Worst-Case Hardness of Approximation for Sparse Optimization with L_0 Norm

immediate

March 29, 2016

Abstract

In this paper, we consider sparse optimization problems with L_0 norm penalty or constraint. We prove that it is strongly NP-hard to find an approximate optimal solution within certain error bound, unless P = NP. This provides a lower bound for the approximation error of any deterministic polynomial-time algorithm. Applying the complexity result to sparse linear regression reveals a gap between computational accuracy and statistical accuracy: It is intractable to approximate the estimator within constant factors of its statistical error. We also show that differentiating between the best k-sparse solution and the best $(k + d^c)$-sparse solution is computationally hard where d is the dimension of the problem and c is any constant in $[0, 1)$. It suggests that tuning the sparsity level is hard.

1 Introduction

Sparsity is a prominent modeling tool for extracting useful information from high-dimensional data. The goal is to minimize the empirical loss using as few features as possible. A most natural way of imposing sparsity is to penalize the objective with the L_0 norm, as follows.

Problem 1 Given the loss function $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$, regularization parameter $\lambda > 0$, consider the problem

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \ell\left(a_i^T x, b_i\right) + \lambda\|x\|_0,$$

where $A = (a_1, \ldots, a_n)^T \in \mathbb{R}^{n\times d}$, $b = (b_1, \ldots, b_n)^T \in \mathbb{R}^n$ are input data.

A related problem is L_0-constrained optimization. It arises from sparse estimation [29] and sparse recovery [9, 27].

Problem 2 Given the loss function $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$, sparsity parameter K, consider the problem

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \ell\left(a_i^T x, b_i\right) \quad \text{s.t. } \|x\|_0 \le K,$$

where $A = (a_1, \ldots, a_n)^T \in \mathbb{R}^{n\times d}$, $b = (b_1, \ldots, b_n)^T \in \mathbb{R}^n$ are input data.

We will study the computational complexity of Problems 1 and 2. We focus on the case where ℓ is a convex loss function. These problems naturally arise from feature selection, compressive sensing, and sparse approximation. For some special cases of Problem 1, it has been shown that finding an *exact solution* is strongly NP-hard [13, 22]. However, these results have not excluded the possibility of the existence of polynomial-time algorithms with small approximation error.

In this paper, we focus on the *approximibility* of sparse optimization by *deterministic* polynomial-time algorithms. We prove that it is strongly NP-hard to approximate Problems 1 and 2 to within certain levels of suboptimality. We show that there exists a lower bound of the suboptimality error that can be achieved by any tractable algorithm. Our results apply to a variety of machine learning and estimation problems, such as sparse classification and sparse logistic regression. Many of these problems have not been considered in the context of complexity. The hardness of approximation is one of the strongest forms of complexity result for continuous optimization. To the authors' best knowledge, this is the first work on the approximation hardness of Problem 1, and also the first work on the complexity of Problem 2.

Our results on optimization complexity provide new insights into the complexity of sparse feature selection [17, 36]. In the case of sparse regression for linear models, our result on Problem 1 shows that the lower bound of approximation error is significantly larger than the desired small statistical error. In the case where practitioners wish to choose the best sparsity level, our result on Problem 2 shows that it is impossible to know how much the loss function would improve if the allowed number of nonzero variables increases by 1. These observations provide strong evidences for the hardness of variable selection.

Section 2 summarizes related literatures from machine learning, statistics, and mathematical programming. Section 3 presents the main results and their implications for sparse regression and feature selection. Section 4 gives the proofs, and Section 5 draws the conclusion.

2 Background and Related Works

Sparse optimization problems are common in machine learning, estimation, and signal processing. The sparsity penalty plays the important role of variable/feature selection. When used appropriately, it allows one to select a small number of features out of exponentially many candidates.

Problem 1 is widely regarded as a computationally hard problem. However, formal characterization of its complexity is lacked. Due to its computational challenge, many have considered relaxations of the L_0 norm to smooth or even convex penalty functions. A well known example is the LASSO using L_1 norm penalty, which is the best convex approximation to the L_0 norm [31]. However, the use of L_1 norm in replacement of L_0 norm usually leads to a biased estimator. To strike a balance between bias and computation efficiency, nonconvex penalty functions have been studied; see [11, 12, 15, 16, 18, 25, 33]. Since nonconvex penalties are often smooth approximations to the L_0 norm, the computational challenge usually remain. Within the scope of this paper, we focus on the L_0 norm. Yet we conjecture that our results on hardness of approximation remain valid for a broader class of smooth nonconvex penalties.

Problem 2 finds applications in sparse estimation and feature selection; see for example [29]. Many greedy algorithms and thresholding techniques have been developed. Satisfactory practical performances and some theoretical guarantees have been reported, e.g., [1, 5, 7, 23, 30, 34, 35]. Moreover, Problem 2 is closely related to the model selection problem in sparse feature selection. For a given sparsity level K, the optimal solution to Problem 2 is the best K-sparse solution that fits the data set. To select the best sparsity level that fits the data, one usually needs to solve a sequence of instances of Problem 2, corresponding to different values of K.

Within the mathematical programming community, the optimization complexity of Problem 1 has been considered in a few works. The work [22] proved the hardness result for L_2 loss and a relaxed family of penalty functions, including L_0 norm, hard-thresholded penalty [3] and SCAD [15]. They show that the decision problem "whether the optimal value is bounded by a given number" is NP-hard. However, NP-hard problems might still be easy to solve using pseudo-polynomial algorithms if the coding size is small [19]. As a result, a stronger notion of hardness called *strong NP-hardness* is considered. A problem is

strongly NP-hard if every problem in NP can be polynomially reduced to it in a way such that input in the reduced instance are written in unary [32]. The work [13] showed that the L_2-L_p minimization is strongly NP-hard when $p \in (0, 1)$. Later, [20] and [8] proved the strong NP-hardness for two broader class of penalty functions. To the best of our knowledge, none of the existing works studies the approximibility of the sparse penalized problem. Also, none of these works considers the complexity of the sparsity-constrained problem.

Within the theoretical computer science community, there have been several early works on the complexity of sparse recovery, beginning with Arora et. al. [4]. Amaldi and Kann [2] proved that the problem $\min\{\|x\|_0 \mid Ax = b\}$ is not approximable within a factor $2^{\log^{1-\epsilon} d}$ for any $\epsilon > 0$. Meanwhile, Natarajan [27] showed that, given $\epsilon > 0, A$ and b, the problem $\min\{\|x\|_0 \mid \|Ax - b\|_2 \leq \epsilon\}$ is NP-hard without addressing the approximability of the problem. In [14], Davis et. al. proved a similar result that given $\epsilon > 0$ and $M > 0$ where $\alpha_1 n \leq M \leq \alpha_2 n$ for some $0 < \alpha_1 < \alpha_2 < 1$, it is NP-complete to find a solution x such that $\|x\|_0 \leq M$ and $\|Ax - b\| \leq \epsilon$.

More recently, there have been several important results on the complexity of sparse estimation. These results are related to ours but have a very different focus. For example, [17] studied models for sparse recovery and sparse linear regression with subgaussian noises. Assuming that the true solution is K-sparse, it showed that no polynomial-time (randomized) algorithm can find a $K \cdot 2^{\log^{1-\delta} d}$-sparse solution x with $\|Ax - b\|_2^2 \leq d^{C_1} n^{1-C_2}$ *with high probability*, where δ, C_1, C_2 are arbitrary positive scalars. Another work [36] showed that under Gaussian linear model, there exists a gap between the mean square loss that can be achieved by polynomial-time algorithms and the statistically optimal mean squared error. These works focus on estimation of linear models. In contrast, we focus on the optimization problem and worst-case complexity. Although worst-case results are no stronger than average-case results, our results apply to a variety of loss functions including logistic loss that are not strongly convex. We do not make distributional assumption regarding the input data. Our optimization hardness results fill in the gap between combinatorial complexity theory and computation in practice. It provides a useful reference for practitioners who want to call an optimization solver.

3 Main Results

In this section, we present the main results on the approximation hardness of L_0 sparse optimization. We make the following assumption about the loss function ℓ.

Assumption 1. *There exists $k \in \mathbb{Z}^+$ and $b \in \mathbb{Q}^k$ such that $h(y) = \sum_{i=1}^k \ell(y, b_i)$ has the following properties: (i) $h(y)$ is convex. (ii) $h(y)$ has a unique positive minimizer y^*. (iii) There exists $N > 0, \delta_0 > 0$ and $C > 0$ such that $h(y^* + \delta) - h(y^*) \geq C|\delta|^N$ for all $\delta \in (-\delta_0, \delta_0)$.*

Assumption 1 may seem unconventional, yet it is a critical and general assumption about the loss function. We will show that it is easily satisfied in a variety of common statistical models in Section 4.

Given an optimization problem $\min_{x \in X} f(x)$, we say that a solution $\bar{x}$ is ϵ-optimal if $\bar{x} \in X$ and $f(\bar{x}) \leq \inf_{x \in X} f(x) + \epsilon$. Our first result establishes the hardness of approximation of Problem 1.

Theorem 1. *Let Assumption 1 hold, and let $c \in [0, 1)$. It is strongly NP-hard to find a $\lambda \cdot d^c$-optimal solution of Problem 1, where d is the dimension of variable space.*

We denote by $\ell_n(x)$ the normalized loss function given by

$$\ell_n(x) = \frac{1}{n} \sum_{i=1}^n \ell\left(a_i^T x, b_i\right),$$

and denote by x_K^* the best K-sparse solution given by

$$x_K^* \in \text{argmin}\left\{\ell_n(x) \mid \|x\|_0 \leq K\right\}.$$

Our second result concerns the sparsity-constrained Problem 2.

Theorem 2. *Let Assumption 1 hold, and let $c \in [0, 1)$. There does not exist a pseudo polynomial-time algorithm that takes the input of Problem 2 and outputs approximate solutions $(\hat{x}_1, \ldots, \hat{x}_d)$ satisfying*

$$\ell_n(\hat{x}_{K+d^c}) \leq \ell_n(x_K^*),$$

for all $K = 0, d^c, 2d^c, ...,$ unless P=NP.

Theorems 1 and 2 validate the long-lasting belief that optimization involving L_0 norm is hard. More importantly, they provide lower bound of the optimization error that can be achieved by any polynomial-time algorithm. This is one of the strongest forms of hardness result for continuous optimization.

Case Study: Hardness of Approximating BIC Estimators for Sparse Linear Regression

Let us try to understand how significant is the non-approximable error of Problem 1. We consider the special case of linear models with subgaussian noise. Let the input data (A, b) be generated by the linear model $A\bar{x} + \varepsilon = b$, where $\bar{x}$ is the unknown *true* sparse coefficients and ε is a zero-mean multivariate subgaussian noise. Given the data size n and variable dimension d, we let the regularization parameter λ be chosen according to the Bayesian information criterion (BIC) [28], i.e.,

$$\lambda = \Omega(1 + \log d).$$

Now consider the resulting instance of Problem 1, with parameter λ and input (A, b). Its optimal solution $x^* \in \text{argmax}\left\{\ell_n(x) + \frac{\lambda}{n}\|x\|_0\right\}$ is often regarded as the BIC estimator of sparse linear regression. It is known that with high probability, the statistical error of the estimator x^* is very small [6, 24], i.e.,

$$\|\bar{x} - x^*\|_2^2 = \mathcal{O}\left(\frac{\log d}{n}\|\bar{x}\|_0 + \frac{1}{n}\right).$$

The error's logarithmic dependence on d makes it possible to select a few nonzero variables from exponentially many candidates, while keeping the statistical error small (as long as the true solution $\bar{x}$ is indeed sparse).

Let x_ϵ^* be an ϵ-optimal solution to Problem 1 with regularization parameter λ and input (A, b). Under additional assumptions (e.g., restricted eigenvalue condition), the optimization error can be translated into estimation error of the approximate optimal solution x_ϵ^*, i.e.,

$$\|\bar{x} - x_\epsilon^*\|_2^2 = \mathcal{O}\left(\frac{\log d}{n}\|\bar{x}\|_0 + \frac{1}{n} + \epsilon\right).$$

Unfortunately, Theorem 1 tells us that it is not possible to compute in polynomial time an ϵ- optimal solution with $\epsilon \leq \frac{\lambda}{n}d^c$ for any $c \in [0, 1)$. When d is large, we have

$$\mathcal{O}\left(\frac{\log d}{n}\|\bar{x}\|_0 + \frac{1}{n}\right) \ll \frac{\Omega(1 + \log d)}{n}d^c = \frac{\lambda}{n}d^c.$$

In other words, *it is strongly NP-hard to approximate the BIC estimator within constant factors of its desired statistical error.* It illustrates a sharp contrast between statistical properties of sparse estimation and the worst-case computational complexity.

Case Study: Hardness of Tuning the Sparsity Level

Suppose that we are given the input data set (A, b) with d variables/features and n samples. Now we want to find a sparse solution x that approximately minimize the empirical loss $\ell_n(x) = \frac{1}{n}\sum_{i=1}^n \ell(a_i^T x, b_i)$. A practical problem is to find the right sparsity level for the approximate solution. This is essentially a model selection problem.

Finding the sparsity level requires computing the K-sparse solutions

$$x_K^* \in \text{argmin}\left\{\ell_n(x) \mid \|x\|_0 \leq K\right\},$$

for a range of values of K. This can be translated into solving a sequence of L_0 constrained problems (of the form Problem 2) with K ranging from 0 to d. Regardless of the specific model selection procedure, it is inevitable to compute x_K^* for many values of K's, and to compare their empirical losses such as $\ell_n(x_K^*)$ and $\ell_n(x_{K+d^c}^*)$.

Now let us interpret the results of Theorem 2 in the setting of tuning parameter K. Theorem 2 can be translated as follows. There always exists some sparsity level K such that: even if the exact K-sparse solution x_K^* is known, the non-approximable optimization error for the $(K + d^c)$-sparse problem is at least

$$\ell_n(x_K^*) - \ell_n(x_{K+d^c}^*) > 0.$$

The minimal empirical loss using K features is the best possible approximation to the minimal loss using $K + d^c$ features. In other words, we cannot decide whether and how much the objective value will change by increasing the sparsity level from K to $K + d^c$. Even if $\ell_n(x_K^*)$ is known as a benchmark, we can not find a better approximation of $\ell_n(x_{K+d^c}^*)$ in polynomial-time. In summary, Theorem 2 tells us that it is computationally intractable to differentiate between the sparsity levels K and $K + d^c$, unless P=NP. This implies that selection of the sparsity level is computationally intractable.

Remarks As illustrated by preceding case studies, the non-approximibility of Problems 1 and 2 suggests that computing the sparse estimator and tuning the sparsity parameter are hard. The results suggest a fundamental conflict between computation efficiency and estimation accuracy in sparse data analysis.

Although the results seem negative, they should not discourage researchers from studying computational perspectives of sparse optimization. We remark that Theorems 1 and 2 are *worst-case* complexity results. They do not exclude the possibility that, under more stringent modeling and distributional assumptions, the problem could be tractable with certain probability or on average. For example, there have been promising computation results [7] that use mixed integer programming for sparse feature selection. It is an example of algorithms with exponential worst-case complexity but good average-case performance.

4 Applications in Statistics and Machine Learning

In this section, we provide a collection of statistical models and sparsity penalties that can be addressed under our framework. We note that the application of our results are not only limited to these examples.

First, we give examples of the loss function ℓ that arise from common statistical models. We show that all of them satisfy Assumption 1.

1. In least squares regression, the loss function is

$$\sum_{i=1}^n \left(a_i^T x - b_i\right)^2,$$

where $a_i \in \mathbb{R}^d$ is the explanatory variable and $b_i \in \mathbb{R}$ is the response variable. Least square regression is the most common model that provides maximum likelihood estimation of linear model with Gaussian noise. The corresponding loss function in our framework is $\ell(y, b) = (y - b)^2$. It is convex with a unique minimizer at b. (iii) is also satisfied for $N = 1, C = 1, \delta = 1$. As a result, the least squares regression with concave sparsity regularization is strongly NP-hard.

2. In the estimation of linear model with Laplacian noise, the negative log-likelihood function is

$$\sum_{i=1}^{n} \left| a_i^T x - b_i \right|.$$

The corresponding loss function takes the form $\ell(y, b) = |y - b|$. It can be easily seen to satisfy Assumption 1. Similar argument also applies if we replace $|\cdot|$ with $|\cdot|^q$ with $q \geq 1$.

3. In robust regression, the Huber loss [21] is widely used as a mixture of L_1 and L_2 norms. It takes the form

$$L_\delta(a^T x - b) = \begin{cases} \frac{1}{2}|a^T x - b|^2 & |a^T x - b| \leq \delta, \\ \delta(|a^T x - b| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

for some $\delta > 0$. Let $h(y) = \ell(y, b') = L_\delta(y - b)$. We can see that $h(y)$ satisfies all the conditions in Assumption 1. Therefore, the estimation problem using Huber loss and concave sparse regularization is strongly NP-hard.

4. In poisson regression [10], the mean estimation problem employs the negative log-likelihood minimization

$$\min_{x \in \mathbb{R}^d} -\log L(x; A, b) = \min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} (\exp(a_i^T x) - b_i \cdot a_i^T x).$$

We claim that the corresponding loss function $\ell(y, b) = e^y - b \cdot y$ satisfies Assumption 1. Let $b = 2$ and $h(y) = \sum_{i=1}^{k} \ell(y, b_i) = r \cdot e^y - 2 \cdot y$. We then verify that all the three conditions in Assumption 1 are satisfied. (i), (ii) are obvious by our construction. To verify (ii), note that $h(y)$ take its minimum at $\ln 2$ by our construction. To verify (iii), consider the second order Taylor expansion of $h(y)$ at $\ln 2$,

$$h(y + \delta) - h(y) = \frac{e^y \delta^2}{2} + o(\delta^2)$$

Therefore, all the conditions are satisfied, implying that poisson regression with nonconvex sparse penalty is strongly NP-hard.

5. In logistic regression, the negative log-likelihood function minimization is

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} \log(1 + \exp(a_i^T x)) - \sum_{i=1}^{n} b_i \cdot a_i^T x.$$

We claim that the loss function $\ell(y, b) = \log(1 + \exp(y)) - b \cdot y$ satisfies Assumption 1. We claim that $h(y) = \sum_{i=1}^{k} \ell(y, b_i) = 3\log(1 + \exp(y)) - 2y$. For (ii), observe that $\ell(y, b)$ take its minimum at $y = \ln 2$. To verify (iii), we just need to consider the second order Taylor expansion at $y = \ln 2$, which is

$$h(y + \delta) - h(y) = \frac{2}{3}\delta^2 + o(\delta^2)$$

As a result, (iii) holds and logistic regression under nonconvex sparsity regularization is strongly NP-hard.

6. In the mean estimation of inverse gaussian models [26], the negative log-likelihood function minimization is

$$\min_{x\in\mathbb{R}^d} -\log L(x;A,b) = \min_{x\in\mathbb{R}^d} \sum_{i=1}^n \frac{\lambda(b_i-\mu)^2}{2\mu^2 b_i} = \min_{x\in\mathbb{R}^d} \sum_{i=1}^n \frac{(b_i\cdot\sqrt{a_i^T x}-1)^2}{b_i}.$$

where λ is a parameter known before and $\mu = 1/\sqrt{a_i^T x}$ by the theory of generalized linear model [26]. We claim that the loss function $l(y,b) = \frac{(b\cdot\sqrt{y}-1)^2}{b}$ satisfies Assumption 1. By setting the derivative to be zero with regard to y, we can see that y take its minimum at $y = 1/b^2$. Choosing $b = 1$ and we can see that $h(y) = \ell(y,b')$ satisfies all the conditions in Assumption 1. To verify (iii), just note that the second order term of the Taylor expansion is $\frac{1}{4}\delta^2$.

7. In the estimation of generalized linear model under the exponential distribution [26], the negative log-likelihood function minimization is

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^n -\log(\mu\cdot\exp(-b_i\cdot\mu)) = \min_{x\in\mathbb{R}^d} \frac{b_i}{a_i^T x} + \log(a_i^T x).$$

We claim that the corresponding loss function $\ell(y,b) = \frac{b}{y} + \log y$ satisfies Assumption 1. By setting the derivative to 0 with regard to y, we can see that $\ell(y,b)$ has a unique minimizer at $y = b$. Thus by choosing $b = 1$, we can readily show that $h(y) = \ell(y,b)$ satisfies all the conditions in Assumption 1.

In summary, any combination of statistical model and L_0 penalty given in the preceding lists is strongly NP-hard. Although the message is somewhat discouraging, it contributes an important piece of complexity theory missing in the current literature.

5 Proof of Theorems 1 and 2

In this section, we prove our main results. The proof idea is to construct a polynomial-time reduction from the *3-partition problem* [19] to the sparse optimization problem. Given a set S of $3m$ integers $s_1, ...s_{3m}$, the three partition problem is to determine whether S can be partitioned into m triplets such that the sum of the numbers in each subset is equal. This problem is known to be strongly NP-hard [19].

We develop the following lemma. It relates the sparsity of a solution with the fact that a specific function of the solution is upper bounded. It relates the sparsity of a vector with a function which grows sufficiently fast around its minimal point. We use $B(\theta,\delta)$ to denote the interval $(\theta-\delta,\theta+\delta)$.

Lemma 1. *Let $f:\Re\mapsto\Re$ be a convex function with a unique minimizer t^* and satisfy $\frac{f(t^*+x)-f(t^*)}{|x|^N} \geq C$ for all $-\delta < x < \delta$, where $N\in\mathbb{Z}^+, \delta\in\mathbb{R}^+, C\in\mathbb{R}^+$. Then there exists $\mu > 0$ such that*

(i) For all $t_1,...,t_n\in\mathbb{R}$: $\sum_{j=1}^n \|t_j\|_0 + \mu\cdot f\left(\sum_{j=1}^n t_j\right) \geq 1 + \mu\cdot f(t^*)$.

(ii) Let $t_1,...,t_n\in\mathbb{R}$ satisfy

$$\sum_{j=1}^n \|t_j\|_0 + \mu\cdot f\left(\sum_{j=1}^n t_j\right) \leq \frac{3}{2} + \mu\cdot f(t^*), \tag{1}$$

then $t_i \in (t^* - \delta, t^* + \delta)$ for exactly one index i and $t_j = 0$ for all $j \neq i$.

Proof. Let $\mu \geq \max(\frac{2}{f(0)-f(t^*)}, \lceil \frac{1}{C\delta^N} \rceil)$. Note that such μ can be computed in polynomial time.
(i) Since $\mu \geq \frac{2}{f(0)-f(t^*)}$, we have $\mu \cdot f(0) - \mu \cdot f(t^*) > \frac{3}{2}$. If $t_i = 0$ for all i, we have $\sum_{j=1}^n \|t_j\|_0 + \mu \cdot f\left(\sum_{j=1}^n t_j\right) = \mu \cdot f(0) \geq 1 + \mu \cdot f(t^*)$. If there is at least one $t_i \neq 0$, then we have $\sum_{j=1}^n \|t_j\|_0 \geq 1$ and $f(\sum_{j=1}^n t_j) \geq f(t^*)$ so that the statement holds.
(ii) We claim that there exists at most one index i such that $t_i \neq 0$. Otherwise, we would have $\sum_{j=1}^n \|t_j\|_0 + \mu \cdot f\left(\sum_{j=1}^n t_j\right) \geq 2 + \mu \cdot f(t^*)$ which is a contradictory to (1). Moreover, since $\mu \geq \frac{2}{f(0)-f(t^*)}$ and $\mu \cdot f(0) - \mu \cdot f(t^*) > \frac{3}{2}$, it is not possible for all t_i's to equal to zero. Therefore, there exists exactly one index i such that $t_i \neq 0$. It remains to show that if $t \neq 0$ and $\|t\|_0 + \mu \cdot f(t) = 1 + \mu \cdot f(t) \leq \frac{3}{2} + \mu \cdot f(t^*)$, we have $t \in (t^* - \delta, t^* + \delta)$.

Suppose that $t \neq 0$ and $\|t\|_0 + \mu \cdot f(t) \leq \frac{3}{2} + \mu \cdot f(t^*)$, then we have $\mu \cdot f(t) - \mu \cdot f(t^*) \leq \frac{1}{2}$. Consider the case where $t \geq t^* + \delta$. We use the convexity of f and the fact $t^* = \text{argmin}_t f(t)$ to obtain

$$\mu \cdot f(t) - \mu \cdot f(t^*) \geq \mu \cdot f(t^* + \delta) - \mu \cdot f(t^*) \geq \frac{f(t^* + \delta) - f(t^*)}{C\delta^N} \geq 1,$$

where the second inequality uses $\mu \geq \lceil \frac{1}{C\delta^N} \rceil$, and the third inequality uses the property of f. Consider the case where $t \leq t^* - \delta$, we can show similarly that $\mu \cdot f(t) - \mu \cdot f(t^*) \geq 1$. In both cases, we obtain a contradiction. Therefore $t \in (t^* - \delta, t^* + \delta)$. □

Now we are ready to prove Theorem 1. The proof idea of Theorem 1 bears a similar spirit as the works by Huo et. al. [22], Chen et. at. [13], and Ge et. al. [20]. Our treatment of the general loss function f and the approximibility is novel.

Proof of Theorem 1. Suppose that we are given the input to the 3-partition problem, i.e., n positive integers $s_1, ..., s_n$. Assume *without loss of generality* that all s_i's are upper bounded by some polynomial function $M(n)$. This restriction on the input space does not weaken our result, because the 3-partition problem is strongly NP-hard.

In what follows, we construct a reduction from the 3-partition problem to Problem 1. For simplicity of notation, we let $n = 3m$. We also assume without loss of generality that $\frac{1}{4m}\sum_{j=1}^{3m} s_j < s_i < \frac{1}{2m}\sum_{j=1}^{3m} s_j$ for all $i = 1, \ldots, n$. Such condition can always be satisfied by adding a sufficiently large integer to all s_i's.

Step 1: The Reduction

The reduction is developed through the following steps.

1. Choose k and b_i's such that $h(y) = \sum_{i=1}^k \ell(y, b_i)$ satisfies Assumption 1 with constants C, N, δ_0. Let $y^* = \text{argmin}_y h(y)$. Let $\delta \leq \min\{\frac{y^*}{9 \cdot M(n)}, \delta_0\}$ and let $\mu \geq \max(\frac{2}{h(0)-h(y^*)}, \lceil \frac{1}{C\delta^N} \rceil)$ such that Lemma 1 is satisfied with function h, δ and μ.

2. Choose $\nu = \left\lceil \frac{1}{2C\delta^N} \right\rceil + 1$ where C and N are the constants defined in Assumption 1 and construct function $f : \mathbb{R}^{3m \times m} \mapsto \mathbb{R}$ where

$$f(x) = \lambda \cdot \sum_{i=1}^{3m}\sum_{j=1}^{m} \|x_{ij}\|_0 + \mu \sum_{i=1}^{3m} h\left(\sum_{j=1}^{m} x_{ij}\right) + \nu \sum_{j=1}^{m} h\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij}\right).$$

3. Let $\Phi_1 = 3m + \mu \cdot 3m \cdot h(y^*)$ and $\Phi_2 = \nu \cdot m \cdot h(y^*)$. We claim that

   (i) If there exists z such that

   $$\Phi_1 + \Phi_2 + \frac{1}{2} \geq \frac{1}{\lambda} f(z) \geq \Phi_1 + \Phi_2, \tag{2}$$

   then we obtain a feasible assignment for the 3-partition problem as follows: If $z_{ij} \in B(y^*, \delta)$, we assign number i to subset j.

   (ii) If the 3-partition problem has a solution, we have $\frac{1}{\lambda} \min_x f(x) = \Phi_1 + \Phi_2$.

4. Choose $d = \left\lceil (6m^2)^{1/(1-c)} \right\rceil$ where c is an arbitrary constant in $[0, 1)$. Choose $r = d/3m^2$. Construct the following instance of Problem 1:

$$\begin{aligned} \min_{x^{(1)},\cdots,x^{(r)} \in \mathbb{R}^{3m \times m}} \sum_{q=1}^{r} f(x^{(q)}) = \min_{x^{(1)},\cdots,x^{(r)} \in \mathbb{R}^{3m \times m}} \lambda \cdot \sum_{q=1}^{r} \sum_{i=1}^{3m} \sum_{j=1}^{m} \left\| x_{ij}^{(q)} \right\|_0 + \\ \mu \sum_{q=1}^{r} \sum_{i=1}^{3m} \sum_{t=1}^{k} \ell \left(\sum_{j=1}^{m} x_{ij}^{(q)}, b_t \right) + \nu \sum_{q=1}^{r} \sum_{j=1}^{m} \sum_{t=1}^{k} \ell \left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij}^{(q)}, b_t \right), \end{aligned} \tag{3}$$

   where the input data are coefficients of x and the values $b_1, \ldots, b_t$. In the regression setting, the variable dimension is d and the sample size is $\mu \cdot r \cdot 3m \cdot k + \nu \cdot r \cdot m \cdot k$. The input size is polynomial with respect to n.

   The parameters $\mu, \nu, \delta, m, r, d$ are bounded by polynomial functions of n. Computing their values also takes polynomial time. The parameter k is a constant determined by the loss function ℓ and is not related to n. As a result, the reduction is polynomial.

5. Let $z^{(1)}, \cdots z^{(r)} \in \mathbb{R}^{3m \times m}$ be a $\lambda \cdot d^c$-optimal solution to problem (3) such that $\sum_{i=1}^{r} f(z^{(i)}) \leq \min_{x^{(1)},\cdots,x^{(r)}} \sum_{i=1}^{r} f(x^{(i)}) + \lambda \cdot d^c$. We claim that

   (iii) If the approximate solution $z^{(1)}, \cdots z^{(r)}$ satisfies

   $$\frac{1}{\lambda} \sum_{i=1}^{r} f(z^{(i)}) \leq r\Phi_1 + r\Phi_2 + d^c, \tag{4}$$

   we can choose one $z^{(i)}$ such that $\Phi_1 + \Phi_2 + \frac{1}{2} \geq \frac{1}{\lambda} f(z^{(i)}) \geq \Phi_1 + \Phi_2$ and obtain a feasible assignment: If $z_{ij}^{(i)} \in B(y^*, \delta)$, we assign number i to subset j. If the $\lambda \cdot d^c$-optimal solution $z^{(1)}, \cdots z^{(r)}$ does not satisfy (4), the 3-partition problem has no feasible solution.

We have constructed a polynomial reduction from the 3-partition problem to finding an $\lambda \cdot d^c$-optimal solution to problem (3). In what follows, we prove that the reduction works.

Step 2: Proof of Claim (i)

We begin with the proof (i). By our choice of μ and Lemma 1(i), we can see that for all $x \in \mathbb{R}^{3m \times m}$,

$$\sum_{i=1}^{3m} \sum_{j=1}^{m} \|x_{ij}\|_0 + \mu \cdot \sum_{i=1}^{3m} h \left(\sum_{j=1}^{m} x_{ij} \right) \geq 3m + \mu \cdot 3m \cdot h(y^*) = \Phi_1,$$

By the fact $y^* = \text{argmin}_\theta h(\theta)$, we have for all $x \in \mathbb{R}^{3m \times m}$ that

$$\nu \cdot \sum_{j=1}^{m} h\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij}\right) \geq \nu \cdot m \cdot h(y^*) = \Phi_2.$$

Thus we always have $\min_z \frac{1}{\lambda} f(z) \geq \Phi_1 + \Phi_2$. Now if there exists z such that $\Phi_1 + \Phi_2 + \frac{1}{2} \geq \frac{1}{\lambda} f(z) \geq \Phi_1 + \Phi_2$, we must have

$$\Phi_1 + \frac{1}{2} \geq \sum_{i=1}^{3m} \sum_{j=1}^{m} \|z_{ij}\|_0 + \mu \cdot \sum_{i=1}^{3m} h\left(\sum_{j=1}^{m} z_{ij}\right) \geq \Phi_1, \tag{5}$$

and

$$\Phi_2 + \frac{1}{2} \geq \nu \cdot \sum_{j=1}^{m} h\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij}\right) \geq \Phi_2. \tag{6}$$

In order for equation (5) to hold, we have that for all i,

$$\mu \cdot h(y^*) + \frac{3}{2} \geq \sum_{j=1}^{m} \|z_{ij}\|_0 + \mu \cdot h\left(\sum_{j=1}^{m} z_{ij}\right) \geq \mu \cdot h(y^*) + 1.$$

Consider an arbitrary i. By Lemma 1(i), we have $z_{ij} \in B(y^*, \delta)$ for one j while $z_{ik} = 0$ for all $k \neq j$. If $z_{ij} \in B(y^*, \delta)$, we assign number i to subset j. Each number index i is assigned to exactly one subset index j. Therefore the assignment is feasible.

We claim that every subset sum must equal to $\sum_{i=1}^{3m} s_i/m$. Assume that the jth subset sum is greater than or equal to $\sum_{i=1}^{3m} s_i/m + 1$. Let $I_j = \{i \mid z_{ij} \in B(y^*, \delta)\}$. Thus, $\sum_{i \in I_j} s_i \geq \sum_{i=1}^{3m} s_i/m + 1$. As a result, we have

$$\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij} \geq \sum_{i \in I_1} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} (y^* - \delta) \geq \frac{\sum_{i=1}^{3m} s_i/m + 1}{\sum_{i=1}^{3m} s_i/m} (y^* - \delta) \geq y^* + \frac{y^*}{\sum_{i=1}^{3m} s_i/m} - 2\delta.$$

Because $s_i \leq M(n)$ for all i and $\delta \leq \frac{y^*}{9M(n)}$, we have

$$\frac{y^*}{\sum_{i=1}^{3m} s_i/m} - 2\delta \geq \frac{y^*}{3M(n)} - 2\delta = \delta > 0.$$

Since h is a convex function with minimizer y^*, we apply the preceding inequalities and further obtain

$$h\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij}\right) \geq h(y^* + \delta).$$

By the construction $\nu = \left\lceil \frac{1}{2C\delta^N} \right\rceil + 1$ and Assumption 1(iii), we further have

$$\nu \cdot \left(h\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij}\right) - h(y^*)\right) \geq \nu \cdot (h(y^* + \delta) - h(y^*)) > \frac{1}{2}. \tag{7}$$

However, in order for equation (6) to hold, we have that for all j,

$$\nu \cdot h(y^*) + \frac{1}{2} \geq \nu \cdot h\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij}\right) \geq \nu \cdot h(y^*),$$

yielding a contradiction to (7). We could prove similarly that it is not possible for any subset sum to be strictly smaller than $\frac{1}{m}\sum_{i=1}^{3m} s_i$. Therefore, the sum of every subset equals to $\sum_{i=1}^{3m} s_i/m$. Finally, using the assumption that $\frac{1}{4m}\sum_{i=1}^{3m} s_i < s_i < \frac{1}{2m}\sum_{i=1}^{3m} s_i$, each subset has exactly three components. Therefore the assignment is indeed a solution to the 3-partition problem.

Step 3: Proof of Claim (ii)

Suppose we have a solution to the 3-partition problem. Now we construct a solution z to the optimization problem. For all $1 \leq i \leq 3m$, if number i is assigned to subset j, let $z_{ij} = y^*$ and $z_{ik} = 0$ for all $k \neq j$. We can easily verify that

$$\frac{1}{\lambda} f(z) = \sum_{i=1}^{3m}\sum_{j=1}^{m} \|z_{ij}\| + \mu \cdot \sum_{i=1}^{3m} h\left(\sum_{j=1}^{m} z_{ij}\right) + \nu \cdot \sum_{j=1}^{m} h\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij}\right) = \Phi_1 + \Phi_2, \quad (8)$$

which completes the proof of (ii).

Step 4: Proof of Claim (iii)

Suppose that the $\lambda \cdot d^c$-optimal solution satisfies (4), i.e., $\frac{1}{\lambda}\sum_{i=1}^{r} f(z^{(i)}) \leq r\Phi_1 + r\Phi_2 + d^c$. It follows that there exists at least one term $z^{(i)}$ such that

$$\frac{1}{\lambda} f(z^{(i)}) \leq \Phi_1 + \Phi_2 + \frac{1}{r} d^c \leq \Phi_1 + \Phi_2 + 1/2. \quad (9)$$

where the second inequality equality uses $r = d/3m^2$ and $d = \left\lceil (6m^2)^{1/(1-c)} \right\rceil$. Therefore, by claim (i), we can find a solution to the 3-partition problem.

Suppose that the 3-partition problem has a solution. By claim (ii), there exists $\bar{x}$ such that $\frac{1}{\lambda} f(\bar{x}) = \Phi_1 + \Phi_2$. Thus we have

$$\min_{x^{(1)},\cdots,x^{(r)}} \frac{1}{\lambda}\sum_{i=1}^{r} f(x^{(i)}) \leq \frac{r}{\lambda} f(\bar{x}) = r\Phi_1 + r\Phi_2. \quad (10)$$

Thus if $z^{(1)}, \cdots z^{(r)}$ is a $\lambda \cdot d^c$-optimal solution to (3), the relation (4) must hold. If (4) is not satisfied, the 3-partition problem has no solution. □

Next we study the complexity of Problem 2. The proof uses a basic duality between Problem 1 and Problem 2.

Proof of Theorem 2. We will use a reduction from Problem 1 to prove the theorem. For the simplicity of notation, we use $f(x)$ to denote $f(x) = \sum_{i=1}^{n} \ell\left(a_i^T x, b_i\right)$. Let $r = \lfloor d^{c/(1-c)} \rfloor$. Given the input of Problem 1, we consider the following instance of Problem 2

$$\min_{x^{(1)},\ldots,x^{(r)} \in \mathbb{R}^d} \sum_{i=1}^{r} f(x^{(i)}) \quad \text{s.t.} \quad \sum_{i=1}^{r} \left\|x^{(i)}\right\|_0 \leq K. \quad (11)$$

We claim that

$$\min_{x^{(1)},\ldots,x^{(r)}\in\mathbb{R}^d}\left\{\sum_{i=1}^r f(x^{(i)})+\lambda\sum_{i=1}^r\left\|x^{(i)}\right\|_0\right\}=\min_{K\in\{0,r,2r,\ldots,dr\}}\left\{\sum_{i=1}^r f(x_K^{(i)})+\lambda\left\|x_K^{(i)}\right\|_0\right\},\qquad(12)$$

where $\{x_K^{(1)},\ldots,x_K^{(r)}\}\in\operatorname{argmin}\{\sum_{i=1}^r f(x^{(i)})\mid\sum_{i=1}^r\left\|x^{(i)}\right\|_0\le K\}$. It is easy to see that the left side of (12) is smaller than or equal to the right side. To show the other direction, we denote the optimal solution of $\min_{x\in\mathbb{R}^d} f(x)+\lambda\|x\|_0$ to be x^* and let $\bar{K}=\|x^*\|_0$. By the definition of $x_K^{(i)}$, we have $\sum_{i=1}^r\left\|x_{r\bar{K}}^{(i)}\right\|_0\le r\bar{K}=r\|x^*\|_0$ and $\sum_{i=1}^r f(x_{r\bar{K}}^{(i)})\le\sum_{i=1}^r f(x^*)$. Thus we have $\sum_{i=1}^r f(x_{r\bar{K}}^{(i)})+\lambda\sum_{i=1}^r\left\|x_{r\bar{K}}^{(i)}\right\|_0\le\sum_{i=1}^r f(x^*)+\lambda\sum_{i=1}^r\|x^*\|_0$, which means that the right side of (12) is smaller than or equal to the left side. Thus we have proved (12) and that $\sum_{i=1}^r f(x^*)=\sum_{i=1}^r f(x_{r\bar{K}}^{(i)})$.

Assume to the contrary that we have a pseudo polynomial-time algorithm which takes the input $(a_i,b_i)_{i=1}^n$ and outputs $(\hat{x}_1,\ldots,\hat{x}_d)$ satisfying

$$\sum_{i=1}^r f(\hat{x}_{K+r}^{(i)})\le\sum_{i=1}^r f(x_K^{(i)}),\qquad\forall\ K=0,r,\ldots,dr.$$

Replacing K by $r\bar{K}$, we get

$$\begin{aligned}\sum_{i=1}^r f(\hat{x}_{r\bar{K}+r}^{(i)})+\lambda\sum_{i=1}^r\left\|\hat{x}_{r\bar{K}+r}^{(i)}\right\|_0&\le\sum_{i=1}^r f(x_{r\bar{K}}^{(i)})+\lambda\sum_{i=1}^r\left\|x_{r\bar{K}}^{(i)}\right\|_0+\lambda r\\&=\min_{x^{(1)},\ldots,x^{(r)}\in\mathbb{R}^d}\left\{\sum_{i=1}^r f(x^{(i)})+\lambda\sum_{i=1}^r\left\|x^{(i)}\right\|_0\right\}+\lambda r,\end{aligned}\qquad(13)$$

where the first inequality uses the facts $\sum_{i=1}^r f(\hat{x}_{K+r}^{(i)})\le\sum_{i=1}^r f(x_K^{(i)})=\sum_{i=1}^r f(x^*)$ and $\sum_{i=1}^r\left\|\hat{x}_{r\bar{K}+r}^{(i)}\right\|_0\le r\bar{K}+r=r\|x^*\|_0+r$. By solving $d+1$ instances of Problem 2, we can choose one out of the $d+1$ solutions $(\hat{x}_0,\ldots,\hat{x}_{dr})$ with the smallest penalized objective, such that

$$\min_{K\in\{0,r,\ldots,dr\}}\left\{\sum_{i=1}^r f(\hat{x}_K^{(i)})+\lambda\sum_{i=1}^r\left\|\hat{x}_K^{(i)}\right\|_0\right\}\le\min_{x^{(1)},\ldots,x^{(r)}\in\mathbb{R}^d}\left\{\sum_{i=1}^r f(x^{(i)})+\lambda\sum_{i=1}^r\left\|x^{(i)}\right\|_0\right\}+\lambda r$$

This yields a λ-optimal solution to Problem 1, leading to a contradiction with Theorem 1. Note that the dimension of (11) is $d'=d\cdot r=d^{1/(1-c)}$ and thus $r=d'^c$. It follows that there does not exist a pseudo polynomial-time algorithm that outputs x_K such that $\ell_n(\hat{x}_K+d^c)\le\ell_n(x_K^*)$. □

6 Conclusion

We have studied the worst-case computation complexity for L_0-regularized/constrained optimization. In contrast to existing results on proving NP-hardness, we focus on the approximation hardness. Although it is broadly believed that optimization problems involving L_0 norm are intractable, this is the first work on its approximibility and approximation error.

For future research, one direction is to relax the L_0 norm to other nonconvex norms and study the corresponding approximation error. Another direction is to study the average-case computational complexity instead of the worst-case complexity, and to study randomized algorithm and/or randomized input. It remains an open problem whether or when the sparse optimization is hard on average.

References

[1] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[2] Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1):237–260, 1998.

[3] Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455), 2001.

[4] Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedy. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 724–733. IEEE, 1993.

[5] Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *The Journal of Machine Learning Research*, 14(1):807–841, 2013.

[6] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[7] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *arXiv preprint arXiv:1507.03133*, 2015.

[8] Wei Bian and Xiaojun Chen. Optimality conditions and complexity for non-lipschitz constrained optimization problems. *Preprint*, 2014.

[9] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[10] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[11] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[12] Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *Signal Processing Letters, IEEE*, 14(10):707–710, 2007.

[13] Xiaojun Chen, Dongdong Ge, Zizhuo Wang, and Yinyu Ye. Complexity of unconstrained l_2-l_p minimization. *Mathematical Programming*, 143(1-2):371–383, 2014.

[14] Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[15] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[16] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

[17] Dean Foster, Howard Karloff, and Justin Thaler. Variable selection is hard. In *COLT*, pages 696–709, 2015.

[18] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[19] Michael R Garey and David S Johnson. "strong"np-completeness results: Motivation, examples, and implications. *Journal of the ACM (JACM)*, 25(3):499–508, 1978.

[20] Dongdong Ge, Zizhuo Wang, Yinyu Ye, and Hao Yin. Strong NP-hardness result for regularized l_q-minimization problems with concave penalty functions. *arXiv preprint arXiv:1501.00622*, 2015.

[21] Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[22] Xiaoming Huo and Jie Chen. Complexity of penalized likelihood estimation. *Journal of Statistical Computation and Simulation*, 80(7):747–759, 2010.

[23] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

[24] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. In *Annales de l'IHP Probabilités et statistiques*, volume 45, pages 7–57, 2009.

[25] Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[26] Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

[27] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[28] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[29] Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

[30] Ambuj Tewari, Pradeep K Ravikumar, and Inderjit S Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In *Advances in Neural Information Processing Systems*, pages 882–890, 2011.

[31] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[32] Vijay V Vazirani. *Approximation algorithms*. Springer Science & Business Media, 2001.

[33] Lingzhou Xue, Hui Zou, Tianxi Cai, et al. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429, 2012.

[34] Xiaotong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *ICML*, pages 127–135, 2014.

[35] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems*, pages 1921–1928, 2009.

[36] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *COLT*, 2014.