}\right)^{T}_{\left[\left(\mathcal{T}_{1}^{(r-1)}\right)^{c}\right]},\ldots,\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r-1)}\right)^{T}_{\left[\left(\mathcal{T}_{p}^{(r-1)}\right)^{c}\right]}\right)^{T}=0$.

By (A8) and Proposition A2, we obtain

$$\begin{aligned}
&\left\|\left(\boldsymbol{w}^{*T}_{1\left[\mathcal{T}_{1}^{(r-1)}\right]},\ldots,\boldsymbol{w}^{*T}_{p\left[\mathcal{T}_{p}^{(r-1)}\right]},(\boldsymbol{\Theta}^{*})^{T}\right)^{T}-\left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r-1)}\right)^{T},\ldots,\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r-1)}\right)^{T},\left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)^{T}\right)^{T}\right\|_{2}\\
&\quad=\left\|\left(\boldsymbol{\Delta}^{*}-\boldsymbol{\Upsilon}^{(r-1)}\right)_{\left[\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}\right]}\right\|\\
&\quad\leq\left\|\boldsymbol{G}^{-1}\nabla\mathcal{L}\left(\boldsymbol{W}^{*},\boldsymbol{\Theta}^{*}\right)_{\left[\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}\right]}\right\|_{2}\\
&\qquad+\left\|\boldsymbol{G}^{-1}\left(\int_{0}^{1}\boldsymbol{P}^{T}_{\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}}\boldsymbol{H}_{\mathcal{L}}\left(t\boldsymbol{\Delta}^{*}+(1-t)\boldsymbol{\Upsilon}^{(r-1)}\right)\boldsymbol{P}_{\bigcup_{i=1}^{p}\left(\mathcal{T}_{i}^{(r-1)}\right)^{c}}dt\right)\right.
\end{aligned}$$

$$\times \left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T} \right)^T \Bigg\|_2$$

$$\leq \frac{\left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}\right]} \right\|_2}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)} + \frac{\gamma_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)}$$

$$\times \left\| \left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T} \right)^T \right\|_2 .$$

■

Lemma A3. *The estimation error in the* $(r-1)$*th iteration,* $\|\boldsymbol{\Upsilon}^{(r-1)} - \boldsymbol{\Delta}^*\|_2$*, and that in the rth iteration,* $\|\boldsymbol{\Upsilon}^{(r)} - \boldsymbol{\Delta}^*\|_2$*, are related by the inequality*

$$\left\|\boldsymbol{\Delta}^{(r)} - \boldsymbol{\Delta}^*\right\|_2 \leq \frac{\gamma_{4s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right) + \gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)} \left(1 + \frac{\gamma_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)}\right) \times \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2$$

$$+ \frac{2\left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p \mathcal{T}_i^{(r-1)}\right) \cup \mathcal{S}_{\boldsymbol{\Theta}}\right]} \right\|_2}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)} + \left(1 + \frac{\gamma_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)}, \boldsymbol{\Delta}^*\right)}\right)$$

$$\times \frac{\left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^p \left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)} \setminus \mathcal{V}_i^{(r-1)}\right)\right]} \right\|_2 + \left\| \nabla \mathcal{L}\left(\boldsymbol{W}^*, \boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^p \left(\mathcal{V}_i^{(r-1)} \setminus \mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]} \right\|_2}{\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}.\boldsymbol{\Delta}^*\right)}$$

Proof of Lemma A3. Since $\mathcal{V}_i^{(r-1)} \subset \mathcal{T}_i^{(r-1)}$, we have $\left(\mathcal{T}_i^{(r-1)}\right)^c \subset \left(\mathcal{V}_i^{(r-1)}\right)^c$ for $i = 1, \ldots, p$. Therefore

$$\left\| \left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T} \right)^T \right\|_2$$

$$= \left\| \left(\left(\boldsymbol{w}_1^{(r-1)} - \boldsymbol{w}^*\right)_{\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^T, \ldots, \left(\boldsymbol{w}_p^{(r-1)} - \boldsymbol{w}^*\right)_{\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^T \right)^T \right\|_2$$

$$\leq \left\| \left(\left(\boldsymbol{w}_1^{(r-1)} - \boldsymbol{w}^*\right)_{\left[\left(\mathcal{V}_1^{(r-1)}\right)^c\right]}^T, \ldots, \left(\boldsymbol{w}_p^{(r-1)} - \boldsymbol{w}^*\right)_{\left[\left(\mathcal{V}_p^{(r-1)}\right)^c\right]}^T \right)^T \right\|_2 .$$

Applying Lemma A1, we have

$$\left\| \left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T}, \ldots, \boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T} \right)^T \right\|_2$$

$$\leq \left\|\boldsymbol{\Delta}^{(r-1)} - \boldsymbol{\Delta}^*\right\|_2 \frac{\gamma_{4s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right) + \gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}{2\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)}, \boldsymbol{\Delta}^*\right)}$$

$$+\frac{\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*,\boldsymbol{\Theta}^*\right)_{\left[\bigcup_{i=1}^p\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\setminus\mathcal{V}_i^{(r-1)}\right)\right]}\right\|_2+\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*,\boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p\mathcal{V}_i^{(r-1)}\setminus\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]}\right\|_2}{\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)},\boldsymbol{\Delta}^*\right)}. \quad \text{(A9)}$$

Furthermore

$$\begin{aligned}
&\left\|\boldsymbol{\Delta}^{(r)}-\boldsymbol{\Delta}^*\right\|_2\\
&\leq\left\|\left(\left(\boldsymbol{w}_1^{(r)}\right)^T,\ldots,\left(\boldsymbol{w}_p^{(r)}\right)^T,\left(\boldsymbol{\Theta}^{(r)}\right)^T\right)^T-\left(\boldsymbol{w}_{1\left[\mathcal{T}_1^{(r-1)}\right]}^{*T},\ldots,\boldsymbol{w}_{p\left[\mathcal{T}_p^{(r-1)}\right]}^{*T},\left(\boldsymbol{\Theta}^*\right)^T\right)^T\right\|_2\\
&\quad+\left\|\left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T},\ldots,\boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T}\right)^T\right\|_2\\
&\leq\left\|\left(\boldsymbol{w}_{1\left[\mathcal{T}_1^{(r-1)}\right]}^{*T},\ldots,\boldsymbol{w}_{p\left[\mathcal{T}_p^{(r-1)}\right]}^{*T},(\boldsymbol{\Theta}^*)^T\right)^T-\left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r-1)}\right)^T,\ldots,\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r-1)}\right)^T,\left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)^T\right)^T\right\|_2\\
&\quad+\left\|\boldsymbol{\Delta}^{(r)}-\boldsymbol{\Upsilon}^{(r-1)}\right\|_2+\left\|\left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T},\ldots,\boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T}\right)^T\right\|_2\\
&\leq 2\left\|\left(\boldsymbol{w}_{1\left[\mathcal{T}_1^{(r-1)}\right]}^{*T},\ldots,\boldsymbol{w}_{p\left[\mathcal{T}_p^{(r-1)}\right]}^{*T},(\boldsymbol{\Theta}^*)^T\right)^T-\left(\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_1}^{(r-1)}\right)^T,\ldots,\left(\boldsymbol{\Upsilon}_{\boldsymbol{w}_p}^{(r-1)}\right)^T,\left(\boldsymbol{\Upsilon}_{\boldsymbol{\Theta}}^{(r-1)}\right)^T\right)^T\right\|_2\\
&\quad+\left\|\left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T},\ldots,\boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T}\right)^T\right\|_2,
\end{aligned}$$

where the last inequality holds because $\left\|\boldsymbol{w}_{i\left[\mathcal{T}_i^{(r-1)}\right]}^*\right\|_0\leq s_i$ and $\boldsymbol{w}_i^{(r)}$ is the best s_i-term approximation to $\boldsymbol{\Upsilon}_{\boldsymbol{w}_i}^{(r-1)}$, i.e., keeps the s largest absolute values of $\boldsymbol{\Upsilon}_{\boldsymbol{w}_i}^{(r-1)}$ for $i=1,\ldots,p$. Thus, following Lemma A2

$$\begin{aligned}
\left\|\boldsymbol{\Delta}^{(r)}-\boldsymbol{\Delta}^*\right\|_2\leq&\frac{2\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^*,\boldsymbol{\Theta}^*\right)_{\left[\left(\bigcup_{i=1}^p\mathcal{T}_i^{(r-1)}\right)\cup S_{\boldsymbol{\Theta}}\right]}\right\|_2}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^*\right)}\\
&+\left(1+\frac{\gamma_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^*\right)}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^*\right)}\right)\left\|\left(\boldsymbol{w}_{1\left[\left(\mathcal{T}_1^{(r-1)}\right)^c\right]}^{*T},\ldots,\boldsymbol{w}_{p\left[\left(\mathcal{T}_p^{(r-1)}\right)^c\right]}^{*T}\right)^T\right\|_2. \quad \text{(A10)}
\end{aligned}$$

Combining (A9) and (A10), we get

$$\left\|\boldsymbol{\Delta}^{(r)}-\boldsymbol{\Delta}^*\right\|_2$$

$$\leq \left(1+\frac{\gamma_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}\right)\frac{\gamma_{4s}\left(\boldsymbol{\Delta}^{(r-1)},\boldsymbol{\Delta}^{*}\right)+\gamma_{2s}\left(\boldsymbol{\Delta}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}{2\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}\times\left\|\boldsymbol{\Delta}^{(r-1)}-\boldsymbol{\Delta}^{*}\right\|_{2}$$
$$+\frac{2\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^{*},\boldsymbol{\Theta}^{*}\right)_{\left[\left(\bigcup_{i=1}^{p}\mathcal{T}_{i}^{(r-1)}\right)\cup\mathcal{S}_{\boldsymbol{\Theta}}\right]}\right\|_{2}}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}+\left(1+\frac{\gamma_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}{\beta_{4s}\left(\boldsymbol{\Upsilon}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}\right)$$
$$\times\frac{\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^{*},\boldsymbol{\Theta}^{*}\right)_{\left[\bigcup_{i=1}^{p}\left(\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\setminus\mathcal{V}_{i}^{(r-1)}\right)\right]}\right\|_{2}+\left\|\nabla\mathcal{L}\left(\boldsymbol{W}^{*},\boldsymbol{\Theta}^{*}\right)_{\left[\left(\bigcup_{i=1}^{p}\mathcal{V}_{i}^{(r-1)}\setminus\mathcal{R}_{\boldsymbol{w}_i}^{(r-1)}\right)\right]}\right\|_{2}}{\beta_{2s}\left(\boldsymbol{\Delta}^{(r-1)},\boldsymbol{\Delta}^{*}\right)}.$$

■

Received 1 August 2020
Accepted 6 July 2021