The Reliability of OKRidge Method in Solving Sparse Ridge Regression Problems

Xiyuan Li [1] Youjun Wang [1] Weiwei Liu [1]

[1]School of Computer Science, Wuhan University

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Sparse ridge regression problems play a significant role across various domains. To solve sparse ridge regression, [1] recently proposes an advanced algorithm, Scalable Optimal K-Sparse Ridge Regression (OKRidge), which is both faster and more accurate than existing approaches. However, the absence of theoretical analysis on the error of OKRidge impedes its large-scale applications. In this paper, we reframe the estimation error of OKRidge as a Primary Optimization (**PO**) problem and employ the Convex Gaussian min-max theorem (CGMT) to simplify the **PO** problem into an Auxiliary Optimization (**AO**) problem. Subsequently, we provide a theoretical error analysis for OKRidge based on the **AO** problem. This error analysis improves the theoretical reliability of OKRidge. We also conduct experiments to verify our theorems and the results are in excellent agreement with our theoretical findings.

Sparse Ridge Regression (SRR)

In this paper, we are interested in addressing the following k-sparse linear regression problem with additive noise:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \quad \text{with} \quad \|\boldsymbol{\beta}^*\|_0 \le k, \tag{1}$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ represents the "true" weight parameter, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times d}$ is the input measurement matrix, $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^\top \in \mathbb{R}^n$ is the real output responses, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^\top \in \mathbb{R}^n$ is the noise vector, $k \in \mathbb{Z}^+$ specifies the maximum number of nonzero elements for the model, $\|\cdot\|_0$ denotes the number of nonzero elements of the given vector. Moreover, the entries of $\boldsymbol{X}$ are drawn i.i.d. from $\mathcal{N}(0,1)$; the entries of $\boldsymbol{\epsilon}$ are drawn i.i.d. from $\mathcal{N}(0,\sigma^2)$; and we assume $\frac{k}{d}$ is a constant and $\lim_{d\to\infty} \frac{n(d)}{d} = \delta \in (0,1)$.

The formulation (1) represents a black box model where $\boldsymbol{\beta}^*$ is fixed. Given $\boldsymbol{X}$ and $\boldsymbol{y}$, to determine the target vector $\boldsymbol{\beta}^*$, the most basic method is solving the following k-Sparse Ridge Regression Optimization (k-SRO), as outlined by [1]:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \le k, \tag{2}$$

where $\lambda > 0$ is a regularizer parameter, and $\|\cdot\|_2$ denotes the Euclidean norm. Our paper focuses on the worst-case scenario $\|\boldsymbol{\beta}^*\|_0 = k$. This k-SRO is different from the traditional ridge regression due to the constraint of k-sparse structure for $\boldsymbol{\beta}$. The k-SRO problem (2) is NP-hard, and is more challenging in the presence of highly correlated features.

The Convex Gaussian Min-max Theorem (CGMT)

Definition 3.1[GMT admissible sequence] *The sequence $\{\boldsymbol{G}^{(d)}, \boldsymbol{g}^{(d)}, \boldsymbol{h}^{(d)}, \mathcal{S}_{\boldsymbol{w}}^{(d)}, \mathcal{S}_{\boldsymbol{u}}^{(d)}, \psi^{(d)}\}_{d\in\mathbb{N}}$ indexed by d, with $\boldsymbol{G}^{(d)} \in \mathbb{R}^{n\times d}$, $\boldsymbol{g}^{(d)} \in \mathbb{R}^n$, $\boldsymbol{h}^{(d)} \in \mathbb{R}^d$, $\mathcal{S}_{\boldsymbol{w}}^{(d)} \subset \mathbb{R}^d$, $\mathcal{S}_{\boldsymbol{u}}^{(d)} \subset \mathbb{R}^n$, $\psi^{(d)} : \mathcal{S}_{\boldsymbol{w}}^{(d)} \times \mathcal{S}_{\boldsymbol{u}}^{(d)} \to \mathbb{R}$ and $n = n(d)$, is said to be admissible if, for each $d \in \mathbb{N}$, $\mathcal{S}_{\boldsymbol{w}}^{(d)}$ and $\mathcal{S}_{\boldsymbol{u}}^{(d)}$ are compact sets and $\psi^{(d)}$ is continuous on its domain. Onwards, we will drop the superscript (d) from $\boldsymbol{G}^{(d)}$, $\boldsymbol{g}^{(d)}$, $\boldsymbol{h}^{(d)}$.*

A sequence $\{\boldsymbol{G}^{(d)}, \boldsymbol{g}^{(d)}, \boldsymbol{h}^{(d)}, \mathcal{S}_{\boldsymbol{w}}^{(d)}, \mathcal{S}_{\boldsymbol{u}}^{(d)}, \psi^{(d)}\}_{d\in\mathbb{N}}$ defines a sequence of min-max problems

$$\Phi^{(d)}(\boldsymbol{G}) := \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}^{(d)}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}^{(d)}} \boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{w} + \psi^{(d)}(\boldsymbol{w},\boldsymbol{u}), \tag{3}$$

$$\phi^{(d)}(\boldsymbol{g},\boldsymbol{h}) := \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}^{(d)}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}^{(d)}} \|\boldsymbol{w}\|_2\boldsymbol{g}^\top\boldsymbol{u} + \|\boldsymbol{u}\|_2\boldsymbol{h}^\top\boldsymbol{w} + \psi^{(d)}(\boldsymbol{w},\boldsymbol{u}). \tag{4}$$

Importantly, the formulation (3) is called Primary Optimization (**PO**) and the formulation (4) is called Auxiliary Optimization (**AO**). Based on the GMT admissible sequence and the notation introduced above, we present the CGMT below.

Theorem 3.2 [CGMT[2]] *Let $\{\boldsymbol{G}^{(d)}, \boldsymbol{g}^{(d)}, \boldsymbol{h}^{(d)}, \mathcal{S}_{\boldsymbol{w}}^{(d)}, \mathcal{S}_{\boldsymbol{u}}^{(d)}, \psi^{(d)}\}_{d\in\mathbb{N}}$ be a GMT admissible sequence as in Definition 1, for which additionally the entries of $\boldsymbol{G}$, $\boldsymbol{g}$, $\boldsymbol{h}$ are drawn i.i.d. from $\mathcal{N}(0,1)$. Let $\Phi^{(d)}(\boldsymbol{G})$, $\phi^{(d)}(\boldsymbol{g},\boldsymbol{h})$ be the optimal costs, and, $\boldsymbol{w}_\Phi^{(d)}(\boldsymbol{G})$, $\boldsymbol{w}_\phi^{(d)}(\boldsymbol{g},\boldsymbol{h})$ the corresponding optimal minimizers of the **PO** and **AO** problems in (3) and (4). The following three statements hold*

(i) *For any $d \in \mathbb{N}$ and $c \in \mathbb{R}$,*

$$\mathbb{P}(\Phi^{(d)}(\boldsymbol{G}) < c) \le 2\mathbb{P}(\phi^{(d)}(\boldsymbol{g},\boldsymbol{h}) \le c).$$

(ii) *For any $d \in \mathbb{N}$. If $\mathcal{S}_{\boldsymbol{w}}^{(d)}$, $\mathcal{S}_{\boldsymbol{u}}^{(d)}$ are convex, and, $\psi^{(d)}(\cdot,\cdot)$ is convex-concave on $\mathcal{S}_{\boldsymbol{w}}^{(d)} \times \mathcal{S}_{\boldsymbol{u}}^{(d)}$, then, for any $\mu \in \mathbb{R}$ and $t > 0$,*

$$\mathbb{P}\big(|\Phi^{(d)}(\boldsymbol{G}) - \mu|) > t\big) \le 2\mathbb{P}\big(|\phi^{(d)}(\boldsymbol{g},\boldsymbol{h}) - \mu|) > t\big).$$

(iii) *Assume the conditions of (ii) hold for all $d \in \mathbb{N}$. Let $\|\cdot\|$ denote some norm in $\mathbb{R}^d$. If, there exist constants (independent of d) κ^*, α^* and $\tau > 0$ such that*

(a) $\phi^{(d)}(\boldsymbol{g},\boldsymbol{h}) \xrightarrow{P} \kappa^*$, (b) $\|\boldsymbol{w}_\phi^{(d)}(\boldsymbol{g},\boldsymbol{h})\| \xrightarrow{P} \alpha^*$, (c) *with probability one in the limit $d \to \infty$*

$$\Big\{\upsilon^{(d)}(\boldsymbol{w};\boldsymbol{g},\boldsymbol{h}) \ge \phi^{(d)}(\boldsymbol{g},\boldsymbol{h}) + \tau\big(\|\boldsymbol{w}\| - \boldsymbol{w}_\phi^{(d)}(\boldsymbol{g},\boldsymbol{h})\big)^2, \forall \boldsymbol{w} \in \mathcal{S}_{\boldsymbol{w}}^{(d)}\Big\},$$

then,

$$\|\boldsymbol{w}_\Phi^{(d)}(\boldsymbol{G})\| \xrightarrow{P} \alpha^*. \tag{5}$$

The OKRidge Method for solving SRR

In order to rapidly solve k-SRO problem (2) while ensuring solution optimality, [1] introduces a highly efficient method called OKRidge. Specifically, the optimization (2) can be relaxed as:

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \mathcal{L}_{\text{ridge}}^{\text{saddle}}(\boldsymbol{\beta},\boldsymbol{z}), \quad \text{s.t.} \quad \sum_{j=1}^{d} z_j \le k, \; z_j \in [0,1], \tag{6}$$

where $\mathcal{L}_{\text{ridge}}^{\text{saddle}}(\boldsymbol{\beta},\boldsymbol{z}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1,z_j\neq 0}^{d} \frac{\beta_j^2}{z_j}$. We define a new function $\mathcal{L}(\boldsymbol{\beta})$ as:

$$\mathcal{L}(\boldsymbol{\beta}) = \min_{\boldsymbol{z}} \mathcal{L}_{\text{ridge}}^{\text{saddle}}(\boldsymbol{\beta},\boldsymbol{z}), \quad \text{s.t.} \quad \sum_{j=1}^{d} z_j \le k, \; z_j \in [0,1]. \tag{7}$$

For any $\boldsymbol{\beta}$, $\mathcal{L}(\boldsymbol{\beta})$ serves as a valid lower bound for problem (6). Then, we choose $\boldsymbol{z}$ such that this lower bound $\mathcal{L}(\boldsymbol{\beta})$ is tight.

Theorem 4.2 *The function $\mathcal{L}(\boldsymbol{\beta})$ defined in Equation (7) is lower bounded by*

$$\mathcal{L}(\boldsymbol{\beta}) \ge \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta}). \tag{8}$$

where $\odot$ is Hadamard product, and $\text{SumTop}_k(\cdot)$ denotes the summation of the largest k elements of a given vector.

If we define

$$\mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta}),$$

OKRidge solves k-SRO problem (2) with

$$\min_{\boldsymbol{\beta}} \mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta}). \tag{9}$$

So far, we transform the constrained k-SRO problem (2) into the unconstrained optimization problem (9). Let $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta})$, OKRidge regards $\hat{\boldsymbol{\beta}}$ as the estimation of $\boldsymbol{\beta}^*$ in problem (1). Next, we apply CGMT to analyze the error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ for OKRidge by transforming the optimization (9) into a **PO** problem.

The Error Analysis for OKRidge

Based on (2), the estimation error of OKRidge can be obtained by normalized **AO** problem:

$$\min_{\boldsymbol{w}} \frac{1}{\sqrt{n}}\Big[\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{\epsilon}\|_2^2 + \lambda \text{SumTop}_k\big((\boldsymbol{w}+\boldsymbol{\beta}^*) \odot (\boldsymbol{w}+\boldsymbol{\beta}^*)\big)\Big], \tag{10}$$

where $\boldsymbol{w} := \boldsymbol{\beta} - \boldsymbol{\beta}^*$, and the estimation error can be measured by $\|\boldsymbol{w}\|_2$. Subsequently, we transform the optimization (10) into **PO** (11) about the error of OKRidge, using the Fenchel-Moreau theorem.

$$\max_{\boldsymbol{u}} \frac{1}{\sqrt{n}}\Big[\boldsymbol{u}^\top\boldsymbol{X}\boldsymbol{w} - \boldsymbol{u}^\top\boldsymbol{\epsilon} - \frac{\|\boldsymbol{u}\|_2^2}{4} + \lambda \text{SumTop}_k\big((\boldsymbol{w}+\boldsymbol{\beta}^*) \odot (\boldsymbol{w}+\boldsymbol{\beta}^*)\big)\Big], \tag{11}$$

Then, we employ the CGMT framework to substitute the complex **PO** problem with a simplified **AO** problem (12) that only involves two scalar variables: α and η.

$$\max_{\eta\ge 0} \min_{\alpha\ge 0} \eta\sqrt{\alpha^2+\sigma^2} - \alpha\eta\sqrt{\bar{D}(\frac{\lambda}{\eta})} - \Gamma(\eta). \tag{12}$$

where $\alpha = \|\boldsymbol{w}\|_2$ and $\eta = \|\boldsymbol{u}\|_2$. Finally, we present the following theoretical error analysis of OKRidge based on the **AO** problem (12).

Theorem 5.2 *Suppose $\boldsymbol{\beta}^*$ is the true weight parameter of the problem (1), $\hat{\boldsymbol{\beta}}$ is the optimal solution to the objective function (9) of OKRidge, $\frac{D(\tau)}{n} \to \bar{D}(\tau) \in (0,1)$, aNSE $:= \lim_{\sigma^2\to 0}$ NSE $= \lim_{\sigma^2\to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2/\sigma^2$. Define λ_{map} is the solution of $map(\tau) = 0$ for $\tau > 0$, then, the estimation error of OKRidge is given by the following probability limit:*

$$\lim_{d\to 0} aNSE \xrightarrow{P} \Delta(\hat{\lambda}), \tag{13}$$

where $\Delta(\hat{\lambda}) = \frac{\bar{D}(\hat{\lambda})}{1-\bar{D}(\hat{\lambda})}$, and $\hat{\lambda} = \lambda_{map}$.

Numerical Experiments

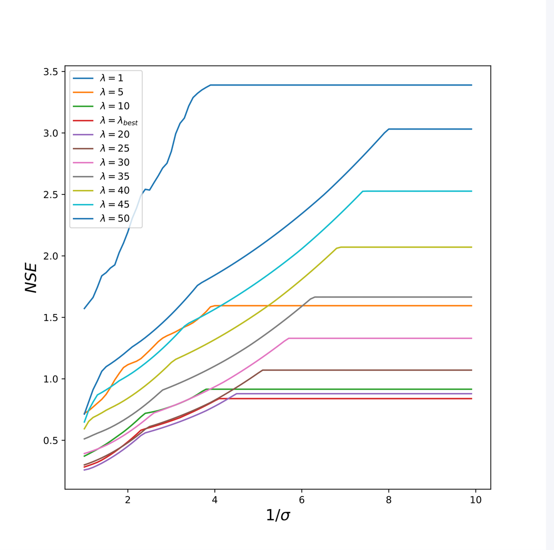

Figure 1. The change of NSE with $1/\sigma$ for OKRidge under different λ. The red curve at the bottom corresponds to the case $\lambda = \lambda_{best}$.

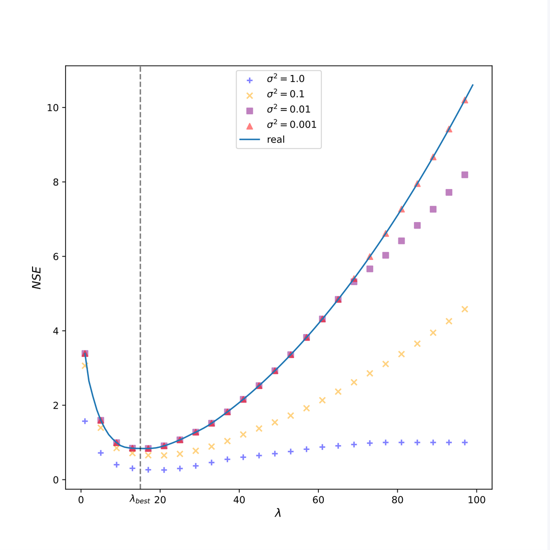

Figure 2. The change of NSE with λ for OKRidge under different σ. The blue curve corresponds to the real change of $\Delta(\hat{\lambda})$. Here, λ_{best} is the optimal weight of the regularizer.

References

[1] Jiachang Liu, Sam Rosen, Chudi Zhong, and Cynthia Rudin. Okridge: Scalable optimal k-sparse ridge regression for learning dynamical systems. In *NeurIPS*, 2023.

[2] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Emails: Lee_xiyuan@outlook.com; youjunw1208@gmail.com; liuweiwei863@gmail.com