The Reliability of OKRidge Method in Solving Sparse Ridge Regression Problems

Xiyuan Li Youjun Wang Weiwei Liu*
School of Computer Science, Wuhan University
National Engineering Research Center for Multimedia Software, Wuhan University
Institute of Artificial Intelligence, Wuhan University
Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University
Lee_xiyuan@outlook.com, youjunw1208@gmail.com, liuweiwei863@gmail.com

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.

1 Introduction

Sparse Ridge Regression (SRR) has achieved notable success across various machine learning applications, including statistics [2], signal processing [3], dynamical systems [4, 5], and others. In this paper, we are interested in addressing the following k-sparse linear regression problem with additive noise:

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

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, 6, 7]:

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

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

*Corresponding author: Weiwei Liu (liuweiwei863@gmail.com).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-2420

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 [8].

Two main types of algorithms are commonly employed for solving k-SRO problem (2): heuristic algorithms [9, 10] and optimal algorithms [11]. However, heuristic algorithms lack the ability to assess the solution quality, while the optimal algorithms are slow. In order to rapidly solve k-SRO problem (2) while ensuring solution optimality, [1] introduces a highly efficient method called OKRidge. Therefore, a complete algorithm of OKRidge, including how to choose hyper-parameters can be seen in the original paper [1]. OKRidge substitutes k-SRO problem (2) with an unconstrained optimization on a novel tight lower bound. The experiment results in [1] show that OKRidge is superior to heuristic algorithms, optimal algorithms, and existing mixed-integer programming (MIP) formulations solved by the commercial solver Gurobi. Nevertheless, the absence of theoretical error analysis for OKRidge impedes its scalability in practical applications.

In this paper, we provide theoretical error analysis for OKRidge utilizing the framework of the CGMT [12]. Specifically, we propose another novel tight lower bound $\mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta})$ to replace k-SRO problem (2):

$$\mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta}), \tag{3}$$

where $\odot$ denotes Hadamard product, and $\text{SumTop}_k(\cdot)$ represents the summation of the largest k elements of a given vector. The tight lower bound (3) is equivalent to that proposed by [1]. Thus, $\mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta})$ can replace the objective function of OKRidge. It is noteworthy that our proposed regularizer, defined as $\gamma(\boldsymbol{\beta}) = \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta})$, differs from any previously proposed instances by [12]. Then, the optimal solution obtained by OKRidge is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta}). \tag{4}$$

[1] utilizes $\hat{\boldsymbol{\beta}}$ as the estimate of $\boldsymbol{\beta}^*$ in problem (1). By combining formulations (1) and (3), the estimation error of OKRidge can be obtained through the following normalized optimization problem:

$$\min_{\boldsymbol{w}} \frac{1}{\sqrt{n}} \Big[\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{\epsilon}\|_2^2 + \lambda \text{SumTop}_k\big((\boldsymbol{w} + \boldsymbol{\beta}^*) \odot (\boldsymbol{w} + \boldsymbol{\beta}^*)\big) \Big], \tag{5}$$

where $\boldsymbol{w} := \boldsymbol{\beta} - \boldsymbol{\beta}^*$ is a random variable with randomness from the random variables $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$, and the estimation error can be measured by $\|\boldsymbol{w}\|_2$. Subsequently, we transform the optimization (5) into a **PO** problem about the error of OKRidge, using the Fenchel-Moreau theorem [13]. Then, we employ the CGMT framework to substitute the complex **PO** problem with a simplified **AO** problem. Finally, we present the theoretical error analysis of OKRidge based on the **AO** problem. Our theoretical results focus on the Normalized Squared Error (NSE) of OKRidge and can be summarized as:

$$\lim_{d \to \infty} \lim_{\sigma \to 0} \text{NSE} \xrightarrow{P} \Delta(\hat{\lambda}), \tag{6}$$

where $\text{NSE} := \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 / \sigma^2$, and $\Delta(\hat{\lambda})$ is a function of λ. These theoretical results indicate that if the regularizer parameter λ used in OKRidge is constant, the NSE limit of OKRidge is also fixed. Moreover, $\hat{\boldsymbol{\beta}}$ learned by OKRidge is reliable to estimate $\boldsymbol{\beta}^*$, due to $\lim_{d\to\infty} \lim_{\sigma\to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \xrightarrow{P} 0$. The comprehensive experiments of OKRidge on real-world examples were conducted by the NeurIPS 2023 paper [1] (see Figure 3 and Appendix H in [1]), which demonstrates that the error of OKRidge tends to zero. Our analysis explains the experimental phenomenon observed in [1], strengthens the theoretical underpinnings of OKRidge, and provides theoretical reliability for its broad application.

We also conduct numerical experiments to validate our theorems. The findings demonstrate that the NSE converges to a fixed constant determined by λ, aligning excellently with our theoretical predictions.

1.1 Outline

The structure of the remaining sections in this paper is as follows. Section 2 provides a review of related work. Section 3 offers background information on OKRidge, CGMT, and basic concepts. Section 4 introduces an alternative tight lower bound for the objective function of OKRidge. In Section 5, we convert the estimation error of OKRidge into a **PO** problem and simplify it into an **AO** problem using CGMT. Subsequently, an estimation error analysis of OKRidge based on the **AO** problem is conducted. Section 6 presents the experimental results. Finally, we conclude with a summary in Section 7. Additionally, the limitaion and impact of our work are detailed in Appendix A

https://doi.org/10.52202/079017-2420

2 Related work

2.1 Heuristic and Optimal Methods

Heuristic methods approximate solutions to optimization problems based on practical experience [14], including ensemble methods [15], swapping features [16], greedy methods [17], etc. While heuristic methods are fast, they often become trapped in local minima, and their solution quality cannot be assessed due to the absence of a lower bound on performance. Optimal methods aim to precisely solve sparse regression problems, such as the big-M method [18], the conditional-value-at-risk (CVaR) approach [19], big-M free mixed integer second order conic (MISOC) method [19], and so on. However, exact optimal methods are slow, particularly for large instances, to achieve near-optimality [20, 21]. To address the limitations of heuristic and optimal methods, [1] proposes an efficient approximation algorithm, OKRidge. Experimental results in [1] demonstrate that OKRidge outperforms heuristic algorithms, optimal algorithms, and existing MIP formulations solved by the commercial solver Gurobi.

2.2 Lower Bound Methods

Lower bound methods are capable of solving the NP-hard k-SRO problems. Several algorithms utilize the lower bound method, such as SOS1 formulation [18], big-M formulation [18], Subset Selection CIO method [22], and others. However, the SOS1 formulation lacks scalability in high dimensions, the big-M formulation is sensitive to hyperparameters, and the Subset Selection CIO method runs slowly. Recently, the perspective formulation [6, 23, 24] has been employed to induce a convex relaxed lower bound that is easier to solve. Building upon the perspective formulation, [1] proposes a novel lower bound used as the objective function for the OKRidge method.

2.3 Normalized Squared Error

NSE, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2/\sigma^2$, serves as a natural measure of the estimation error. NSE is an important indicator in signal-to-noise ratio scenes [25, 26]. Bounds on NSE have been derived by [27, 28]. Additionally, [29] is the first to precisely formulate the limiting behavior of NSE. These studies primarily consider a Gaussian sensing matrix $\boldsymbol{X}$ and utilize the Approximate Message Passing (AMP) framework for analysis [30, 31]. These achievements motivate us to utilize NSE for evaluating the estimation error of the OKRidge method.

3 Preliminary

3.1 Relaxed Transformation of k-SRO

According to [1], the k-SRO problem (2) can be reformulated as the following optimization problem:

$$\min_{\boldsymbol{\beta}} \mathcal{L}_{\text{ridge}}(\boldsymbol{\beta}), \quad \text{s.t.} \begin{cases} (1-z_j)\beta_j = 0, \ j = 1, 2, \cdots, d, \\ \sum_{j=1}^{d} z_j \leq k, \ z_j \in \{0, 1\}, \end{cases} \tag{7}$$

where $\boldsymbol{z} = (z_1, z_2, \cdots, z_n)^\top \in \mathbb{R}^d$, and $\mathcal{L}_{\text{ridge}}(\boldsymbol{\beta}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{d} \beta_j^2$. This problem (7) remains NP-hard under the sparsity constraint [32]. Existing methods such as SOS1, big-M, or the perspective formulation do not leverage the k-sparse structure of the problem. [1] develops a novel method, OKRidge, to preserve the special structure through the following relaxed transformation.

By employing the perspective formulation [33, 34] and the Fenchel conjugate [35], [1] transforms the problem (7) to a new perspective optimization problem:

$$\min_{\boldsymbol{\beta}, \boldsymbol{z}} \max_{\boldsymbol{c}} \mathcal{L}_{\text{ridge}}^{\text{Fenchel}}(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{c}), \quad \text{s.t.} \sum_{j=1}^{d} z_j \leq k, \ z_j \in \{0, 1\}. \tag{8}$$

where $\boldsymbol{c} = (c_1, c_2, \cdots, c_n) \in \mathbb{R}^d$, and $\mathcal{L}_{\text{ridge}}^{\text{Fenchel}}(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{c}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{d} (\beta_j c_j - \frac{c_j^2}{4} z_j)$. This transformation does not change the optimal solution of the problem (7) [1, 35], indicating that problem (7) can be replaced by problem (8). To efficiently solve problem (8), [1] further relaxs

https://doi.org/10.52202/079017-2420

the binary constraint $\{0,1\}$ to the interval $[0,1]$, ultimately yielding the following relaxed convex optimization problem:

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \max_{\boldsymbol{c}} \mathcal{L}^{\text{Fenchel}}_{\text{ridge}}(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{c}), \quad \text{s.t.} \sum_{j=1}^{d} z_j \leq k,\ z_j \in [0,1]. \tag{9}$$

According to problem (7) , to preserve the special sparse structure of $\boldsymbol{\beta}$, we always have $\beta_j = 0$ if $z_j = 0$. Directly solving the min-max problem (9) is computationally challenging. [1] utilizes the relaxed problem (9) to obtain a tight lower bound for the problem (8), where the lower bound corresponds to the objective function of OKRidge.

3.2 The Convex Gaussian Min-max Theorem

The CGMT framework, introduced by [28], has been utilized to analyze the performance of solutions to non-smooth regularized convex optimization problems. It has achieved significant success in various practical applications, including regularized logistic regression [36], max-margin classifiers [37], adversarial training [38, 39] and others. These achievements inspire us to apply the CGMT framework to analyze the NSE of the OKRidge method.

CGMT originates from Gordon's Gaussian Min-max Theorem (GMT) [40], which provides probabilistic bounds on the optimal cost of **PO** problem via a simpler **AO** problem. CGMT further tightens the bounds under convexity assumptions. According to GMT, [41] introduces the following asymptotic sequence and notation.

Definition 3.1 (GMT admissible sequence)**.** The sequence $\{\boldsymbol{G}^{(d)}, \boldsymbol{g}^{(d)}, \boldsymbol{h}^{(d)}, \mathcal{S}^{(d)}_{\boldsymbol{w}}, \mathcal{S}^{(d)}_{\boldsymbol{u}}, \psi^{(d)}\}_{d\in\mathbb{N}}$ indexed by d, with $\boldsymbol{G}^{(d)} \in \mathbb{R}^{n\times d}$, $\boldsymbol{g}^{(d)} \in \mathbb{R}^n$, $\boldsymbol{h}^{(d)} \in \mathbb{R}^d$, $\mathcal{S}^{(d)}_{\boldsymbol{w}} \subset \mathbb{R}^d$, $\mathcal{S}^{(d)}_{\boldsymbol{u}} \subset \mathbb{R}^n$, $\psi^{(d)} : \mathcal{S}^{(d)}_{\boldsymbol{w}} \times \mathcal{S}^{(d)}_{\boldsymbol{u}} \to \mathbb{R}$ and $n = n(d)$, is said to be admissible if, for each $d \in \mathbb{N}$, $\mathcal{S}^{(d)}_{\boldsymbol{w}}$ and $\mathcal{S}^{(d)}_{\boldsymbol{u}}$ 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 $\{\boldsymbol{G}^{(d)}, \boldsymbol{g}^{(d)}, \boldsymbol{h}^{(d)}, \mathcal{S}^{(d)}_{\boldsymbol{w}}, \mathcal{S}^{(d)}_{\boldsymbol{u}}, \psi^{(d)}\}_{d\in\mathbb{N}}$ defines a sequence of min-max problems

$$\Phi^{(d)}(\boldsymbol{G}) := \min_{\boldsymbol{w}\in\mathcal{S}^{(d)}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}^{(d)}_{\boldsymbol{u}}} \boldsymbol{u}^\top \boldsymbol{G}\boldsymbol{w} + \psi^{(d)}(\boldsymbol{w},\boldsymbol{u}), \tag{10}$$

$$\phi^{(d)}(\boldsymbol{g},\boldsymbol{h}) := \min_{\boldsymbol{w}\in\mathcal{S}^{(d)}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}^{(d)}_{\boldsymbol{u}}} \|\boldsymbol{w}\|_2 \boldsymbol{g}^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \boldsymbol{h}^\top \boldsymbol{w} + \psi^{(d)}(\boldsymbol{w},\boldsymbol{u}). \tag{11}$$

Importantly, the formulation (10) is called Primary Optimization **(PO)** and the formulation (11) is called Auxiliary Optimization **(AO)**. Additionally, let $\boldsymbol{w}^{(d)}_{\Phi}(\boldsymbol{G})$ denote the optimal minimizer of **PO** problem (10), and $\boldsymbol{w}^{(d)}_{\phi}(\boldsymbol{g},\boldsymbol{h})$ denote the optimal minimizer of **AO** problem (11) . Define $\upsilon^{(d)} : \mathcal{S}^{(d)}_{\boldsymbol{w}} \to \mathbb{R}$ as follows,

$$\upsilon^{(d)}(\boldsymbol{w};\boldsymbol{g},\boldsymbol{h}) := \max_{\boldsymbol{u}\in\mathcal{S}^{(d)}_{\boldsymbol{u}}} \|\boldsymbol{w}\|_2 \boldsymbol{g}^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \boldsymbol{h}^\top \boldsymbol{w} + \psi^{(d)}(\boldsymbol{w},\boldsymbol{u}). \tag{12}$$

Clearly, $\phi^{(d)}(\boldsymbol{g},\boldsymbol{h}) := \min_{\boldsymbol{w}\in\mathcal{S}^{(d)}_{\boldsymbol{w}}} \upsilon^{(d)}(\boldsymbol{w};\boldsymbol{g},\boldsymbol{h})$. For a sequence of random variables $\{\mathcal{X}^{(d)}\}_{d\in\mathbb{N}}$ and a constant $c \in \mathbb{R}$, $\mathcal{X}^{(d)} \xrightarrow{P} c$ denotes convergence in probability, i.e. $\forall \epsilon > 0, \quad \lim_{d\to\infty} \mathbb{P}\big(|\mathcal{X}^{(d)} - c| > \epsilon\big) = 0$. Based on the GMT admissible sequence and the notation introduced above, we present the CGMT below.

Theorem 3.2 (CGMT [12])**.** *Let $\{\boldsymbol{G}^{(d)}, \boldsymbol{g}^{(d)}, \boldsymbol{h}^{(d)}, \mathcal{S}^{(d)}_{\boldsymbol{w}}, \mathcal{S}^{(d)}_{\boldsymbol{u}}, \psi^{(d)}\}_{d\in\mathbb{N}}$ be a GMT admissible sequence as in Definition 3.1, 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, $\boldsymbol{w}^{(d)}_{\Phi}(\boldsymbol{G})$, $\boldsymbol{w}^{(d)}_{\phi}(\boldsymbol{g},\boldsymbol{h})$ the corresponding optimal minimizers of the **PO** and **AO** problems in (10) and (11). The following three statements hold*

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

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

(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 $t > 0$,

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

(iii) Assume the conditions of (ii) hold for all $d \in \mathbb{N}$. Let $\|\cdot\|$ denote some norm in $\mathbb{R}^d$ and recall (12). 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 $d \to \infty$

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

then,

$$\|\boldsymbol{w}_{\Phi}^{(d)}(\boldsymbol{G})\| \xrightarrow{P} \alpha^*. \tag{13}$$

Theorem 3.2 indicates that, if the optimal cost $\phi(\boldsymbol{g},\boldsymbol{h})$ of (11) concentrates to some value μ, the same holds true for $\Phi(\boldsymbol{G})$ of (10). Furthermore, under appropriate additional assumptions, the optimal solutions of the **AO** and **PO** problems are also closely related by $\|\boldsymbol{w}_{\Phi}(\boldsymbol{G})\| = \|\boldsymbol{w}_{\phi}(\boldsymbol{g},\boldsymbol{h})\|$, as $n \to \infty$. This suggests that, within the CGMT framework, a challenging **PO** problem can be replaced with a simplified **AO** problem, from which the optimal solution of the **PO** problem can be accurately inferred [12]. Subsequently, we rewrite the lower bound of problem (9) in the form of **PO** problem (10) and analyze the minimizer of the simplified **AO** problem instead.

3.3 Basic Concept

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, the Fenchel conjugate of f is defined as $f^*(\boldsymbol{u}) = \sup_{\boldsymbol{v}} \boldsymbol{v}^\top \boldsymbol{u} - f(\boldsymbol{v})$. Additionally, f^* is always convex and lower semi-continuous. According to the Fenchel-Moreau theorem [13], if f is convex and continuous, we have $f(\boldsymbol{v}) = \sup_{\boldsymbol{u}} \boldsymbol{u}^\top \boldsymbol{v} - f^*(\boldsymbol{u})$. In this paper, we utilize the following conjugate pairs

$$f(\boldsymbol{v}) = \|\boldsymbol{v}\|_2^2 \leftrightarrow f^*(\boldsymbol{u}) = \frac{\|\boldsymbol{u}\|_2^2}{4}. \tag{14}$$

If $\gamma(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a convex function of $\boldsymbol{\beta}$, the subdifferential of $\gamma(\cdot)$ at $\boldsymbol{\beta}^*$ is the set of vectors: $\partial\gamma(\boldsymbol{\beta}^*) = \{\boldsymbol{s} \in \mathbb{R}^d \big| \gamma(\boldsymbol{\beta}^* + \boldsymbol{u}) \geq \gamma(\boldsymbol{\beta}^*) + \boldsymbol{s}^\top \boldsymbol{u}\}$. According to [13], $\partial\gamma(\boldsymbol{\beta}^*)$ is nonempty, convex and compact. Given $\boldsymbol{h} \in \mathbb{R}^d$, we define $\text{dist}(\boldsymbol{h}, \partial\gamma(\boldsymbol{\beta}^*)) = \min_{\boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)} \|\boldsymbol{h} - \boldsymbol{s}\|_2$. Then, the Gaussian squared distance corresponding to the scaled subdifferential is defined as $D(\tau) := D_{\partial\gamma(\boldsymbol{\beta}^*)}(\tau) := \mathbb{E}_{\boldsymbol{h}}\big[\text{dist}^2(\boldsymbol{h}, \tau\partial\gamma(\boldsymbol{\beta}^*))\big]$, where $\tau > 0$. Suppose $C(\tau) = -\frac{\tau}{2}\frac{\partial D(\tau)}{\partial \tau}$, $\lim_{d\to\infty} \frac{n}{d} \to \delta \in (0,1)$, $\lim_{d\to\infty} \frac{D(\tau)}{n} \to \bar{D}(\tau) \in (0,1)$, $\lim_{d\to\infty} \frac{C(\tau)}{n} \to \bar{C}(\tau)$. Based on the Gaussian squared distance, we define a map function:

$$\text{map}(\tau) := \frac{1 - \bar{C}(\tau) - \bar{D}(\tau)}{\sqrt{1 - \bar{D}(\tau)}}, \tau > 0. \tag{15}$$

We denote λ_{map} as the solution of $\text{map}(\tau) - \lambda/2 = 0$. Since $map(\tau)$ depends on $\gamma(\cdot)$ and $\boldsymbol{\beta}^*$, when the form of $\gamma(\cdot)$ and the value of $\boldsymbol{\beta}^*$ are determined, the λ_{map} is fixed.

4 Tight Lower Bound in OKRidge

In this section, we utilize problem (9) to derive another novel lower bound for problem (8), serving as used as the objective function of OKRidge. Our lower bound is equivalent to the tight lower bound provided by [1]. Specifically, [1] eliminates the parameter $\boldsymbol{c}$ in problem (9) by setting the gradient of $\boldsymbol{\beta}$ to $\mathbf{0}$, while we eliminate the parameter $\boldsymbol{c}$ by setting the gradient of $\boldsymbol{c}$ to $\mathbf{0}$. These two methods

https://doi.org/10.52202/079017-2420

are equivalent due to the independence and convexity of $\boldsymbol{c}$ and $\boldsymbol{\beta}$. Given any $\boldsymbol{\beta}$ and $\boldsymbol{z}$, the optimality condition for $\boldsymbol{c}$ in problem (9) is taking $\partial\mathcal{L}_{\text{ridge}}^{\text{Fenchel}}(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{c})/\partial\boldsymbol{c}=\mathbf{0}$. Therefore, we have

$$\frac{\partial\mathcal{L}_{\text{ridge}}^{\text{Fenchel}}(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{c})}{\partial\boldsymbol{c}}=\boldsymbol{\beta}-\frac{\text{diag}(\boldsymbol{z})\boldsymbol{c}}{2}=\mathbf{0}, \tag{16}$$

$$\Rightarrow c_j=\begin{cases}\rho\in\mathbb{R} & \text{, if } z_j=0,\\ \frac{2\beta_j}{z_j} & \text{, if } z_j\neq 0,\end{cases} \tag{17}$$

where $\text{diag}(\boldsymbol{z})$ is a diagonal matrix with $\boldsymbol{z}$ on the diagonal. Inspired by this optimality condition, we present the following theorem.

Theorem 4.1. *If we define the parameter* $\mathbf{c}$ *as (16), the problem (9) is equivalent to the following optimization problem:*

$$\min_{\boldsymbol{\beta},\boldsymbol{z}}\mathcal{L}_{\textit{ridge}}^{\textit{saddle}}(\boldsymbol{\beta},\boldsymbol{z}),\quad \textit{s.t.}\ \sum_{j=1}^{d}z_j\leq k,\ z_j\in[0,1], \tag{18}$$

where

$$\mathcal{L}_{\textit{ridge}}^{\textit{saddle}}(\boldsymbol{\beta},\boldsymbol{z}):=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1,z_j\neq 0}^{d}\frac{\beta_j^2}{z_j}. \tag{19}$$

The proof of Theorem 4.1 follows Theorem 3.1 of [1] and is included in Appendix B for completeness. Following the approach by [1], we can approximately solve the problem (18) while still obtaining a feasible lower bound. We define a new function $\mathcal{L}(\boldsymbol{\beta})$ as:

$$\mathcal{L}(\boldsymbol{\beta})=\min_{\boldsymbol{z}}\mathcal{L}_{\text{ridge}}^{\text{saddle}}(\boldsymbol{\beta},\boldsymbol{z}),\quad \text{s.t.}\ \sum_{j=1}^{d}z_j\leq k,\ z_j\in[0,1]. \tag{20}$$

For any $\boldsymbol{\beta}$, $\mathcal{L}(\boldsymbol{\beta})$ serves as a valid lower bound for problem (7). We should 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 (20) is lower bounded by*

$$\mathcal{L}(\boldsymbol{\beta})\geq\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\textit{SumTop}_k(\boldsymbol{\beta}\odot\boldsymbol{\beta}). \tag{21}$$

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

The proof of Theorem 4.2 can be seen in Appendix C. Based on (9), (18) and (20), the tight lower bound (21) is equivalent to the one provided by [1], as both are derived through equivalent processes. OKRidge solves the original k-sparse problem (7) using this tight lower bound (21) as its objective function. 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

$$\min_{\boldsymbol{\beta}}\mathcal{L}_{\text{OKRidge}}(\boldsymbol{\beta}), \tag{22}$$

where we obtain $\mathcal{L}_{\text{OKRidge}}$ of formulation (3). So far, we transform the constrained k-SRO problem (2) into the unconstrained optimization problem (22). 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.

5 The Error Analysis for OKRidge

5.1 From PO to AO

As discussed in Section 4, the estimation error of OKRidge is characterized by $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$. Taking formulation (1) into the properly normalized objective (22), OKRidge (22) can be equivalently transformed to the following optimization:

$$\min_{\boldsymbol{\beta}} \frac{1}{\sqrt{n}} \left[\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \boldsymbol{\epsilon}\|_2^2 + \lambda \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta}) \right]. \tag{23}$$

The crucial step is to convert (23) into a **PO** problem within the framework of CGMT. We introduce the new variable $\boldsymbol{w} := \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and apply the Fenchel-Moreau theorem (14) to formulation (23),

$$\begin{aligned} &\frac{1}{\sqrt{n}} \left[\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{\epsilon}\|_2^2 + \lambda \text{SumTop}_k\big((\boldsymbol{w} + \boldsymbol{\beta}^*) \odot (\boldsymbol{w} + \boldsymbol{\beta}^*)\big) \right] \\ =& \max_{\boldsymbol{u}} \frac{1}{\sqrt{n}} \Big[\boldsymbol{u}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{u}^\top \boldsymbol{\epsilon} - \frac{\|\boldsymbol{u}\|_2^2}{4} + \lambda \text{SumTop}_k\big((\boldsymbol{w} + \boldsymbol{\beta}^*) \odot (\boldsymbol{w} + \boldsymbol{\beta}^*)\big) \Big], \end{aligned} \tag{24}$$

where $\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{u} \in \mathbb{R}^n$. Based on (10) and (24), the **PO** problem corresponding to the estimation error of OKRidge is

$$\Phi_{\text{OKRidge}}(\boldsymbol{X}) = \min_{\boldsymbol{w}} \max_{\boldsymbol{u}} \frac{1}{\sqrt{n}} \big(\boldsymbol{u}^\top \boldsymbol{X}\boldsymbol{w} + \psi(\boldsymbol{w}, \boldsymbol{u}) \big), \tag{25}$$

where

$$\psi(\boldsymbol{w}, \boldsymbol{u}) := -\boldsymbol{u}^\top \boldsymbol{\epsilon} - \frac{\|\boldsymbol{u}\|_2^2}{4} + \lambda \text{SumTop}_k\big((\boldsymbol{w} + \boldsymbol{\beta}^*) \odot (\boldsymbol{w} + \boldsymbol{\beta}^*)\big). \tag{26}$$

Since the entries of $\boldsymbol{X}$ are drawn i.i.d. from $\mathcal{N}(0,1)$, to replace the challenging **PO** problem (25) with a simplified **AO** problem through CGMT, $\psi(\boldsymbol{w}, \boldsymbol{u})$ should be a convex-concave function. The following Lemma illustrates that the $\psi(\boldsymbol{w}, \boldsymbol{u})$ satisfies the conditions of Theorem 3.2.

Lemma 5.1. *Suppose $\psi(\boldsymbol{w}, \boldsymbol{u})$ is defined as in formulation (26). Then, $\psi(\boldsymbol{w}, \boldsymbol{u})$ is convex-concave function.*

The proof of Lemma 5.1 can be seen in Appendix D. Define

$$\gamma(\boldsymbol{\beta}) := \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta}).$$

Because the **PO** problem (25) satisfies the assumptions of CGMT, we transform it to the following **AO** problem:

$$\begin{aligned} \phi_{\text{OKRidge}}(\boldsymbol{g}, \boldsymbol{h}) &= \min_{\boldsymbol{w}} \max_{\boldsymbol{u}} \frac{1}{\sqrt{n}} \Big[\|\boldsymbol{w}\|_2 \boldsymbol{g}^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \boldsymbol{h}^\top \boldsymbol{w} - \boldsymbol{\epsilon}^\top \boldsymbol{u} - \frac{\|\boldsymbol{u}\|_2^2}{4} + \lambda \gamma(\boldsymbol{\beta}^* + \boldsymbol{w}) \Big] \\ &= \min_{\boldsymbol{w}} \max_{\boldsymbol{u}} \frac{1}{\sqrt{n}} \Big[(\|\boldsymbol{w}\|_2 \boldsymbol{g} - \boldsymbol{\epsilon})^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \boldsymbol{h}^\top \boldsymbol{w} - \frac{\|\boldsymbol{u}\|_2^2}{4} + \lambda \gamma(\boldsymbol{\beta}^* + \boldsymbol{w}) \Big], \end{aligned} \tag{27}$$

where the entries $\boldsymbol{g}$, $\boldsymbol{h}$ are drawn i.i.d. from $\mathcal{N}(0,1)$, due to the property of $\boldsymbol{X}$. Suppose $\boldsymbol{w}_{\Phi_{\text{OKRidge}}}$ is the of optimal solutions of the **PO** problem (25), and $\boldsymbol{w}_{\phi_{\text{OKRidge}}}$ is the optimal solutions of the **AO** problem (27). According to Theorem 3.2, if $\|\boldsymbol{w}_{\phi_{\text{OKRidge}}}\|_2 \xrightarrow{P} \alpha^*$, we have $\|\boldsymbol{w}_{\Phi_{\text{OKRidge}}}\|_2 \xrightarrow{P} \alpha^*$. Thus, we can analyze the minimizer of **AO** problem (27) instead of **PO** problem (25).

5.2 Simplification for AO

In this chapter, we simplify the **AO** problem (27) into ones involving only scalar quantities. Since $\gamma(\boldsymbol{\beta})$ is a convex (see Lemma 5.1), $\partial\gamma(\boldsymbol{\beta}^*)$ is nonempty, convex and compact. According to [13, Theorem 23.4], we have $\gamma(\boldsymbol{\beta}^* + \boldsymbol{w}) = \gamma(\boldsymbol{\beta}^*) + \max_{\boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)} \boldsymbol{s}^\top \boldsymbol{w} + O(\|\boldsymbol{w}\|_2^2)$. The first-order approximation of $\gamma(\boldsymbol{\beta})$ around the vector of interest $\boldsymbol{\beta}^*$ is

$$\hat{\gamma}(\boldsymbol{\beta}^* + \boldsymbol{w}) := \gamma(\boldsymbol{\beta}^*) + \max_{\boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)} \boldsymbol{s}^\top \boldsymbol{w}. \tag{28}$$

https://doi.org/10.52202/079017-2420

where $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{w}$. Then, following the approach from [12], the **AO** problem (27) can be simplified by the first-order approximation (28):

$$\begin{aligned}\hat{\phi}_{\text{OKRidge}}(\boldsymbol{g}, \boldsymbol{h}) &= \min_{\boldsymbol{w}} \max_{\boldsymbol{u}} \frac{1}{\sqrt{n}} \Big[(\|\boldsymbol{w}\|_2 \boldsymbol{g} - \boldsymbol{\epsilon})^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \boldsymbol{h}^\top \boldsymbol{w} + \lambda\big(\gamma(\boldsymbol{\beta}^*) + \max_{\boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)} \boldsymbol{s}^\top \boldsymbol{w}\big) - \frac{\|\boldsymbol{u}\|_2^2}{4} \Big] \\ &= \min_{\boldsymbol{w}} \max_{\substack{\|\boldsymbol{u}\|_2 \geq 0 \\ \boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)}} \frac{1}{\sqrt{n}} \Big[(\|\boldsymbol{w}\|_2 \boldsymbol{g} - \boldsymbol{\epsilon})^\top \boldsymbol{u} + (\|\boldsymbol{u}\|_2 \boldsymbol{h} + \lambda \boldsymbol{s})^\top \boldsymbol{w} + \lambda\gamma(\boldsymbol{\beta}^*) - \frac{\|\boldsymbol{u}\|_2^2}{4} \Big]. \end{aligned} \tag{29}$$

Suppose $f(\boldsymbol{\beta})$ and $\hat{f}(\boldsymbol{\beta})$ denote the objective functions of the original and the approximated **AO** problems (27) and (29), respectively,

$$\begin{aligned} f(\boldsymbol{\beta}) &= (\|\boldsymbol{w}\|_2 \boldsymbol{g} - \boldsymbol{\epsilon})^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \boldsymbol{h}^\top \boldsymbol{w} - \frac{\|\boldsymbol{u}\|_2^2}{4} + \lambda\gamma(\boldsymbol{\beta}^* + \boldsymbol{w}), \\ \hat{f}(\boldsymbol{\beta}) &= (\|\boldsymbol{w}\|_2 \boldsymbol{g} - \boldsymbol{\epsilon})^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \boldsymbol{h}^\top \boldsymbol{w} - \frac{\|\boldsymbol{u}\|_2^2}{4} + \lambda\big(\gamma(\boldsymbol{\beta}^*) + \max_{\boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)} \boldsymbol{s}^\top \boldsymbol{w}\big). \end{aligned}$$

Then, based on (28), we have

$$\lim_{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \to 0} \hat{f}(\boldsymbol{\beta}) = f(\boldsymbol{\beta}). \tag{30}$$

Compared with **AO** problem (27), the approximated **AO** problem (29) is tight when $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \to 0$, and we later demonstrate that this condition is satisfied as $\sigma^2 \to 0$, independent of the original **AO** problem (27). This fact allows us to translate the analysis on the optimal solution $\boldsymbol{w}_{\hat{\phi}_{\text{OKRidge}}}$ of the approximated **AO** problem (29) to the analysis on the optimal solution $\boldsymbol{w}_{\phi_{\text{OKRidge}}}$ of the corresponding original **AO** problem (27). Because $\gamma(\boldsymbol{\beta}^*)$ is a constant, the approximated **AO** problem (29) is equivalent to the following optimization problem:

$$\min_{\boldsymbol{w}} \max_{\substack{\|\boldsymbol{u}\|_2 \geq 0 \\ \boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)}} \frac{1}{\sqrt{n}} \Big[(\|\boldsymbol{w}\|_2 \boldsymbol{g} - \boldsymbol{\epsilon})^\top \boldsymbol{u} + (\|\boldsymbol{u}\|_2 \boldsymbol{h} + \lambda \boldsymbol{s})^\top \boldsymbol{w} - \frac{\|\boldsymbol{u}\|_2^2}{4} \Big], \tag{31}$$

where we have approximated γ in the first order. Since $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$, the term $\|\boldsymbol{w}\|_2 \boldsymbol{g} - \boldsymbol{\epsilon}$ above is statistically identical to a random vector with entries drawn i.i.d. from $\mathcal{N}(0, \|\boldsymbol{w}\|_2^2 + \sigma^2)$, where $\boldsymbol{I}$ is the unit matrix. Following the method used by [42], we substitute the first term in the objective (31) with $\sqrt{\|\boldsymbol{w}\|_2^2 + \sigma^2}\boldsymbol{g}^\top \boldsymbol{u}$. Then, we obtain:

$$\min_{\boldsymbol{w}} \max_{\substack{\|\boldsymbol{u}\|_2 \geq 0 \\ \boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)}} \frac{1}{\sqrt{n}} \Big[\sqrt{\|\boldsymbol{w}\|_2^2 + \sigma^2} \cdot \boldsymbol{g}^\top \boldsymbol{u} + (\|\boldsymbol{u}\|_2 \boldsymbol{h} + \lambda \boldsymbol{s})^\top \boldsymbol{w} - \frac{\|\boldsymbol{u}\|_2^2}{4} \Big]. \tag{32}$$

Let $\eta = \|\boldsymbol{u}\|_2$. Since $\max_{\|\boldsymbol{u}\|_2 = \eta} \boldsymbol{g}^\top \boldsymbol{u} = \|\boldsymbol{g}\|_2 \cdot \|\boldsymbol{u}\|_2$ and $\boldsymbol{h} \sim \mathcal{N}(0, \boldsymbol{I})$, in term of mathematical expectation, the optimization (32) can be equivalently expressed as:

$$\min_{\boldsymbol{w}} \max_{\substack{\eta \geq 0 \\ \boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)}} \frac{1}{\sqrt{n}} \Big[\sqrt{\|\boldsymbol{w}\|_2^2 + \sigma^2} \|\boldsymbol{g}\|_2 \eta - \eta\big(\boldsymbol{h} - \frac{\lambda}{\eta}\boldsymbol{s}\big)^\top \boldsymbol{w} - \frac{\eta^2}{4} \Big]. \tag{33}$$

The objective (33) is strongly convex in $\boldsymbol{w}$ and (jointly) concave in η, $\boldsymbol{s}$, and the constraint sets are bounded. Therefore, we can reverse the order of min-max in problem (33) based on [13, Corollary 37.3.2]. Let $\alpha = \|\boldsymbol{w}\|_2$. Since $\min_{\|\boldsymbol{w}\|_2 = \alpha} (-\boldsymbol{h} + \frac{\lambda}{\eta}\boldsymbol{s})^\top \boldsymbol{w} = -\alpha \|\boldsymbol{h} - \frac{\lambda}{\eta}\boldsymbol{s}\|_2$, the optimization (33) can be equivalently reformulated as:

$$\max_{\substack{\eta \geq 0 \\ \boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)}} \min_{\alpha \geq 0} \frac{1}{\sqrt{n}} \Big(\sqrt{\alpha^2 + \sigma^2} \cdot \|\boldsymbol{g}\|_2 \eta - \alpha\eta \|\boldsymbol{h} - \frac{\lambda}{\eta}\boldsymbol{s}\|_2 - \frac{\eta^2}{4} \Big). \tag{34}$$

Next, we further reverse the order of min-max, as the objective (34) exhibits the desired concave-convex structure. Then, we proceed to maximize over $\boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)$. Since $\min_{\boldsymbol{s} \in \partial\gamma(\boldsymbol{\beta}^*)} \|\boldsymbol{h} - \frac{\lambda}{\eta}\boldsymbol{s}\|_2 = \text{dist}\big(\boldsymbol{h}, \frac{\lambda}{\eta}\partial\gamma(\boldsymbol{\beta}^*)\big)$, the optimization problem (34) can alternatively be formulated as:

$$\min_{\alpha \geq 0} \max_{\eta \geq 0} \frac{1}{\sqrt{n}} \Big(\sqrt{\alpha^2 + \sigma^2} \cdot \|\boldsymbol{g}\|_2 \eta - \alpha\eta \cdot \text{dist}\big(\boldsymbol{h}, \frac{\lambda}{\eta}\boldsymbol{s}\big) - \frac{\eta^2}{4} \Big). \tag{35}$$

Because both the random components $\|\boldsymbol{g}\|_2$ and dist$(\boldsymbol{h}, \frac{\lambda}{\eta}\boldsymbol{s})$ are Lipschitz, $\|\boldsymbol{g}\|_2$ concentrates around $\sqrt{n}$ and dist$(\boldsymbol{h}, \frac{\lambda}{\eta}\boldsymbol{s})$ around $\sqrt{D(\frac{\lambda}{\eta})}$ [43, Lemma B.2]. Suppose, as $d \to \infty$, $\frac{D(\tau)}{n} \to \bar{D}(\tau) \in (0,1)$, $\frac{C(\tau)}{n} \to \bar{C}(\tau)$, and $\Gamma(\eta) := \lim_{n\to\infty} \frac{\eta^2}{4\sqrt{n}}$. Then, the optimal minimizer of (35) converges to the optimal minimizer of the following deterministic optimization in probability [44]:

$$\max_{\eta \geq 0} \min_{\alpha \geq 0} \eta\sqrt{\alpha^2 + \sigma^2} - \alpha\eta\sqrt{\bar{D}(\frac{\lambda}{\eta})} - \Gamma(\eta). \tag{36}$$

Here, we complete the simplifications by reducing the **AO** problem (27) to an equivalent optimization (36) that now only involves two scalar variables: α and η.

5.3 Error Analysis

Based on the analysis above, if the optimal solution of optimization (36) is $\alpha = \alpha^*$, we have $\|\boldsymbol{w}_{\hat{\phi}_{\text{OKRidge}}}\|_2 \xrightarrow{P} \alpha^*$ for approximated **AO** problem (29). If α^* further tends to 0, according to formulation (30) and CGMT, $\|\boldsymbol{w}_{\Phi_{\text{OKRidge}}}\|_2 \xrightarrow{P} \alpha^*$ holds for **PO** problem (25). Then, for the estimation error of OKRidge produced by (22), we have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \xrightarrow{P} \alpha^*$. Therefore, it only remains to obtain the optimal value of α in optimization (36) that plays the role of $\|\boldsymbol{w}\|_2$. Following [45], we conclude the estimation error of OKRidge with Theorem 5.2 below.

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 (22) 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:*

$$\lim_{d\to 0} aNSE \xrightarrow{P} \Delta(\hat{\lambda}), \tag{37}$$

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

The proof of Theorem 5.2 can be seen in Appendix E.

Remark 5.3. In the objective (24) concerning estimation error of OKRidge, $\gamma(\boldsymbol{\beta}) = \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta})$ and the value of $\boldsymbol{\beta}^*$ is assumed to be known. Then, the analysis on $map(\cdot)$ in Section 3.3 reveals that the form of $map(\cdot)$ and the value of λ_{map} are fixed. Thus, $\Delta(\hat{\lambda})$ is a function of λ. In other words, if the regularizer parameter λ of OKRidge is fixed, the NSE limit of OKRidge $\Delta(\hat{\lambda})$ is also fixed. Additionally, Theorem 5.2 also indicates that $\lim_{d\to\infty} \lim_{\sigma\to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \xrightarrow{P} 0$, which guarantees the effectiveness of $\hat{\boldsymbol{\beta}}$ learned by OKRidge in accurately estimating $\boldsymbol{\beta}^*$. These results substantiate the theoretical reliability of OKRidge and promote its broad application in the real world.

6 Numerical Experiments

In this section, we conduct experiments to verify Theorem 5.2. The experiments contain two aspects: (i) When λ is fixed, NSE tends to a fixed constant as $\sigma \to 0$. (ii) When $\sigma \to 0$, NSE is determined by the weight λ of the regularizer. In other words, NSE is a function of λ.

In our experiments, $\boldsymbol{\beta}^*$ is randomly generated with $\|\boldsymbol{\beta}^*\|_0 \leq k$. For $i \in \{1, 2, \cdots, n\}$, $\boldsymbol{x}_i$ is drawn i.i.d. from $\mathcal{N}(0, \boldsymbol{I})$, and ϵ_i is drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$. According to the k-sparse linear regression (1), $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i$, we get dataset $(\boldsymbol{x}_i, y_i)$ with $i = 1, 2, \cdots, n$. Then, we appy OKRidge to get the estimator $\hat{\boldsymbol{\beta}}$ and calculate the NSE by $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2/\sigma^2$. The NSE is averaged over 10 trials to evaluate the effectiveness of the OKRidge algorithm. In the main paper, we set $\frac{n}{d} = 0.5$, $\frac{k}{d} = 0.1$, $d = 100$. The computer resources are detailed in Appendix F.1. More experiments with various settings about $\frac{n}{d}$ and $\frac{k}{d}$ can be seen in Appendix F.2.

https://doi.org/10.52202/079017-2420

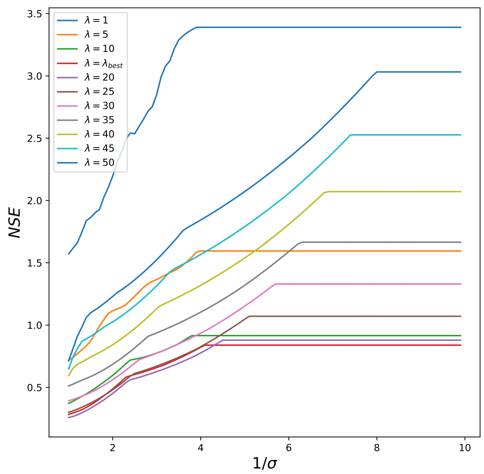


Figure 1: The change of NSE with $1/\sigma$ for OKRidge under different λ. The red curve at the bottom corresponds to the case $\lambda = \lambda_{best}$.

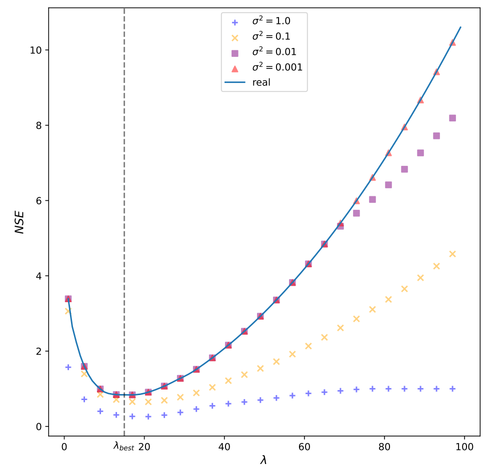


Figure 2: The change of NSE with λ for OKRidge under different σ. The blue curve corresponds to the real change of $\Delta(\hat{\lambda})$. Here, λ_{best} is the optimal weight of the regularizer.

6.1 The Change of NSE with σ

We investigate the change of NSE with $1/\sigma$ under $\lambda = 1, 5, 10, \lambda_{best}, \cdots$. The results are illustrated in Figure 1. As depicted in Figure 1, when λ is constant, NSE converges to a fixed value as $\sigma \to 0$. This observation validates aspect (i) of Theorem 5.2.

6.2 The Change of NSE with λ

We analyze the change of NSE with λ under $\sigma^2 = 1, 0.1, 0.01, 0.001$. The outcomes are depicted in Figure 2. As shown in Figure 2, the curves converge towards the real blue curve as $\sigma \to 0$, where the blue curve relies on λ. This observation confirms aspect (ii) of Theorem 5.2.

7 Conclusion

In this paper, we present a theoretical high-dimensional error analysis of the OKRidge algorithm in idealized settings using the CGMT framework. Specifically, when OKRidge tackles a k-sparse linear model with $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I})$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $\lim_{d\to\infty} \frac{n}{d} = \delta \in (0, 1)$, we have

$$\lim_{d\to\infty} \lim_{\sigma^2\to 0} \frac{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2}{\sigma^2} \xrightarrow{P} \Delta(\hat{\lambda}), \text{ and } \lim_{d\to\infty} \lim_{\sigma\to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \xrightarrow{P} 0.$$

where $\Delta(\hat{\lambda})$ depends on λ. This indicates that (i) the NSE limit of OKRidge remains constant when λ is fixed; (ii) $\hat{\boldsymbol{\beta}}$ learned by OKRidge is effective in estimating $\boldsymbol{\beta}^*$. Our experimental findings support these theoretical assertions. This theoretical error analysis substantiates the reliability of OKRidge and provides guidelines on the error analysis of other algorithms.

Acknowledgments and Disclosure of Funding

This work is supported by the Key R&D Program of Hubei Province under Grant 2024BAB038, National Key R&D Program of China under Grant 2023YFC3604702, and the Fundamental Research Fund Program of LIESMARS.

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] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, page 2313–2351, 2007.

[3] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] Xiyuan Li, Xin Zou, and Weiwei Liu. Residual network with self-adaptive time step size. *Pattern Recognition*, 158:111008, 2025.

[5] Xiyuan Li, Zou Xin, and Weiwei Liu. Defending against adversarial attacks via neural dynamic system. In *NeurIPS*, 2022.

[6] Weijun Xie and Xinwei Deng. Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization*, 30(4):3359–3386, 2020.

[7] Mert Pilanci, Martin J. Wainwright, and Laurent El Ghaoui. Sparse learning via boolean relaxations. *Mathematical Programming*, 151(1):63–87, 2015.

[8] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Scientific Computing*, 24(2):227–234, 1995.

[9] Kathleen P. Champion, Peng Zheng, Aleksandr Y. Aravkin, Steven L. Brunton, and J. Nathan Kutz. A unified sparse optimization framework to learn parsimonious physics-informed models from data. *IEEE Access*, 8:169259–169271, 2020.

[10] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2009.

[12] 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.

[13] R. Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1997.

[14] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[15] Urban Fasel, J Nathan Kutz, Bingni W Brunton, and Steven L Brunton. Ensemble-sindy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A*, 478(2260):20210904, 2022.

[16] Amir Beck and Yonina C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[17] T. Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

[18] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

[19] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.

[20] Dimitris Bertsimas and Wes Gurnee. Learning sparse nonlinear dynamics via mixed-integer optimization. *Nonlinear Dynamics*, 111(7):6585–6604, 2023.

[21] Weiwei Liu and Ivor W. Tsang. Making decision trees feasible in ultrahigh feature and label dimensions. *Journal Of Machine Learning Research*, 18:81:1–81:36, 2017.

https://doi.org/10.52202/079017-2420

[22] Dimitris Bertsimas, Jean Pauphilet, and Bart Van Parys. Sparse regression: Scalable algorithms and empirical performance. *Statistical Science*, 35(4):555–578, 2020.

[23] Alper Atamtürk, Andrés Gómez, and Shaoning Han. Sparse and smooth signal estimation: Convexification of l0-formulations. *Journal of Machine Learning Research*, 22:52:1–52:43, 2021.

[24] Oktay Günlük and Jeff Linderoth. Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Mathematical programming*, 124(1):183–205, 2010.

[25] Chenglin Yu, Xinsong Ma, and Weiwei Liu. Delving into noisy label detection with clean data. In *ICML*, volume 202, pages 40290–40305.

[26] Lianghe Shi and Weiwei Liu. Adversarial self-training improves robustness and generalization for gradual domain adaptation. In *NeurIPS*, 2023.

[27] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):pp. 1705–1732, 2009.

[28] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *COLT*, pages 1683–1709, 2015.

[29] David L. Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

[30] Arian Maleki, Laura Anitori, Zai Yang, and Richard G. Baraniuk. Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Transactions on Information Theory*, 59(7):4290–4308, 2013.

[31] Christopher A. Metzler, Arian Maleki, and Richard G. Baraniuk. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016.

[32] Ryuhei Miyashiro and Yuichi Takano. Subset selection by mallows' c_p: A mixed integer programming approach. *Expert Systems with Applications*, 42(1):325–331, 2015.

[33] Antonio Frangioni and Claudio Gentile. Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106(2):225–236, 2006.

[34] Patrick L Combettes. Perspective functions: Properties, constructions, and examples. *Set-Valued and Variational Analysis*, 26(2):247–264, 2018.

[35] Alper Atamturk and Andrés Gómez. Safe screening rules for l0-regression from perspective relaxations. In *ICML*, pages 421–430, 2020.

[36] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *NeurIPS*, volume 32, 2019.

[37] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The performance analysis of generalized margin maximizers on separable data. In *ICML*, pages 8417–8426, 2020.

[38] Yanbo Chen and Weiwei Liu. A theory of transfer-based black-box attacks: Explanation and implications. In *NeurIPS*, 2023.

[39] Jingyuan Xu and Weiwei Liu. Characterization of overfitting in robust multiclass classification. In *NeurIPS*, 2023.

[40] Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*, 1988.

[41] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

[42] Christos Thrampoulidis, Ashkan Panahi, and Babak Hassibi. Asymptotically exact error analysis for the generalized equation-lasso. In *ISIT*, pages 2021–2025, 2015.

[43] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized LASSO: A precise analysis. In *Allerton*, pages 1002–1009, 2013.

[44] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

[45] Christos Thrampoulidis, Ashkan Panahi, Daniel Guo, and Babak Hassibi. Precise error analysis of the LASSO. In *ICASSP*, pages 3467–3471, 2015.

https://doi.org/10.52202/079017-2420

The Reliability of OKRidge (Appendix)

A Limitaion and Impact

Limitaion: Our results rely on Gaussian input settings. For non-Gaussian settings, we can utilize Fisher transformation, Box-Cox transformation, or inversion sampling to transform non-Gaussian distribution to Gaussian distribution. In our future work, we will discuss the potential extensions of our findings to non-Gaussian input settings, providing insights into the universality of the results.

Impact: Our work provides theoretical support for the broad application of OKRidge, which does not require proprietary software or expensive licenses [1], unlike its main competitor. This can significantly impact various regression applications.

B Proof of Theorem 4.1

Theorem 4.1. If we define the parameter $\boldsymbol{c}$ as (16), the problem (9) is equivalent to the following saddle point optimization problem:

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \mathcal{L}_{\text{ridge}}^{\text{saddle}}(\boldsymbol{\beta},\boldsymbol{z}) \quad \text{s.t.} \quad \sum_{j=1}^{d} z_j \le k, z_j \in [0,1], \tag{38}$$

where

$$\mathcal{L}_{\text{ridge}}^{\text{saddle}}(\boldsymbol{\beta},\boldsymbol{z}) := \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1,z_j\neq 0}^{d} \frac{\beta_j^2}{z_j}. \tag{39}$$

Proof. The problem (9) is

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

where

$$\mathcal{L}_{\text{ridge}}^{\text{Fenchel}}(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{c}) := \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{d} (\beta_j c_j - \frac{c_j^2}{4} z_j).$$

The parameter $\boldsymbol{c}$ in (16) is

$$c_j = \begin{cases} \rho \in \mathbb{R} & ,\text{if} \quad z_j = 0, \\ \frac{2\beta_j}{z_j} & ,\text{if} \quad z_j \neq 0. \end{cases}$$

Substitute parameter $\boldsymbol{c}$ of (16) into (9),

$$\begin{aligned} \mathcal{L}_{\text{ridge}}^{\text{Fenchel}}(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{c}) &= \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{d} (\beta_j c_j - \frac{c_j^2}{4} z_j) \\ &= \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1,z_j\neq 0}^{d} \left[\beta_j \cdot \frac{2\beta_j}{z_j} - (\frac{2\beta_j}{z_j})^2/4 \cdot z_j\right] \\ &= \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1,z_j\neq 0}^{d} \left[\frac{2\beta_j^2}{z_j} - \frac{\beta_j^2}{z_j}\right] \\ &= \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1,z_j\neq 0}^{d} \frac{2\beta_j^2}{z_j} \end{aligned}$$

Define

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

Then, the problem (9) can be equivalently written as

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

which is our saddle point optimization (38). □

C Proof of Theorem 4.2

Theorem 4.2. The function $\mathcal{L}(\boldsymbol{\beta})$ defined in Equation (20) is lower bounded by

$$\mathcal{L}(\boldsymbol{\beta}) \geq \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta}). \tag{40}$$

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

Proof. The Equation (20) is

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

According to (20) and (39),

$$\mathcal{L}(\boldsymbol{\beta}) = \min_{\boldsymbol{z}} \mathcal{L}_{\text{ridge}}^{\text{saddle}}(\boldsymbol{\beta}, \boldsymbol{z}) = \min_{\boldsymbol{z}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1, z_j \neq 0}^{d} \frac{\beta_j^2}{z_j} \tag{41}$$

Following the method of [1], if we minimize z in the optimization (41) under the constraints $\sum_{j=1}^{p} z_j \leq k$ and $z_j \in [0, 1]$ for $\forall j$, we have $z_j = 1$ for the top k terms of β_j^2 and $z_j = 0$ otherwise. Then, we have

$$\mathcal{L}(\boldsymbol{\beta}) \geq \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1, z_j \neq 0}^{d} \frac{\beta_j^2}{1} = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \text{SumTop}_k(\boldsymbol{\beta} \odot \boldsymbol{\beta}).$$

□

D Proof of Lemma 5.1

Lemma 5.1. Suppose $\psi(\boldsymbol{w}, \boldsymbol{u})$ is defined as formulation (26). Then, $\psi(\boldsymbol{w}, \boldsymbol{u})$ is convex-concave on $\mathbb{R}^d \times \mathbb{R}^n$.

Proof. Obviously, $\psi(\boldsymbol{w}, \boldsymbol{u})$ is concave about $\boldsymbol{u}$. Next, we indicate that $\psi(\boldsymbol{w}, \boldsymbol{u})$ is convex about $\boldsymbol{w}$. For any $\bar{\boldsymbol{w}}, \widetilde{\boldsymbol{w}} \in \mathbb{R}^d$ and $\forall \theta \in (0, 1)$,

$$[\theta \bar{w}_j + (1-\theta)\widetilde{w}_j]^2 \leq \theta \bar{w}_j^2 + (1-\theta)\widetilde{w}_j^2.$$

Then, we have

$$\begin{aligned}
&\text{SumTop}_k\big([\theta\bar{\boldsymbol{w}} + (1-\theta)\widetilde{\boldsymbol{w}}] \odot [\theta\bar{\boldsymbol{w}} + (1-\theta)\widetilde{\boldsymbol{w}}]\big) \\
\leq &\text{SumTop}_k\big(\theta\bar{\boldsymbol{w}} \odot \bar{\boldsymbol{w}} + (1-\theta)\widetilde{\boldsymbol{w}} \odot \widetilde{\boldsymbol{w}}\big) \\
\leq &\text{SumTop}_k\big(\theta\bar{\boldsymbol{w}} \odot \bar{\boldsymbol{w}}\big) + \text{SumTop}_k\big((1-\theta)\widetilde{\boldsymbol{w}} \odot \widetilde{\boldsymbol{w}}\big) \\
\leq &\theta\text{SumTop}_k\big(\bar{\boldsymbol{w}} \odot \bar{\boldsymbol{w}}\big) + (1-\theta)\text{SumTop}_k\big(\widetilde{\boldsymbol{w}} \odot \widetilde{\boldsymbol{w}}\big).
\end{aligned}$$

$\text{SumTop}_k(\cdot)$ is a convex operator. Therefore, $\psi(\boldsymbol{w}, \boldsymbol{u})$ is concave about $\boldsymbol{u}$ and convex about $\boldsymbol{w}$.

□

https://doi.org/10.52202/079017-2420

E Proof of Theorem 5.2

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 (22) of OKRidge, $\frac{D(\tau)}{n} \to \bar{D}(\tau) \in (0,1)$,

$$\text{aNSE} := \lim_{\sigma^2 \to 0} \text{NSE} = \lim_{\sigma^2 \to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 / \sigma^2.$$

Define λ_{map} is the solution of $\text{map}(\tau) = 0$ for $\tau > 0$, then, the estimation error of OKRidge is given by the following probability limit:

$$\lim_{d \to 0} \text{aNSE} \xrightarrow{P} \Delta(\hat{\lambda}), \tag{42}$$

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

Proof. Starting from the simplified **AO** problem (36), let's define

$$\kappa(\alpha, \eta) := \eta\sqrt{\alpha^2 + \sigma^2} - \alpha\eta\sqrt{\bar{D}(\frac{\lambda}{\eta})} - \Gamma(\eta). \tag{43}$$

Given that $\kappa(\alpha, \eta)$ is strongly convex in α and concave in η, we denote (α^*, η^*) as the Nash equilibrium of $\kappa(\alpha, \eta)$. Then, α^* is unique and we use duality to compute α^*. For any η,

$$\frac{\partial \kappa(\alpha, \eta)}{\partial \alpha} = \frac{\alpha\eta}{\sqrt{\alpha^2 + \sigma^2}} - \eta\sqrt{\bar{D}(\frac{\lambda}{\eta})} = 0, \Rightarrow \alpha^*(\eta) = \sigma\sqrt{\frac{\bar{D}(\frac{\lambda}{\eta})}{1 - \bar{D}(\frac{\lambda}{\eta})}}, \tag{44}$$

which is well-defined, due to $1 - \bar{D}(\frac{\lambda}{\eta}) > 0$. Substituting $\alpha^*(\eta)$ back into (43) gives:

$$\begin{aligned}
\kappa(\alpha^*, \eta) &= \eta\sqrt{\alpha^{*2} + \sigma^2} - \alpha^*\eta\sqrt{\bar{D}(\frac{\lambda}{\eta})} - \Gamma(\eta) \\
&= \eta\sqrt{\sigma^2 \frac{\bar{D}(\frac{\lambda}{\eta})}{1 - \bar{D}(\frac{\lambda}{\eta})} + \sigma^2} - \sigma\eta\sqrt{\frac{\bar{D}(\frac{\lambda}{\eta})}{1 - \bar{D}(\frac{\lambda}{\eta})}} \cdot \sqrt{\bar{D}(\frac{\lambda}{\eta})} - \Gamma(\eta) \\
&= \frac{\eta\sigma}{\sqrt{1 - \bar{D}(\frac{\lambda}{\eta})}} - \frac{\eta\sigma\bar{D}(\frac{\lambda}{\eta})}{\sqrt{1 - \bar{D}(\frac{\lambda}{\eta})}} - \Gamma(\eta) = \eta\sigma\sqrt{1 - \bar{D}(\frac{\lambda}{\eta})} - \Gamma(\eta).
\end{aligned}$$

Differentiating $\kappa(\alpha^*, \eta)$ with respect to η, we have

$$\begin{aligned}
\frac{\partial \kappa(\alpha^*, \eta)}{\partial \eta} &= \sigma\sqrt{1 - \bar{D}(\frac{\lambda}{\eta})} + \frac{1}{2}\eta\sigma\frac{C(\frac{\lambda}{\eta})}{\sqrt{1 - \bar{D}(\frac{\lambda}{\eta})}} \cdot \frac{\lambda}{\eta^2} - \Gamma'(\eta) = \sigma\frac{1 - \bar{C}(\frac{\lambda}{\eta}) - \bar{D}(\frac{\lambda}{\eta})}{\sqrt{1 - \bar{D}(\frac{\lambda}{\eta})}} - \Gamma'(\eta) \\
&= \sigma\text{map}(\frac{\lambda}{\eta}) - \lim_{n \to \infty} \frac{\eta}{2\sqrt{n}}.
\end{aligned} \tag{45}$$

Here, $\text{map}(\tau) = 0$ when $\tau = \lambda_{map}$. Hence, $\eta^* = \frac{\lambda}{\lambda_{map}}$ is the optimal solution, due to

$$\frac{\partial \kappa(\alpha^*, \eta)}{\partial \eta}\Big|_{\eta = \eta^*} = \sigma\text{map}(\lambda_{map}) - \lim_{n \to \infty} \frac{\lambda}{2\lambda_{map}\sqrt{n}} = 0.$$

Because the form of $\gamma(\cdot)$ and the value of $\boldsymbol{\beta}^*$ are determined, the form of $map(\cdot)$ and the value of λ_{map} are fixed. In other words, λ_{map} is a function about λ. Therefore, we can take $\eta^* = \frac{\lambda}{\lambda_{map}}$ to formulation (44),

$$\alpha^*(\eta^*) = \sigma\sqrt{\frac{\bar{D}(\frac{\lambda}{\eta^*})}{1 - \bar{D}(\frac{\lambda}{\eta^*})}} = \sigma\sqrt{\frac{\bar{D}(\lambda_{map})}{1 - \bar{D}(\lambda_{map})}}.$$

Moreover $\kappa(\alpha^*, \eta^*) = \frac{\lambda}{\lambda_{map}} \sigma \sqrt{1 - \bar{D}(\lambda_{map})}$. Denote $\hat{\lambda} = \lambda_{map}$,

$$\alpha^* = \sigma \sqrt{\frac{\bar{D}(\hat{\lambda})}{1 - \bar{D}(\hat{\lambda})}}. \tag{46}$$

Based on the analysis above, if the optimal solution of optimization (36) is $\alpha = \alpha^*$, we have $\|\boldsymbol{w}_{\hat{\phi}_{\text{OKRidge}}}\|_2 \xrightarrow{P} \alpha^*$ for approximated **AO** problem (29). Furthermore, according to formulation (46), $\lim_{\sigma \to 0} \alpha^* = 0$ occurs for the approximated **AO** problem (29) and is independent of the original **AO** problem (27). Thus, formulation (30) holds:

$$\lim_{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \to 0} \hat{f}(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) \Leftrightarrow \lim_{d \to \infty} \lim_{\sigma \to 0} \hat{f}(\boldsymbol{\beta}) = f(\boldsymbol{\beta}). \tag{47}$$

In the case $n \to \infty$ and $\sigma \to 0$, formulation (47) allows us to translate the optimal error α^* of the approximated **AO** problem (29) to the optimal error of the original **AO** problem (27). Combining formulations (46), (47), and CGMT, $\|\boldsymbol{w}_{\Phi_{\text{OKRidge}}}\|_2 \xrightarrow{P} \alpha^*$ holds for **PO** problem (25). Then, for the estimation error of OKRidge produced by (22), we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \xrightarrow{P} \alpha^*.$$

Therefore, combining **PO** problem (25), **AO** problem (27), the relation (47), and the formulations (23) and (46), according to CGMT, we obtain::

$$\lim_{d \to \infty} \lim_{\sigma \to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 / \sigma \xrightarrow{P} \sqrt{\frac{\bar{D}(\hat{\lambda})}{1 - \bar{D}(\hat{\lambda})}} \Rightarrow \lim_{d \to \infty} \lim_{\sigma \to 0} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 / \sigma^2 \xrightarrow{P} \frac{\bar{D}(\hat{\lambda})}{1 - \bar{D}(\hat{\lambda})}.$$

To sum up,

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

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

□

F Experiments Appendix

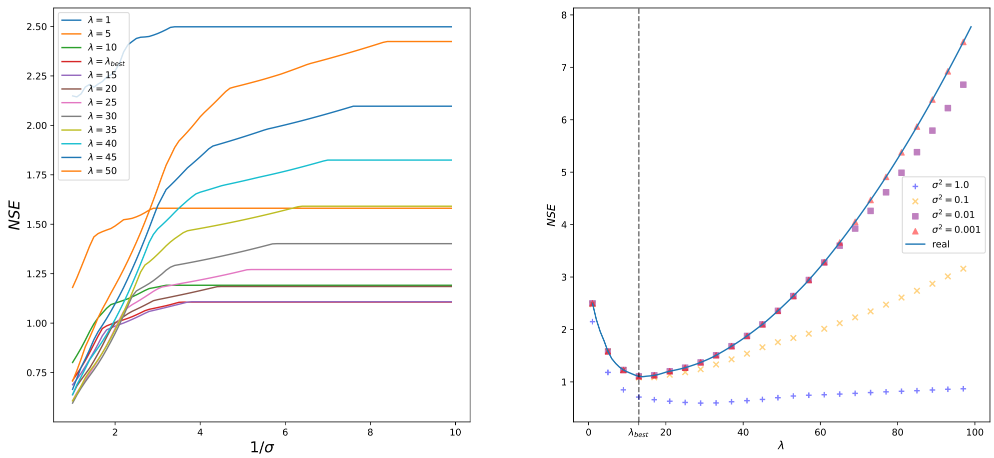
F.1 Computing Platform

All experiments were run on the 10x TensorEX TS2-673917-DPN Intel Xeon Gold 6226 Processor, 2.7Ghz. We set the memory limit to be 100GB.

F.2 More Experiments

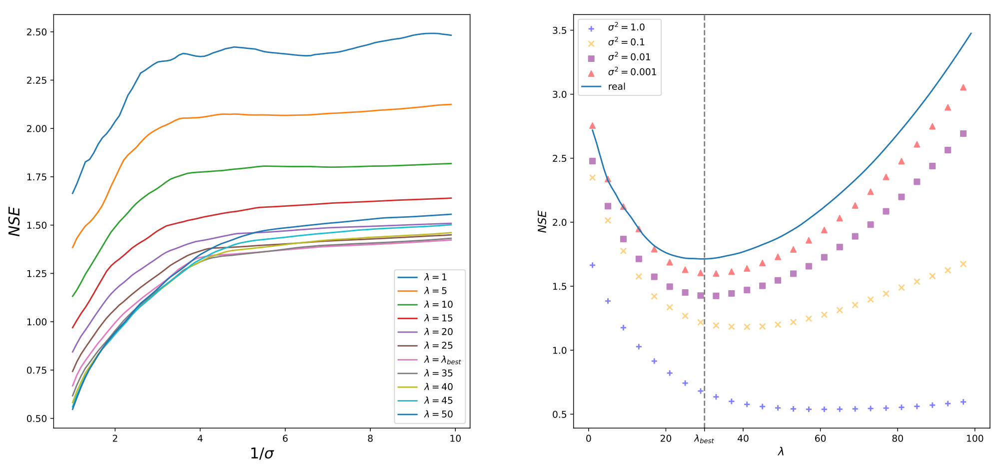
Figures F1 $\sim$ F7 (a) demonstrate that when λ is fixed, NSE converges to a constant as $\sigma \to 0$. In Figures F1 $\sim$ F7 (b), the curves converge towards the real blue curve as $\sigma \to 0$, with the real blue curve representing $\Delta(\hat{\lambda})$. These observation validate Theorem 5.2.

https://doi.org/10.52202/079017-2420



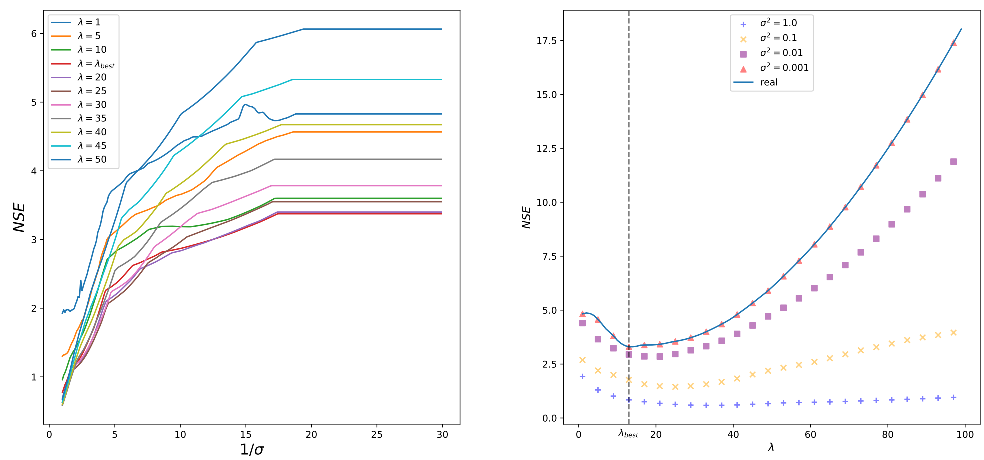
(a) The change of NSE with $1/\sigma$ under different λ. (b) The change of NSE with λ under different σ.

Figure F1: The change of NSE under $\frac{n}{d} = 0.5$, $\frac{k}{d} = 0.1$, $d = 200$



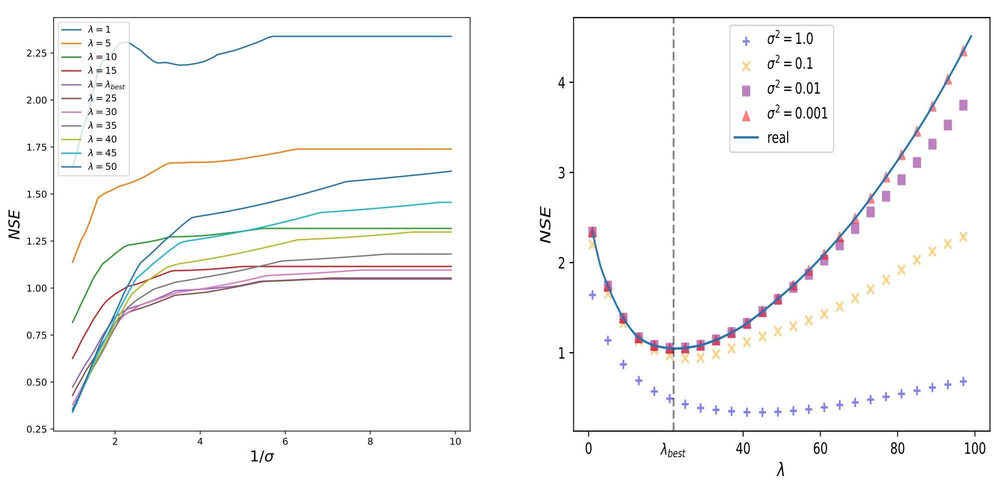
(a) The change of NSE with $1/\sigma$ under different λ. (b) The change of NSE with λ under different σ.

Figure F2: The change of NSE under $\frac{n}{d} = 0.5$, $\frac{k}{d} = 0.1$, $d = 1000$



(a) The change of NSE with $1/\sigma$ under different λ. (b) The change of NSE with λ under different σ.

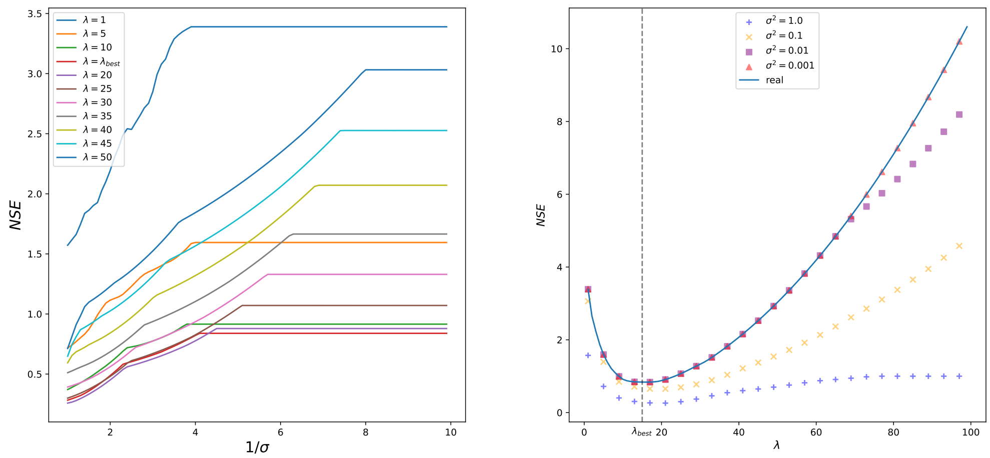
Figure F3: The change of NSE under $\frac{n}{d} = 0.4$, $\frac{k}{d} = 0.1$, $d = 1000$



(a) The change of NSE with $1/\sigma$ under different λ. (b) The change of NSE with λ under different σ.

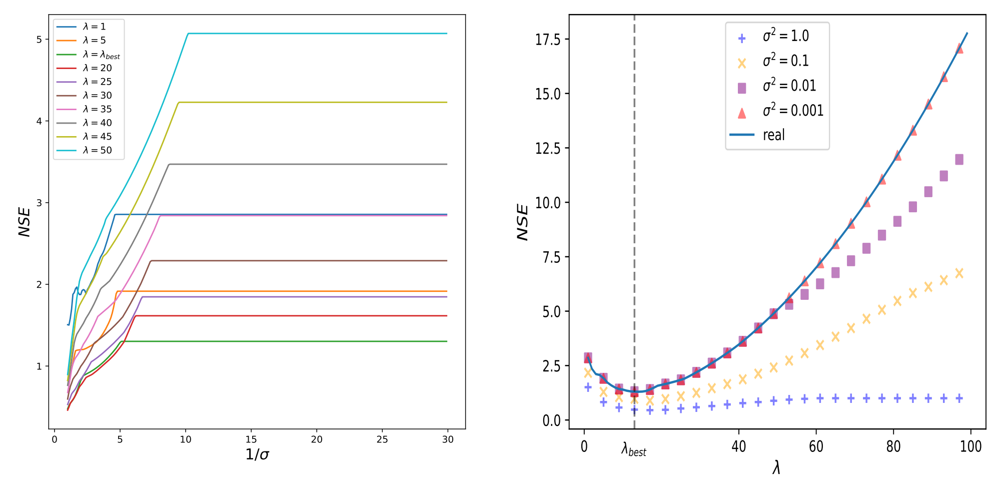
Figure F4: The change of NSE under $\frac{n}{d} = 0.6$, $\frac{k}{d} = 0.1$, $d = 1000$

https://doi.org/10.52202/079017-2420



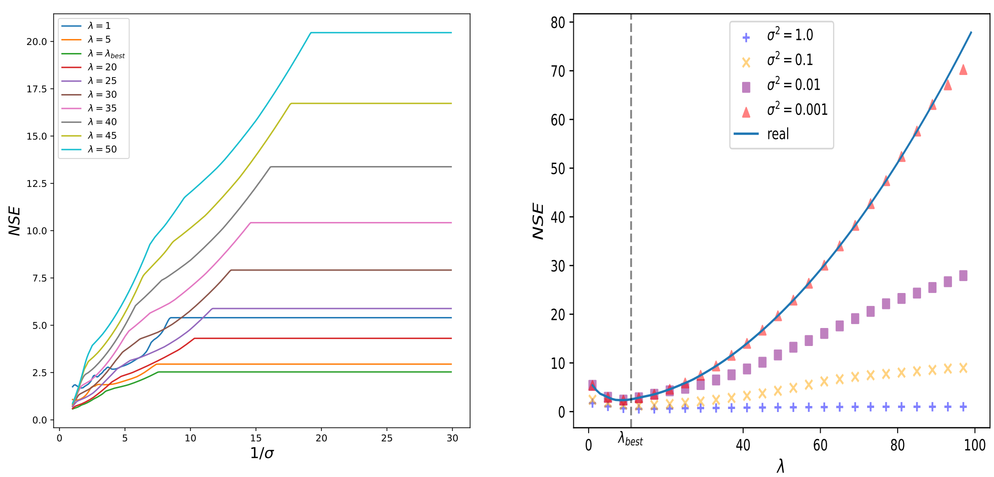
(a) The change of NSE with $1/\sigma$ under different λ. (b) The change of NSE with λ under different σ.

Figure F5: The change of NSE under the sparsity $\frac{k}{d} = 0.10$



(a) The change of NSE with $1/\sigma$ under different λ. (b) The change of NSE with λ under different σ.

Figure F6: The change of NSE under the sparsity $\frac{k}{d} = 0.15$



(a) The change of NSE with $1/\sigma$ under different λ. (b) The change of NSE with λ under different σ.

Figure F7: The change of NSE under the sparsity $\frac{k}{d} = 0.20$

 https://doi.org/10.52202/079017-2420

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We highly summarize what we do in the abstract. The introduction clearly introduces the concepts and issues that are related to our main results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our paper explicitly acknowledges the dependency on Gaussian inputs as a limitation, prompting further investigation into its applicability to non-Gaussian contexts.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-2420

Answer: [Yes]

Justification: We provide the assumptions of the theorems and all our theorems are followed by their proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our paper meticulously details all methodologies, parameters, and experimental conditions necessary for replication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-2420

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: We focus on the OKRidge algorithm in our paper. The codes of the OKRidge algorithm are from [1], which is open.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Our paper comprehensively details all training and testing protocols, including data management, hyperparameter selection, and optimization techniques.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our results are based on averaging the results from 10 trials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our paper specifies the computing platform, processor details, and memory constraints, ensuring accurate replication of the experimental setup and results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research strictly adheres to the NeurIPS Code of Ethics, focusing on theoretical advancements without ethical concerns in implementation or application.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our paper assesses the broad accessibility and potential application impacts of OKRidge, highlighting its advantages over costlier alternatives.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-2420

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper focuses on theoretical aspects of the OKRidge algorithm and uses synthetic data, mitigating risks associated with real-world data misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our paper properly credits the original creators of the OKRidge algorithm and adheres to the licensing terms of the cited open resources required by NeurIPS paper [1].

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We don't introduce new assets in our paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] ,

Justification: Our paper is theoretical and includes numerical verification experiments. It does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper is theoretical and includes numerical verification experiments. It does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-2420

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2420