Sparse Learning for Large-Scale and High-Dimensional Data: A Randomized Convex-Concave Optimization Approach

Lijun Zhang[1(✉)], Tianbao Yang[2], Rong Jin[3], and Zhi-Hua Zhou[1]

[1] National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China
{zhanglj,zhouzh}@lamda.nju.edu.cn

[2] Department of Computer Science, The University of Iowa, Iowa City 52242, USA
tianbao-yang@uiowa.edu

[3] Alibaba Group, Seattle, USA
jinrong.jr@alibaba-inc.com

Abstract. In this paper, we develop a randomized algorithm and theory for learning a sparse model from large-scale and high-dimensional data, which is usually formulated as an empirical risk minimization problem with a sparsity-inducing regularizer. Under the assumption that there exists a (approximately) sparse solution with high classification accuracy, we argue that the dual solution is also sparse or approximately sparse. The fact that both primal and dual solutions are sparse motivates us to develop a randomized approach for a general convex-concave optimization problem. Specifically, the proposed approach combines the strength of random projection with that of sparse learning: it utilizes random projection to reduce the dimensionality, and introduces ℓ_1-norm regularization to alleviate the approximation error caused by random projection. Theoretical analysis shows that under favored conditions, the randomized algorithm can accurately recover the optimal solutions to the convex-concave optimization problem (i.e., recover both the primal and dual solutions).

Keywords: Random projection · Sparse learning · Convex-concave optimization · Primal solution · Dual solution

1 Introduction

Learning the sparse representation of a predictive model has received considerable attention in recent years [4]. Given a set of training examples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, the optimization problem is generally formulated as

$$\min_{\mathbf{w} \in \Omega} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \mathbf{x}_i^\top \mathbf{w}) + \gamma \psi(\mathbf{w}) \tag{1}$$

where $\ell(\cdot)$ is a convex function such as the logistic loss to measure the empirical error, and $\psi(\cdot)$ is a sparsity-inducing regularizer such as the elastic net [38]

R. Ortner et al. (Eds.): ALT 2016, LNAI 9925, pp. 83–97, 2016.
DOI: 10.1007/978-3-319-46379-7_6

to avoid overfitting [13]. When both d and n are very large, directly solving (1) could be computationally expensive. A straightforward way to address this challenge is first reducing the dimensionality of the data, then solving a low-dimensional problem, and finally mapping the solution back to the original space. The limitation of this approach is that the final solution, after mapping from the low-dimensional space to the original high-dimensional space, may not be sparse.

The goal of this paper is to develop an efficient algorithm for solving the problem in (1), and at the same time preserve the (approximate) sparsity of the solution. Our approach is motivated by the following simple observation:

> If there exists a sparse model with high prediction accuracy, the dual solution to (1) is also sparse or approximately sparse.

To see this, let us formulate (1) as a convex-concave optimization problem. By writing $\ell(z)$ in its convex conjugate form, i.e.,

$$\ell(z) = \max_{\lambda \in \mathit{\Gamma}} \lambda z - \ell_*(\lambda),$$

where $\ell_*(\cdot)$ is the Fenchel conjugate of $\ell(\cdot)$ [27] and $\mathit{\Gamma}$ is the domain of the dual variable, we get the following convex-concave formulation:

$$\max_{\boldsymbol{\lambda} \in \mathit{\Gamma}^n} \min_{\mathbf{w} \in \mathit{\Omega}} \ \gamma n \psi(\mathbf{w}) - \sum_{i=1}^{n} \ell_*(\lambda_i) + \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i^\top \mathbf{w}. \tag{2}$$

Denote the optimal solutions to (2) by $(\mathbf{w}_*, \boldsymbol{\lambda}_*)$. By the Fenchel conjugate theory [9, Lemma 11.4], we have

$$[\boldsymbol{\lambda}_*]_i = \ell'(y_i \mathbf{x}_i^\top \mathbf{w}_*).$$

Let us consider the squared hinge loss for classification [31], where $\ell(z) = \max(0, 1-z)^2$. Therefore, $y_i \mathbf{x}_i^\top \mathbf{w}_* \geq 1$ indicates that $[\boldsymbol{\lambda}_*]_i = 0$. As a result, when most of the examples can be classified by a large margin (which is likely to occur in large-scale and high-dimensional setting), it is reasonable to assume that the dual solution is sparse. Similarly, for logistic regression, we can argue the dual solution is approximately sparse.

Abstracting (2) slightly, in the following, we will study a general convex-concave optimization problem:

$$\max_{\boldsymbol{\lambda} \in \mathit{\Delta}} \min_{\mathbf{w} \in \mathit{\Omega}} \ g(\mathbf{w}) - h(\boldsymbol{\lambda}) - \mathbf{w}^\top A \boldsymbol{\lambda} \tag{3}$$

where $\mathit{\Delta} \subseteq \mathbb{R}^n$ and $\mathit{\Omega} \subseteq \mathbb{R}^d$ are the domains for $\boldsymbol{\lambda}$ and $\mathbf{w}$, respectively, $g(\cdot)$ and $h(\cdot)$ are two convex functions, and $A \in \mathbb{R}^{d \times n}$ is a matrix. The benefit of analyzing (3) instead of (1) is that the convex-concave formulation allows us to exploit the prior knowledge that *both* $\mathbf{w}_*$ and $\boldsymbol{\lambda}_*$ are sparse or approximately sparse. The problem in (3) has been widely studied in the optimization community, and when n and d are medium size, it can be solved iteratively by gradient based methods [21, 22].

We assume the two convex functions $g(\cdot)$ and $h(\cdot)$ are relatively simple such that evaluating their values or gradients takes $O(d)$ and $O(n)$ complexities, respectively. The bottleneck is the computations involving the bilinear term $\mathbf{w}^\top A\boldsymbol{\lambda}$, which have $O(nd)$ complexity in both time and space. To overcome this difficulty, we develop a randomized algorithm that solves (3) approximately but at a significantly lower cost. The proposed algorithm combines two well-known techniques—*random projection* and *ℓ_1-norm regularization* in a principled way. Specifically, random projection is used to find a low-rank approximation of A, which not only reduces the storage requirement but also accelerates the computations. The role of ℓ_1-norm regularization is twofold. One one hand, it is introduced to compensate for the distortion caused by randomization, and on the other hand it enforces the sparsity of the final solutions. Under mild assumptions about the optimization problem in (3), the proposed algorithm has a small recovery error provided the optimal solutions to (3) are sparse or approximately sparse.

2 Related Work

Random projection has been widely used as an efficient algorithm for dimensionality reduction [6,16]. In the case of unsupervised learning, it has been proved that random projection is able to preserve the distance [11], inner product [3], volumes and distance to affine spaces [18]. In the case of supervised learning, random projection is generally used as a preprocessing step to find a low-dimensional representation of the data, and thus reduces the computational cost of training. For classification, theoretical studies mainly focus on examining the generalization error or the preservation of classification margin in the low-dimensional space [5,24,28]. For regression, there do exist theoretical guarantees for the recovery error, but they only hold for the least squares problem [19].

Our work is closely related to Dual Random Projection (DRP) [35,36] and Dual-sparse Regularized Randomized Reduction (DSRR) [34], which also investigate random projection from the perspective of optimization. However, both DRP and DSRR are limited to the special case that $\psi(\mathbf{w}) = \|\mathbf{w}\|_2^2$, which leads to a simple dual problem. In contrast, our algorithm is designed for the case that $\psi(\cdot)$ is a sparsity-inducing regularizer, and built upon the convex-concave formulation. Similar to DSRR, our algorithm makes use of the sparsity of the dual solution, but we further exploit the sparsity of the primal solution. A noticeable advantage of our analysis is the mild assumption about the data matrix A. To recover the primal solution, DRP assumes the data matrix is low-rank and DSRR assumes it satisfies the restricted eigenvalue condition, in contrast, our algorithm only requires columns or rows of A are bounded.

There are many literatures that study the statistical property of the sparse learning problem in (1) [2,23,33,37]. For example, in the context of compressive sensing [12], it has been established that a sparse signal can be recovered up to an $O(\sqrt{s \log d/n})$ error, where s is the sparsity of the unknown signal. We note that the statistical error is not directly comparable to the optimization error derived in this paper. That is because the analysis of statistical error relies on heavy assumptions about the data, e.g., the RIP condition [8]. On the other hand, the optimization error is derived under very weak conditions.

3 Algorithm

To reduce the computational cost of (3), we first generate a random matrix $R \in \mathbb{R}^{n\times m}$, where $m \ll \min(d, n)$. Define $\widehat{A} = AR \in \mathbb{R}^{d\times m}$, we propose to solve the following problem

$$\max_{\boldsymbol{\lambda}\in\Delta} \min_{\mathbf{w}\in\Omega} \; g(\mathbf{w}) - h(\boldsymbol{\lambda}) - \mathbf{w}^\top \widehat{A} R^\top \boldsymbol{\lambda} + \gamma_w \|\mathbf{w}\|_1 - \gamma_\lambda \|\boldsymbol{\lambda}\|_1 \tag{4}$$

where γ_w and γ_λ are two regularization parameters. The construction of the random matrix R, as well as the values of the two regularization parameters γ_w and γ_λ will be discussed later. The optimization problem in (4) can be solved by algorithms designed for composite convex-concave problems [10, 14].

Compared to (3), the main advantage of (4) is that it only needs to load $\widehat{A}$ and R into the memory, making it convenient to deal with large-scale problems. With the help of random projection, the computational complexity for evaluating the value and gradient is reduced from $O(dn)$ to $O(dm+nm)$. Compared to previous randomized algorithms [5, 34, 35], (4) has two new features: (i) the optimization is still performed in the original space; and (ii) the ℓ_1-norm is introduced to regularize both primal and dual solutions. As we will prove later, the combination of these two features will ensure the solutions to (4) are approximately sparse. Finally, note that in (4) $RR^\top$ is inserted at the right side of A, it can also be put at the left side of A. In this case, we have the following optimization problem

$$\max_{\boldsymbol{\lambda}\in\Delta} \min_{\mathbf{w}\in\Omega} \; g(\mathbf{w}) - h(\boldsymbol{\lambda}) - \mathbf{w}^\top R \widehat{A} \boldsymbol{\lambda} + \gamma_w \|\mathbf{w}\|_1 - \gamma_\lambda \|\boldsymbol{\lambda}\|_1 \tag{5}$$

where $R \in \mathbb{R}^{d\times m}$ is a random matrix, and $\widehat{A} = R^\top A \in \mathbb{R}^{m\times n}$.

Let $(\mathbf{w}_*, \boldsymbol{\lambda}_*)$ and $(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\lambda}})$ be the optimal solution to the convex-concave optimization problem in (3) and (4)/(5), respectively. Under suitable conditions, we will show that

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \le O\left(\sqrt{\frac{\|\mathbf{w}_*\|_0 \|\boldsymbol{\lambda}_*\|_0 \log n}{m}}\right) \text{ and}$$

$$\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2 \le O\left(\sqrt{\frac{\|\mathbf{w}_*\|_0 \|\boldsymbol{\lambda}_*\|_0 \log d}{m}}\right)$$

implying a small recovery error when $\mathbf{w}_*$ and $\boldsymbol{\lambda}_*$ are sparse. A similar recovery guarantee also holds when the optimal solutions to (3) are approximately sparse, i.e., when they can be well-approximated by sparse vectors.

4 Main Results

We first introduce common assumptions that we make, and then present theoretical guarantees.

4.1 Assumptions

Assumptions About (3). We make the following assumptions about (3).

- $g(\mathbf{w})$ is α-strongly convex with respect to the Euclidean norm. Let's take the optimization problem in (2) as an example. (2) will satisfy this assumption if some strongly convex function (e.g., $\|\mathbf{w}\|_2^2$) is a part of the regularizer $\psi(\mathbf{w})$.
- $h(\boldsymbol{\lambda})$ is β-strongly convex with respect to the Euclidean norm. For the problem in (2), if $\ell(\cdot)$ is a smooth function (e.g., the logistic loss), then its convex conjugate $\ell_*(\cdot)$ will be strongly convex [15,27].
- Either columns or rows of A have bounded ℓ_2-norm. Without loss of generality, we assume

$$\|A_{i*}\|_2 \leq 1, \ \forall i \in [d], \tag{6}$$

$$\|A_{*j}\|_2 \leq 1, \ \forall j \in [n]. \tag{7}$$

  The above assumption can be satisfied by normalizing rows or columns of A.

Assumptions About R. We assume the random matrix $R \in \mathbb{R}^{n \times m}$ has the following property.

- With a high probability, the linear operator $R^\top : \mathbb{R}^n \mapsto \mathbb{R}^m$ is able to preserve the ℓ_2-norm of its input. In mathematical terms, we need the following property.

Property 1. There exists a constant $c > 0$, such that

$$\Pr\left\{(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \|R^\top \mathbf{x}\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2\right\} \geq 1 - 2\exp(-m\varepsilon^2/c)$$

for any fixed $\mathbf{x} \in \mathbb{R}^d$ and $0 < \epsilon \leq 1/2$.

The above property is widely used to prove the famous Johnson–Lindenstrauss lemma [11]. Let $R = \frac{1}{\sqrt{m}} S$. Previous studies [1,3] have proved that Property 1 is true if $\{S_{ij}\}$ are independent random variables sampled from the Gaussian distribution $\mathcal{N}(0,1)$, uniform distribution over $\{\pm 1\}$, or the following database-friendly distribution

$$X = \begin{cases} \sqrt{3}, & \text{with probability } 1/6; \\ 0, & \text{with probability } 2/3; \\ -\sqrt{3}, & \text{with probability } 1/6. \end{cases}$$

More generally, a sufficient condition for Property 1 is that columns of R are independent, isotropic, and subgaussian vectors [20].

4.2 Theoretical Guarantees

Sparse Solutions. We first consider the case that both $\mathbf{w}_*$ and $\boldsymbol{\lambda}_*$ are sparse. Define

$$s_w = \|\mathbf{w}_*\|_0, \text{ and } s_\lambda = \|\boldsymbol{\lambda}_*\|_0.$$

We have the following theorem for the optimization problem in (4).

Theorem 1. *Let* $(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\lambda}})$ *be the optimal solution to the problem in* (4). *Set*

$$\gamma_\lambda \geq 2\|A^\top \mathbf{w}_*\|_2 \sqrt{\frac{c}{m} \log \frac{4n}{\delta}}, \tag{8}$$

$$\gamma_w \geq 2\|\boldsymbol{\lambda}_*\|_2 \sqrt{\frac{c}{m} \log \frac{4d}{\delta}} + \frac{6\gamma_\lambda \sqrt{s_\lambda}}{\beta} \left(1 + 7\sqrt{\frac{c}{m}\left(\log \frac{4d}{\delta} + 16 s_\lambda \log \frac{9n}{8 s_\lambda}\right)}\right). \tag{9}$$

With a probability at least $1 - 3\delta$, *we have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \leq \frac{3\gamma_w \sqrt{s_w}}{\alpha}, \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1 \leq \frac{12\gamma_w s_w}{\alpha}, \ \textit{and} \ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1}{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2} \leq 4\sqrt{s_w}$$

provided

$$m \geq 4c \log \frac{4}{\delta} \tag{10}$$

where c *is the constant in Property* 1.

Notice that $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1 / \|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \leq 4\sqrt{s_w}$ indicates that $\widehat{\mathbf{w}} - \mathbf{w}_*$ is approximately sparse [25,26]. Combining with the fact $\mathbf{w}_*$ is sparse, we conclude that $\widehat{\mathbf{w}}$ is also approximately sparse.

Then, we discuss the recovery guarantee for the sparse learning problem in (1) or (2). Since $A^\top \mathbf{w}_* \in \mathbb{R}^n$, we can take $\|A^\top \mathbf{w}_*\|_2 = O(\sqrt{n})$. Since $\|\boldsymbol{\lambda}_*\|_0 = s_\lambda$, we can assume $\|\boldsymbol{\lambda}_*\|_2 = O(\sqrt{s_\lambda})$. According to the theoretical analysis of regularized empirical risk minimization [17,29,32], the optimal γ, that minimizes the generalization error, can be chosen as $\gamma = O(1/\sqrt{n})$, and thus $\alpha = O(\gamma n) = O(\sqrt{n})$. When the loss $\ell(\cdot)$ is smooth, we have $\beta = O(1)$. The following corollary provides a simplified result based on the above discussions.

Corollary 1. *Assume* $\|A^\top \mathbf{w}_*\|_2 = O(\sqrt{n})$, $\|\boldsymbol{\lambda}_*\|_2 = O(\sqrt{s_\lambda})$, $\alpha = O(\sqrt{n})$, *and* $\beta = O(1)$. *When* $m \geq O(s_\lambda \log n)$, *we can choose*

$$\gamma_\lambda = O\left(\sqrt{\frac{n \log n}{m}}\right) \ \textit{and} \ \gamma_w = O\left(\sqrt{\frac{s_\lambda \log d}{m}} + \gamma_\lambda \sqrt{s_\lambda}\right) = O\left(\sqrt{\frac{n s_\lambda \log n}{m}}\right)$$

such that with a high probability

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \leq O\left(\frac{\gamma_w \sqrt{s_w}}{\sqrt{n}}\right) = O\left(\sqrt{\frac{s_w s_\lambda \log n}{m}}\right) \ \textit{and} \ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1}{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2} \leq 4\sqrt{s_w}.$$

A natural question to ask is whether similar recovery guarantees for $\widehat{\boldsymbol{\lambda}}$ can be proved under the conditions in Theorem 1. Unfortunately, we are not able to give a positive answer, and only have the following theorem.

Theorem 2. *Assume* γ_λ *satisfies the condition in* (8). *With a probability at least* $1 - \delta$, *we have*

$$\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2 \leq \frac{3\gamma_\lambda \sqrt{s_\lambda}}{\beta} + \frac{2}{\beta}\left(1 + \|RR^\top - I\|_2\right) \|A^\top (\widehat{\mathbf{w}} - \mathbf{w}_*)\|_2$$

provided (10) *holds.*

The upper bound in the above theorem is quite loose, because $\|RR^\top - I\|_2$ is roughly on the order of $n \log n/m$ [30].

Due to the symmetry between $\boldsymbol{\lambda}$ and $\mathbf{w}$, we can recover $\boldsymbol{\lambda}_*$ via (5) instead of (4). Then, by replacing $\mathbf{w}_*$ in Theorem 1 with $\boldsymbol{\lambda}_*$, $\widehat{\mathbf{w}}$ with $\widehat{\boldsymbol{\lambda}}$, n with d, and so on, we obtain the following theoretical guarantee.

Theorem 3. *Let $(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\lambda}})$ be the optimal solution to the problem in (5). Set*

$$\begin{aligned}
&\gamma_w \geq 2\|A\boldsymbol{\lambda}_*\|_2 \sqrt{\frac{c}{m} \log \frac{4d}{\delta}},\\
&\gamma_\lambda \geq 2\|\mathbf{w}_*\|_2 \sqrt{\frac{c}{m} \log \frac{4n}{\delta}} + \frac{6\gamma_w\sqrt{s_w}}{\alpha}\left(1 + 7\sqrt{\frac{c}{m}\left(\log\frac{4n}{\delta} + 16 s_w \log \frac{9d}{8s_w}\right)}\right).
\end{aligned}$$

With a probability at least $1 - 3\delta$, we have

$$\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2 \leq \frac{3\gamma_\lambda\sqrt{s_\lambda}}{\beta}, \ \|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1 \leq \frac{12\gamma_\lambda s_\lambda}{\beta}, \ \textit{and}\ \frac{\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1}{\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \leq 4\sqrt{s_\lambda}$$

provided (10) holds.

To simplify the above theorem, we can take $\|A\boldsymbol{\lambda}_*\|_2 = O(\sqrt{d})$ since $A\boldsymbol{\lambda}_* \in \mathbb{R}^d$. Because (1) has both a constraint and a regularizer, we can assume the optimal primal solution is well-bounded, that is, $\|\mathbf{w}_*\|_2 = O(1)$. Finally, we assume $d \leq O(n)$, and have the following corollary.

Corollary 2. *Assume $\|A\boldsymbol{\lambda}_*\|_2 = O(\sqrt{d})$, $\|\mathbf{w}_*\|_2 = O(1)$, $\alpha = O(\sqrt{n})$, $\beta = O(1)$, and $d \leq O(n)$. When $m \geq O(s_w \log d)$, we can choose*

$$\gamma_w = O\left(\sqrt{\frac{d \log d}{m}}\right) \ \textit{and}\ \gamma_\lambda = O\left(\sqrt{\frac{\log n}{m}} + \gamma_w\sqrt{\frac{s_w}{n}}\right) \leq O\left(\sqrt{\frac{s_w \log d}{m}}\right)$$

such that with a high probability

$$\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2 \leq O\left(\gamma_\lambda\sqrt{s_\lambda}\right) = O\left(\sqrt{\frac{s_w s_\lambda \log d}{m}}\right) \ \textit{and}\ \frac{\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1}{\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \leq 4\sqrt{s_\lambda}.$$

Approximately Sparse Solutions. We now proceed to study the case that the optimal solutions to (3) are only approximately sparse.

With a slight abuse of notation, we assume $\mathbf{w}_*$ and $\boldsymbol{\lambda}_*$ are two sparse vectors, with $\|\mathbf{w}_*\|_0 = s_w$ and $\|\boldsymbol{\lambda}_*\|_0 = s_\lambda$, that solve (3) approximately in the sense that

$$\|\nabla g(\mathbf{w}_*) - A\boldsymbol{\lambda}_*\|_\infty \leq \varsigma, \tag{11}$$

$$\|\nabla h(\boldsymbol{\lambda}_*) + A^\top \mathbf{w}_*\|_\infty \leq \varsigma, \tag{12}$$

for some small constant $\varsigma > 0$. The above conditions can be considered as sub-optimality conditions [7] of $\mathbf{w}_*$ and $\boldsymbol{\lambda}_*$ measured in the ℓ_∞-norm. After a similar analysis, we have the following theorem.

Theorem 4. *Let $(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\lambda}})$ be the optimal solution to the problem in (4). Assume (11) and (12) hold. Set*

$$\gamma_\lambda \geq 2\|A^\top \mathbf{w}_*\|_2 \sqrt{\frac{c}{m} \log \frac{4n}{\delta}} + 2\varsigma,$$

$$\gamma_w \geq 2\|\boldsymbol{\lambda}_*\|_2 \sqrt{\frac{c}{m} \log \frac{4d}{\delta}} + \frac{6\gamma_\lambda \sqrt{s_\lambda}}{\beta} \left(1 + 7\sqrt{\frac{c}{m}\left(\log \frac{4d}{\delta} + 16 s_\lambda \log \frac{9n}{8 s_\lambda}\right)}\right) + 2\varsigma.$$

With a probability at least $1 - 3\delta$, we have

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \leq \frac{3\gamma_w \sqrt{s_w}}{\alpha}, \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1 \leq \frac{12 \gamma_w s_w}{\alpha}, \ \text{and} \ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1}{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2} \leq 4\sqrt{s_w}$$

provided (10) holds.

When ς is small enough, the upper bound in Theorem 4 is on the same order as that in Theorem 1. To be specific, we have the following corollary.

Corollary 3. *Assume $\|A^\top \mathbf{w}_*\|_2 = O(\sqrt{n})$, $\|\boldsymbol{\lambda}_*\|_2 = O(\sqrt{s_\lambda})$, $\alpha = O(\sqrt{n})$, $\beta = O(1)$, and $\varsigma = O(\sqrt{n \log n / m})$. When $m \geq O(s_\lambda \log n)$, we can choose γ_λ and γ_w as in Corollary 1 such that with a high probability*

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = O\left(\frac{\gamma_w \sqrt{s_w}}{\sqrt{n}}\right) = O\left(\sqrt{\frac{s_w s_\lambda \log n}{m}}\right) \ \text{and} \ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1}{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2} \leq 4\sqrt{s_w}.$$

5 Analysis

Due to the limitation of space, we only provide proofs of Theorem 1 and related lemmas. The omitted proofs will be included in a supplementary.

5.1 Proof of Theorem 1

To facilitate the analysis, we introduce a pseudo optimization problem

$$\max_{\boldsymbol{\lambda} \in \Delta} \ -h(\boldsymbol{\lambda}) - \mathbf{w}_*^\top \widehat{A} R^\top \boldsymbol{\lambda} - \gamma_\lambda \|\boldsymbol{\lambda}\|_1$$

whose optimal solution is denoted by $\widetilde{\boldsymbol{\lambda}}$. In the following, we will first discuss how to bound the difference between $\widetilde{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda}_*$, and then bound the difference between $\widehat{\mathbf{w}}$ and $\mathbf{w}_*$ in a similar way.

From the optimality of $\widetilde{\boldsymbol{\lambda}}$ and $\boldsymbol{\lambda}_*$, we derive the following lemma to bound their difference.

Lemma 1. *Denote*

$$\rho_\lambda = \left\|(RR^\top - I) A^\top \mathbf{w}_*\right\|_\infty . \tag{13}$$

By choosing $\gamma_\lambda \geq 2\rho_\lambda$, we have

$$\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2 \leq \frac{3\gamma_\lambda \sqrt{s_\lambda}}{\beta}, \ \|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1 \leq \frac{12 \gamma_\lambda s_\lambda}{\beta}, \ \text{and} \ \frac{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1}{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \leq 4\sqrt{s_\lambda}.$$

Based on the property of the random matrix R described in Property 1, we have the following lemma to bound ρ_λ in (13).

Lemma 2. *With a probability at least* $1-\delta$*, we have*

$$\rho_\lambda = \left\|(RR^\top - I)A^\top \mathbf{w}_*\right\|_\infty \leq \|A^\top \mathbf{w}_*\|_2 \sqrt{\frac{c}{m}\log\frac{4n}{\delta}}$$

provided (10) *holds.*

Combining Lemma 1 with Lemma 2, we immediately obtain the following lemma.

Lemma 3. *Set*

$$\gamma_\lambda \geq 2\|A^\top \mathbf{w}_*\|_2 \sqrt{\frac{c}{m}\log\frac{4n}{\delta}}.$$

With a probability at least $1-\delta$*, we have*

$$\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2 \leq \frac{3\gamma_\lambda\sqrt{s_\lambda}}{\beta}, \ \|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1 \leq \frac{12\gamma_\lambda s_\lambda}{\beta}, \ \textit{and} \ \frac{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1}{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \leq 4\sqrt{s_\lambda}$$

provided (10) *holds.*

We are now in a position to formulate the key lemmas that lead to Theorem 1. Similar to Lemma 1, we introduce the following lemma to characterize the relation between $\widehat{\mathbf{w}}$ and $\mathbf{w}_*$.

Lemma 4. *Denote*

$$\rho_w = \left\|A(I - RR^\top)\boldsymbol{\lambda}_*\right\|_\infty + \left\|ARR^\top(\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}})\right\|_\infty. \tag{14}$$

By choosing $\gamma_w \geq 2\rho_w$*, we have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \leq \frac{3\gamma_w\sqrt{s_w}}{\alpha}, \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1 \leq \frac{12\gamma_w s_w}{\alpha}, \ \textit{and} \ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_1}{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2} \leq 4\sqrt{s_w}.$$

The last step of the proof is to derive an upper bound for ρ_w based on Property 1 and Lemma 3.

Lemma 5. *Assume the conclusion in Lemma 3 happens. With a probability at least* $1-2\delta$*, we have*

$$\rho_w \leq \|\boldsymbol{\lambda}_*\|_2 \sqrt{\frac{c}{m}\log\frac{4d}{\delta}} + \frac{3\gamma_\lambda\sqrt{s_\lambda}}{\beta}\left(1 + 7\sqrt{\frac{c}{m}\left(\log\frac{4d}{\delta} + 16 s_\lambda \log\frac{9n}{8s_\lambda}\right)}\right)$$

provided (10) *holds.*

5.2 Proof of Lemma 1

Notations. For a vector $\mathbf{x} \in \mathbb{R}^d$ and a set $\mathcal{D} \subseteq [d]$, we denote by $\mathbf{x}_{\mathcal{D}}$ the vector which coincides with $\mathbf{x}$ on $\mathcal{D}$ and has zero coordinates outside $\mathcal{D}$.

Let Ω_λ include the subset of non-zeros entries in $\boldsymbol{\lambda}_*$ and $\bar{\Omega}_\lambda = [n] \setminus \Omega_\lambda$. Define

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}) &= -h(\boldsymbol{\lambda}) + \min_{\mathbf{w} \in \Omega} g(\mathbf{w}) - \mathbf{w}^\top A \boldsymbol{\lambda}, \\ \widetilde{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) &= -h(\boldsymbol{\lambda}) - \mathbf{w}_*^\top \widehat{A} R^\top \boldsymbol{\lambda} - \gamma_\lambda \|\boldsymbol{\lambda}\|_1. \end{aligned}$$

Let $\mathbf{v} \in \partial \|\boldsymbol{\lambda}_*\|_1$ be any subgradient of $\|\cdot\|_1$ at $\boldsymbol{\lambda}_*$. Then, we have[1]

$$\mathbf{u} = -\nabla h(\boldsymbol{\lambda}_*) - R R^\top A^\top \mathbf{w}_* - \gamma_\lambda \mathbf{v} \in \partial \widetilde{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_*).$$

Using the fact that $\widetilde{\boldsymbol{\lambda}}$ maximizes $\widetilde{\boldsymbol{\lambda}}(\cdot)$ over the domain Δ and $h(\cdot)$ is β-strongly convex, we have

$$\begin{aligned} 0 \geq \widetilde{\boldsymbol{\lambda}}(\boldsymbol{\lambda}_*) - \widetilde{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\lambda}}) \geq & \langle -(\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*), \mathbf{u} \rangle + \frac{\beta}{2} \|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2^2 \\ = & \left\langle \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*, \nabla h(\boldsymbol{\lambda}_*) + R R^\top A^\top \mathbf{w}_* + \gamma_\lambda \mathbf{v} \right\rangle + \frac{\beta}{2} \|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2^2. \end{aligned} \tag{15}$$

By setting $v_i = \operatorname{sign}(\widetilde{\lambda}_i)$, $\forall i \in \bar{\Omega}_\lambda$, we have $\langle \widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}, \mathbf{v}_{\bar{\Omega}_\lambda} \rangle = \|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1$. As a result,

$$\langle \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*, \mathbf{v} \rangle = \langle \widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}, \mathbf{v}_{\bar{\Omega}_\lambda} \rangle + \langle \widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*, \mathbf{v}_{\Omega_\lambda} \rangle \geq \|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1 - \|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1. \tag{16}$$

Combining (15) with (16), we have

$$\left\langle \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*, \nabla h(\boldsymbol{\lambda}_*) + R R^\top A^\top \mathbf{w}_* \right\rangle + \frac{\beta}{2} \|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2^2 + \gamma_\lambda \|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1 \leq \gamma_\lambda \|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1. \tag{17}$$

From the fact that $\boldsymbol{\lambda}_*$ maximizes $\mathcal{L}(\cdot)$ over the domain Δ, we have

$$\langle \nabla \mathcal{L}(\boldsymbol{\lambda}_*), \boldsymbol{\lambda} - \boldsymbol{\lambda}_* \rangle = \langle -\nabla h(\boldsymbol{\lambda}_*) - A^\top \mathbf{w}_*, \boldsymbol{\lambda} - \boldsymbol{\lambda}_* \rangle \leq 0, \ \forall \boldsymbol{\lambda} \in \Delta. \tag{18}$$

Then,

$$\begin{aligned} & \left\langle \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*, \nabla h(\boldsymbol{\lambda}_*) + R R^\top A^\top \mathbf{w}_* \right\rangle \\ = & \left\langle \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*, \nabla h(\boldsymbol{\lambda}_*) + A^\top \mathbf{w}_* \right\rangle + \left\langle \widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*, (R R^\top - I) A^\top \mathbf{w}_* \right\rangle \\ \overset{(18)}{\geq} & - \|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1 \left\| (R R^\top - I) A^\top \mathbf{w}_* \right\|_\infty \\ \overset{(13)}{=} & - \rho_\lambda \|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1 = -\rho_\lambda \left(\|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1 + \|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1 \right). \end{aligned} \tag{19}$$

From (17) and (19), we have

$$\frac{\beta}{2} \|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2^2 + (\gamma_\lambda - \rho_\lambda) \|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1 \leq (\gamma_\lambda + \rho_\lambda) \|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1.$$

[1] In the case that $h(\cdot)$ is non-smooth, $\nabla h(\boldsymbol{\lambda}_*)$ refers to a subgradient of $h(\cdot)$ at $\boldsymbol{\lambda}_*$. In particular, we choose the subgradient that satisfies (18).

Since $\gamma_\lambda \geq 2\rho_\lambda$, we have

$$\frac{\beta}{2}\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2^2 + \frac{\gamma_\lambda}{2}\|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1 \leq \frac{3\gamma_\lambda}{2}\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1.$$

And thus,

$$\begin{aligned}
&\frac{\beta}{2}\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2^2 \leq \frac{3\gamma_\lambda}{2}\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1 \leq \frac{3\gamma_\lambda\sqrt{s_\lambda}}{2}\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_2 \\
\Rightarrow &\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2 \leq \frac{3\gamma_\lambda\sqrt{s_\lambda}}{\beta}, \\
&\frac{\beta}{2s_\lambda}\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1^2 \leq \frac{\beta}{2}\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2^2 \leq \frac{3\gamma_\lambda}{2}\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1 \\
\Rightarrow &\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1 \leq \frac{3\gamma_\lambda s_\lambda}{\beta}, \\
&\frac{\gamma_\lambda}{2}\|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1 \leq \frac{3\gamma_\lambda}{2}\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1 \\
\Rightarrow &\|\widetilde{\boldsymbol{\lambda}}_{\bar{\Omega}_\lambda}\|_1 \leq 3\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1 \Rightarrow \|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1 \leq \frac{12\gamma_\lambda s_\lambda}{\beta}, \\
&\frac{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_1}{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \leq \frac{4\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_1}{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \leq \frac{4\sqrt{s_\lambda}\|\widetilde{\boldsymbol{\lambda}}_{\Omega_\lambda} - \boldsymbol{\lambda}_*\|_2}{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \leq 4\sqrt{s_\lambda}.
\end{aligned}$$

5.3 Proof of Lemma 2

We first introduce one lemma that is central to our analysis. From the property that R preserves the ℓ_2-norm, it is easy to verify that it also preserves the inner product [3]. Specifically, we have the following lemma.

Lemma 6. *Assume R satisfies Property 1. For any two fixed vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^n$, with a probability at least $1 - \delta$, we have*

$$\left|\mathbf{u}^\top RR^\top \mathbf{v} - \mathbf{u}^\top \mathbf{v}\right| \leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2\sqrt{\frac{c}{m}\log\frac{4}{\delta}}.$$

provided (10) *holds.*

Let $\mathbf{e}_j$ be the j-th standard basis vector of $\mathbb{R}^n$. From Lemma 6, we have with a probability at least $1 - \delta$,

$$\left|\left[(RR^\top - I)A^\top \mathbf{w}_*\right]_j\right| = \left|\mathbf{e}_j^\top (RR^\top - I)A^\top \mathbf{w}_*\right| \leq \|A^\top \mathbf{w}_*\|_2\sqrt{\frac{c}{m}\log\frac{4}{\delta}}$$

for each $j \in [n]$. We complete the proof by taking the union bound over all $j \in [n]$.

5.4 Proof of Lemma 5

We first upper bound ρ_w as

$$\rho_w \leq \underbrace{\left\|A(I-RR^\top)\boldsymbol{\lambda}_*\right\|_\infty}_{:=U_1} + \underbrace{\left\|A(\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}})\right\|_\infty}_{:=U_2} + \underbrace{\left\|A(RR^\top - I)(\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}})\right\|_\infty}_{:=U_3}.$$

Bounding U_1. From Lemma 6, we have with a probability at least $1-\delta$,

$$\begin{aligned}\left|\left[A(I-RR^\top)\boldsymbol{\lambda}_*\right]_i\right| &= \left|A_{i*}(I-RR^\top)\boldsymbol{\lambda}_*\right| \\ &\leq \max_{i\in[d]} \|A_{i*}\|_2\|\boldsymbol{\lambda}_*\|_2\sqrt{\frac{c}{m}\log\frac{4}{\delta}} \overset{(6)}{\leq} \|\boldsymbol{\lambda}_*\|_2\sqrt{\frac{c}{m}\log\frac{4}{\delta}}\end{aligned}$$

for each $i\in[d]$. Taking the union bound over all $i\in[d]$, we have with a probability at least $1-\delta$,

$$\left\|A(I-RR^\top)\boldsymbol{\lambda}_*\right\|_\infty \leq \|\boldsymbol{\lambda}_*\|_2\sqrt{\frac{c}{m}\log\frac{4d}{\delta}}.$$

Bounding U_2. From our assumption, we have

$$\left\|A(\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}})\right\|_\infty \leq \max_{i\in[d]}\|A_{i*}\|_2\|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2 \overset{(6)}{\leq} \|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2.$$

Bounding U_3. Notice that the arguments for bounding U_1 cannot be used to upper bound U_3, that is because $\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}$ is a random variable that depends on R and thus we cannot apply Lemma 6 directly. To overcome this challenge, we will exploit the fact that $\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}$ is approximately sparse to decouple the dependence. Define

$$\mathcal{K}_{n,16s_\lambda} = \left\{\mathbf{x}\in\mathbb{R}^n : \|\mathbf{x}\|_2 \leq 1, \|\mathbf{x}\|_1 \leq 4\sqrt{s_\lambda}\right\}.$$

When the conclusion in Lemma 3 happens, we have

$$\frac{\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*}{\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_*\|_2} \in \mathcal{K}_{n,16s_\lambda} \tag{20}$$

and thus

$$\begin{aligned}U_3 &= \|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2\left\|A(RR^\top - I)\frac{\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}}{\|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2}\right\|_\infty \\ &\overset{(20)}{\leq} \|\boldsymbol{\lambda}_* - \widetilde{\boldsymbol{\lambda}}\|_2 \underbrace{\sup_{\mathbf{z}\in\mathcal{K}_{n,16s_\lambda}}\left\|A(RR^\top - I)\mathbf{z}\right\|_\infty}_{:=U_4}.\end{aligned}$$

Then, we will utilize techniques of covering number to provide an upper bound for U_4.

Lemma 7. *With a probability at least* $1-\delta$*, we have*

$$\sup_{\mathbf{z}\in\mathcal{K}_{n,16s_\lambda}} \left\|A(RR^\top - I)\mathbf{z}\right\|_\infty \le 2(2+\sqrt{2})\sqrt{\frac{c}{m}\left(\log\frac{4d}{\delta} + 16s_\lambda \log\frac{9n}{8s_\lambda}\right)}.$$

Putting everything together, we have

$$\begin{aligned}&\rho_w\\ \le&\|\boldsymbol{\lambda}_*\|_2\sqrt{\frac{c}{m}\log\frac{4d}{\delta}}\\ &+\|\boldsymbol{\lambda}_*-\widetilde{\boldsymbol{\lambda}}\|_2\left(1+2(2+\sqrt{2})\sqrt{\frac{c}{m}\left(\log\frac{4d}{\delta}+16s_\lambda\log\frac{9n}{8s_\lambda}\right)}\right)\\ \le&\|\boldsymbol{\lambda}_*\|_2\sqrt{\frac{c}{m}\log\frac{4d}{\delta}}+\frac{3\gamma_\lambda\sqrt{s_\lambda}}{\beta}\left(1+7\sqrt{\frac{c}{m}\left(\log\frac{4d}{\delta}+16s_\lambda\log\frac{9n}{8s_\lambda}\right)}\right).\end{aligned}$$

6 Conclusion and Future Work

In this paper, a randomized algorithm is proposed to solve the convex-concave optimization problem in (3). Compared to previous studies, a distinctive feature of the proposed algorithm is that ℓ_1-norm regularization is introduced to control the damage cased by random projection. Under mild assumptions about the optimization problem, we demonstrate that it is able to accurately recover the optimal solutions to (3) provided they are sparse or approximately sparse.

From the current analysis, we need to solve two different problems if our goal is to recover both $\mathbf{w}_*$ and $\boldsymbol{\lambda}_*$ accurately. It is unclear whether this is an artifact of the proof technique or actually unavoidable. We will investigate this issue in the future. Since the proposed algorithm is designed for the case that the optimal solutions are (approximately) sparse, it is practically important to develop a pre-precessing procedure that can estimate the sparsity of solutions before applying our algorithm. We plan to utilize random sampling to address this problem. Last but not least, we will investigate the empirical performance of the proposed algorithm.

Acknowledgments. This work was partially supported by NSFC (61333014, 61272217), JiangsuFS (BK20160658, BK20131278), NSF (1463988, 1545995), and the Collaborative Innovation Center of Novel Software Technology and Industrialization of Nanjing University.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4), 671–687 (2003)

2. Agarwal, A., Negahban, S., Wainwright, M.J.: Fast global convergence of gradient methods for high-dimensional statistical recovery. Ann. Stat. **40**(5), 2452–2482 (2012)
3. Arriaga, R.I., Vempala, S.: An algorithmic theory of learning: robust concepts and random projection. Mach. Learn. **63**(2), 161–182 (2006)
4. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Found. Trends Mach. Learn. **4**(1), 1–106 (2012)
5. Balcan, M.F., Blum, A., Vempala, S.: Kernels as features: on kernels, margins, and low-dimensional mappings. Mach. Learn. **65**(1), 79–94 (2006)
6. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 245–250 (2001)
7. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
8. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. C.R. Math. **346**(9–10), 589–592 (2008)
9. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, Cambridge (2006)
10. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
11. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and lindenstrauss. Random Struct. Algorithms **22**(1), 60–65 (2003)
12. Davenport, M.A., Duarte, M.F., Eldar, Y.C., Kutyniok, G.: Introduction to compressed sensing (chap. 1). In: Compressed Sensing, Theory and Applications, pp. 1–64. Cambridge University Press (2012)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York (2009)
14. He, Y., Monteiro, R.D.: An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems. Technical report, Georgia Institute of Technology (2014)
15. Kakade, S.M., Shalev-Shwartz, S., Tewari, A.: On the duality of strong convexity and strong smoothness: learning applications and matrix regularization. Technical report, Toyota Technological Institute at Chicago (2009)
16. Kaski, S.: Dimensionality reduction by random mapping: fast similarity computation for clustering. In: Proceedings of the 1998 IEEE International Joint Conference on Neural Networks, vol. 1, pp. 413–418 (1998)
17. Koltchinskii, V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Springer, Heidelberg (2011)
18. Magen, A.: Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications. In: Rolim, J.D.P., Vadhan, S.P. (eds.) RANDOM 2002. LNCS, vol. 2483, pp. 239–253. Springer, Heidelberg (2002)
19. Mahoney, M.W.: Randomized algorithms for matrices and data. Found. Trends Mach. Learn. **3**(2), 123–224 (2011)
20. Mendelson, S., Pajor, A., Tomczak-Jaegermann, N.: Uniform uncertainty principle for Bernoulli and subgaussian ensembles. Constr. Approximation **28**(3), 277–289 (2008)
21. Nemirovski, A.: Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim. **15**(1), 229–251 (2005)
22. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)

23. Omidiran, D., Wainwright, M.J.: High-dimensional variable selection with sparse random projections: measurement sparsity and statistical efficiency. J. Mach. Learn. Res. **11**, 2361–2386 (2010)
24. Paul, S., Boutsidis, C., Magdon-Ismail, M., Drineas, P.: Random projections for support vector machines. In: Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, pp. 498–506 (2013)
25. Plan, Y., Vershynin, R.: One-bit compressed sensing by linear programming. Commun. Pure Appl. Math. **66**(8), 1275–1297 (2013)
26. Plan, Y., Vershynin, R.: Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. IEEE Trans. Inf. Theor. **59**(1), 482–494 (2013)
27. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1997)
28. Shi, Q., Shen, C., Hill, R., van den Hengel, A.: Is margin preserved after random projection? In: Proceedings of the 29th International Conference on Machine Learning (2012)
29. Sridharan, K., Shalev-shwartz, S., Srebro, N.: Fast rates for regularized objectives. Adv. Neural Inf. Process. Syst. **21**, 1545–1552 (2009)
30. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Found. Comput. Math. **12**, 389–434 (2012)
31. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. **6**, 1453–1484 (2005)
32. Wu, Q., Zhou, D.X.: Svm soft margin classifiers: linear programming versus quadratic programming. Neural Comput. **17**(5), 1160–1187 (2005)
33. Xiao, L., Zhang, T.: A proximal-gradient homotopy method for the ℓ_1-regularized least-squares problem. In: Proceedings of the 29th International Conference on Machine Learning, pp. 839–846 (2012)
34. Yang, T., Zhang, L., Jin, R., Zhu, S.: Theory of dual-sparse regularized randomized reduction. In: Proceedings of the 32nd International Conference on Machine Learning (2015)
35. Zhang, L., Mahdavi, M., Jin, R., Yang, T., Zhu, S.: Recovering the optimal solution by dual random projection. In: Proceedings of the 26th Annual Conference on Learning Theory (COLT), pp. 135–157 (2013)
36. Zhang, L., Mahdavi, M., Jin, R., Yang, T., Zhu, S.: Random projections for classification: a recovery approach. IEEE Trans. Inf. Theor. **60**(11), 7300–7316 (2014)
37. Zhang, L., Yang, T., Jin, R., Zhou, Z.H.: A simple homotopy algorithm for compressive sensing. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (2015)
38. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Series B (Stat. Methodol.) **67**(2), 301–320 (2005)