Strong NP-Hardness for Sparse Optimization with Concave Penalty Functions

Yichen Chen[1] Dongdong Ge[2] Mengdi Wang[1] Zizhuo Wang[3] Yinyu Ye[4] Hao Yin[4]

Abstract

Consider the regularized sparse minimization problem, which involves empirical sums of loss functions for n data points (each of dimension d) and a nonconvex sparsity penalty. We prove that finding an $\mathcal{O}(n^{c_1}d^{c_2})$-optimal solution to the regularized sparse optimization problem is strongly NP-hard for any $c_1, c_2 \in [0,1)$ such that $c_1 + c_2 < 1$. The result applies to a broad class of loss functions and sparse penalty functions. It suggests that one cannot even approximately solve the sparse optimization problem in polynomial time, unless P $=$ NP.

Keywords: Nonconvex optimization · Computational complexity · NP-hardness · Concave penalty · Sparsity

1 Introduction

We study the sparse minimization problem, where the objective is the sum of empirical losses over input data and a sparse penalty function. Such problems commonly arise from empirical risk minimization and variable selection. The role of the penalty function is to induce sparsity in the optimal solution, i.e., to minimize the empirical loss using as few nonzero coefficients as possible.

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

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

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 are interested in the computational complexity of Problem 1 under general conditions of the loss function ℓ and the sparse penalty p. In particular, we focus on the case where ℓ is a convex loss function and p is a concave penalty with a unique minimizer at 0. Optimization problems with convex ℓ and concave p are common in sparse regression, compressive sensing, and sparse approximation. A list of applicable examples of ℓ and p is given in Section 3.

For certain special cases of Problem 1, it has been shown that finding an *exact solution* is strongly NP-hard (Huo & Chen, 2010; Chen et al., 2014). However, these results have not excluded the possibility of the existence of polynomial-time algorithms with small approximation error. (Chen & Wang, 2016) established the hardness of approximately solving Problem 1 when p is the L_0 norm.

In this paper, we prove that it is strongly NP-hard to approximately solve Problem 1 within certain optimality error. More precisely, we show that there exists a lower bound on the suboptimality error of any polynomial-time deterministic algorithm. Our results apply to a variety of optimization problems in estimation and machine learning. Examples include sparse classification, sparse logistic regression, and many more. The strong NP-hardness of approximation is one of the strongest forms of complexity result for continuous optimization. To our best knowledge, this paper gives the first and strongest set of hardness results for Problem 1 under very general assumptions regarding the loss and penalty functions.

Our main contributions are three-fold.

1. We prove the strong NP-hardness for Problem 1 with general loss functions. This is the first results that apply to the broad class of problems including but not limited to: least squares regression, linear model with Laplacian noise, robust regression, Poisson regression, logistic regression, inverse Gaussian models, etc.

2. We present a general condition on the sparse penalty function p such that Problem 1 is strongly NP-hard. The condition is a slight weaker version of strict concavity. It is satisfied by typical penalty functions such as the L_q norm ($q \in [0,1)$), clipped L_1 norm, SCAD, etc. To the best of our knowledge, this is the most gen-

[1] Princeton University, NJ, USA [2] Shanghai University of Finance and Economics, Shanghai, China [3] University of Minnesota, MN, USA [4] Stanford University, CA, USA. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

eral condition on the penalty function in the literature.

3. We prove that finding an $\mathcal{O}\left(\lambda n^{c_1} d^{c_2}\right)$-optimal solution to Problem 1 is strongly NP-hard, for any $c_1, c_2 \in [0, 1)$ such that $c_1 + c_2 < 1$. Here the $\mathcal{O}(\cdot)$ hides parameters that depend on the penalty function p, which is to be specified later. It illustrates a gap between the optimization error achieved by any tractable algorithm and the desired statistical precision. Our proof provides a first unified analysis that deals with a broad class of problems taking the form of Problem 1.

Section 2 summarizes related literatures from optimization, machine learning and statistics. Section 3 presents the key assumptions and illustrates examples of loss and penalty functions that satisfy the assumptions. Section 4 gives the main results. Section 5 discusses the implications of our hardness results. Section 6 provides a proof of the main results in a simplified setting. The full proofs are deferred to the appendix.

2 Background and Related Works

Sparse optimization is a powerful machine learning tool for extracting useful information for massive data. In Problem 1, the sparse penalty serves to select the most relevant variables from a large number of variables, in order to avoid overfitting. In recent years, nonconvex choices of p have received much attention; see (Frank & Friedman, 1993; Fan & Li, 2001; Chartrand, 2007; Candes et al., 2008; Fan & Lv, 2010; Xue et al., 2012; Loh & Wainwright, 2013; Wang et al., 2014; Fan et al., 2015).

Within the optimization and mathematical programming community, the complexity of Problem 1 has been considered in a number of special cases. (Huo & Chen, 2010) first proved the hardness result for a relaxed family of penalty functions with L_2 loss. They show that for the penalties in L_0, hard-thresholded (Antoniadis & Fan, 2001) and SCAD (Fan & Li, 2001), the above optimization problem is NP-hard. (Chen et al., 2014) showed that the L_2-L_p minimization is strongly NP-hard when $p \in (0, 1)$. At the same time, (Bian & Chen, 2014) proved the strongly NP-hardness for another class of penalty functions. The preceding existing analyses mainly focused on finding an exact global optimum to Problem 1. For this purpose, they implicitly assumed that all the input and parameters involved in the reduction are rational numbers with a finite numerical representation, otherwise finding a global optimum to a continuous problem would be always intractable. A recent technical report (Chen & Wang, 2016) proves the hardness of obtaining an ϵ-optimal solution when p is the L_0 norm.

Within the theoretical computer science community, there have been several early works on the complexity of sparse recovery, beginning with (Arora et al., 1993). (Amaldi & Kann, 1998) proved that the problem $\min\{\|x\|_0 \mid Ax = b\}$ is not approximable within a factor $2^{\log^{1-\epsilon} d}$ for any $\epsilon > 0$. (Natarajan, 1995) showed that, given $\epsilon > 0, A$ and b, the problem $\min\{\|x\|_0 \mid \|Ax - b\|_2 \le \epsilon\}$ is NP-hard. (Davis et al., 1997) proved a similar result that for some given $\epsilon > 0$ and $M > 0$, it is NP-complete to find a solution x such that $\|x\|_0 \le M$ and $\|Ax - b\| \le \epsilon$. More recently, (Foster et al., 2015) studied 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 $\|Ax - b\|_2^2 \le d^{C_1} n^{1-C_2}$ *with high probability*, where δ, C_1, C_2 are arbitrary positive scalars. Another work (Zhang et al., 2014) showed that under the 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 two works focus on estimation of linear models and impose distributional assumptions regarding the input data. These results on estimation are different in nature with our results on optimization.

In contrast, we focus on the optimization problem itself. Our results apply to a variety of loss functions and penalty functions, not limited to linear regression. Moreover, we do not make any distributional assumption regarding the input data.

There remain several open questions. First, existing results mainly considered least square problems or L_q minimization problems. Second, existing results focused mainly on the L_0 penalty function. The complexity of Problem 1 with general loss function and penalty function is yet to be established. Things get complicated when p is a continuous function instead of the discrete L_0 norm function. The complexity for finding an ϵ-optimal solution with general ℓ and p is not fully understood. We will address these questions in this paper.

3 Assumptions

In this section, we state the two critical assumptions that lead to the strong NP-hardness results: one for the penalty function p, the other one for the loss function ℓ. We argue that these assumptions are essential and very general. They apply to a broad class of loss functions and penalty functions that are commonly used.

3.1 Assumption About Sparse Penalty

Throughout this paper, we make the following assumption regarding the sparse penalty function $p(\cdot)$.

Assumption 1. *The penalty function $p(\cdot)$ satisfies the fol-*

lowing conditions:

(i) (Monotonicity) $p(\cdot)$ is non-decreasing on $[0, +\infty)$.

(ii) (Concavity) There exists $\tau > 0$ such that $p(\cdot)$ is concave but not linear on $[0, \tau]$.

In words, condition (ii) means that the concave penalty $p(\cdot)$ is nonlinear. Assumption 1 is the most general condition on penalty functions in the existing literature of sparse optimization. Below we present a few such examples.

1. In variable selection problems, the L_0 penalization $p(t) = I_{\{t \neq 0\}}$ arises naturally as a penalty for the number of factors selected.

2. A natural generalization of the L_0 penalization is the L_p penalization $p(t) = t^p$ where $(0 < p < 1)$. The corresponding minimization problem is called the bridge regression problem (Frank & Friedman, 1993).

3. To obtain a hard-thresholding estimator, Antoniadis & Fan (2001) use the penalty functions $p_\gamma(t) = \gamma^2 - ((\gamma - t)^+)^2$ with $\gamma > 0$, where $(x)^+ := \max\{x, 0\}$ denotes the positive part of x.

4. Any penalty function that belongs to the folded concave penalty family (Fan et al., 2014) satisfies the conditions in Theorem 1. Examples include the SCAD (Fan & Li, 2001) and the MCP (Zhang, 2010a), whose derivatives on $(0, +\infty)$ are $p'_\gamma(t) = \gamma I_{\{t \leq \gamma\}} + \frac{(a\gamma - t)^+}{a-1} I_{\{t > \gamma\}}$ and $p'_\gamma(t) = (\gamma - \frac{t}{b})^+$, respectively, where $\gamma > 0$, $a > 2$ and $b > 1$.

5. The conditions in Theorem 1 are also satisfied by the clipped L_1 penalty function (Antoniadis & Fan, 2001; Zhang, 2010b) $p_\gamma(t) = \gamma \cdot \min(t, \gamma)$ with $\gamma > 0$. This is a special case of the piecewise linear penalty function:

$$p(t) = \begin{cases} k_1 t & \text{if } 0 \leq t \leq a \\ k_2 t + (k_1 - k_2)a & \text{if } t > a \end{cases}$$

   where $0 \leq k_2 < k_1$ and $a > 0$.

6. Another family of penalty functions which bridges the L_0 and L_1 penalties are the fraction penalty functions $p_\gamma(t) = \dfrac{(\gamma+1)t}{\gamma + t}$ with $\gamma > 0$ (Lv & Fan, 2009).

7. The family of log-penalty functions:

$$p_\gamma(t) = \frac{1}{\log(1+\gamma)} \log(1 + \gamma t)$$

   with $\gamma > 0$, also bridges the L_0 and L_1 penalties (Candes et al., 2008).

3.2 Assumption About Loss Function

We state our assumption about the loss function ℓ.

Assumption 2. *Let M be an arbitrary constant. For any interval $[\tau_1, \tau_2]$ where $0 < \tau_1 < \tau_2 < M$, 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 and Lipschitz continuous on $[\tau_1, \tau_2]$.

(ii) $h(y)$ has a unique minimizer y^* in (τ_1, τ_2).

(iii) There exists $N \in \mathbb{Z}^+, \bar{\delta} \in \mathbb{Q}^+$ and $C \in \mathbb{Q}^+$ such that when $\delta \in (0, \bar{\delta})$, we have

$$\frac{h(y^* \pm \delta) - h(y^*)}{\delta^N} \geq C.$$

(iv) $h(y^*)$, $\{b_i\}_{i=1}^k$ can be represented in $\mathcal{O}(\log \frac{1}{\tau_2 - \tau_1})$ bits.

Assumption 2 is a critical, but very general, assumption regarding the loss function $\ell(y, b)$. Condition (i) requires convexity and Lipschitz continuity within a neighborhood. Conditions (ii), (iii) essentially require that, given an interval $[\tau_1, \tau_2]$, one can artificially pick $b_1, \ldots, b_k$ to construct a function $h(y) = \sum_{i=1}^k \ell(y, b_i)$ such that h has its unique minimizer in $[\tau_1, \tau_2]$ and has enough curvature near the minimizer. This property ensures that a bound on the minimal value of $h(y)$ can be translated to a meaningful bound on the minimizer y^*. The conditions (i), (ii), (iii) are typical properties that a loss function usually satisfies. Condition (iv) is a technical condition that is used to avoid dealing with infinitely-long irrational numbers. It can be easily verified for almost all common loss functions.

We will show that Assumptions 2 is satisfied by a variety of loss functions. An (incomplete) list is given below.

1. In the least squares regression, the loss function has the form
$$\sum_{i=1}^n \left(a_i^T x - b_i\right)^2.$$
   Using our notation, the corresponding loss function is $\ell(y, b) = (y - b)^2$. For all τ_1, τ_2, we choose an arbitrary $b' \in [\tau_1, \tau_2]$. We can verify that $h(y) = \ell(y, b')$ satisfies all the conditions in Assumption 2.

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

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

   So the loss function is $\ell(y, b) = |y - b|$. As in the case of least squares regression, the loss function satisfy

Assumption 2. Similar argument also holds when we consider the L_q loss $|\cdot|^q$ with $q \geq 1$.

3. In robust regression, we consider the Huber loss (Huber, 1964) which is a mixture of L_1 and L_2 norms. The loss function takes the form

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

   for some $\delta > 0$ where $y = a^T x$. We then verify that Assumption 2 is satisfied. For any interval $[\tau_1, \tau_2]$, we pick an arbitrary $b \in [\tau_1, \tau_2]$ and let $h(y) = \ell(y, b)$. We can see that $h(y)$ satisfies all the conditions in Assumption 2.

4. In Poisson regression (Cameron & Trivedi, 2013), the negative log-likelihood minimization is

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

   We now show that $\ell(y, b) = e^y - b \cdot y$ satisfies Assumption 2. For any interval $[\tau_1, \tau_2]$, we choose q and r such that $q/r \in [e^{\tau_1}, e^{\tau_2}]$. Note that $e^{\tau_2} - e^{\tau_1} = e^{\tau_1 + \tau_2 - \tau_1} - e^{\tau_1} \geq \tau_2 - \tau_1$. Also, e^{τ_2} is bounded by e^M. Thus, q, r can be chosen to be polynomial in $\lceil 1/(\tau_2 - \tau_1) \rceil$ by letting $r = \lceil 1/(\tau_2 - \tau_1) \rceil$ and q be some number less than $r \cdot e^M$. Then, we choose $k = r$ and $b \in \mathbb{Z}^k$ such that $h(y) = \sum_{i=1}^k \ell(y, b_i) = r \cdot e^y - q \cdot y$. Let us verify Assumption 2. (i), (iv) are straightforward by our construction. For (ii), note that $h(y)$ take its minimum at $\ln(q/r)$ which is inside $[\tau_1, \tau_2]$ by our construction. To verify (iii), consider the second order Taylor expansion of $h(y)$ at $\ln(q/r)$,

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

   We can see that (iii) is satisfied. Therefore, Assumption 2 is satisfied.

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 2. By a similar argument as the one in Poisson regression, we can verify that $h(y) = \sum_{i=1}^r \ell(y, b_i) = r \log(1 + \exp(y)) - qy$ where $q/r \in [\frac{e^{\tau_1}}{1+e^{\tau_1}}, \frac{e^{\tau_2}}{1+e^{\tau_2}}]$ and q, r are polynomial in $\lceil 1/(\tau_2 - \tau_1) \rceil$ satisfies all the conditions in Assumption 2. For (ii), observe that $\ell(y, b)$ take its minimum at $y = \ln \frac{q/r}{1-q/r}$. To verify (iii), we consider the second order Taylor expansion at $y = \ln \frac{q/r}{1-q/r}$, which is

$$h(y + \delta) - h(y) = \frac{q}{2(1 + e^y)} \delta^2 + o(\delta^2)$$

   where $y \in [\tau_1, \tau_2]$. Note that e^y is bounded by e^M, which can be computed beforehand. As a result, (iii) holds as well.

6. In the mean estimation of inverse Gaussian models (McCullagh, 1984), the negative log-likelihood function minimization is

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^n \frac{(b_i \cdot \sqrt{a_i^T x} - 1)^2}{b_i}.$$

   Now we show that the loss function $\ell(y, b) = \frac{(b \cdot \sqrt{y} - 1)^2}{b}$ satisfies Assumption 2. By setting the derivative to be zero with regard to y, we can see that y take its minimum at $y = 1/b^2$. Thus for any $[\tau_1, \tau_2]$, we choose $b' = q/r \in [1/\sqrt{\tau_2}, 1/\sqrt{\tau_1}]$. We can see that $h(y) = \ell(y, b')$ satisfies all the conditions in Assumption 2.

7. In the estimation of generalized linear model under the exponential distribution (McCullagh, 1984), the negative log-likelihood function minimization is

$$\min_{x \in \mathbb{R}^d} -\log L(x; A, b) = \min_{x \in \mathbb{R}^d} \frac{b_i}{a_i^T x} + \log(a_i^T x).$$

   By setting the derivative to 0 with regard to y, we can see that $\ell(y, b) = \frac{b}{y} + \log y$ has a unique minimizer at $y = b$. Thus by choosing $b' \in [\tau_1, \tau_2]$ appropriately, we can readily show that $h(y) = \ell(y, b')$ satisfies all the conditions in Assumption 2.

To sum up, the combination of *any* loss function given in Section 3.1 and *any* penalty function given in Section 3.2 results in a strongly NP-hard optimization problem.

4 Main Results

In this section, we state our main results on the strong NP-hardness of Problem 1. We warm up with a preliminary result for a special case of Problem 1.

Theorem 1 (A Preliminary Result)**.** *Let Assumption 1 hold, and let $p(\cdot)$ be twice continuously differentiable in $(0, \infty)$. Then the minimization problem*

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_q^q + \lambda \sum_{j=1}^d p(|x_j|), \tag{1}$$

is strongly NP-hard.

The result shows that many of the penalized least squares problems, e.g., (Fan & Lv, 2010), while enjoying small estimation errors, are hard to compute. It suggests that there does not exist a fully polynomial-time approximation scheme for Problem 1. It has not answered the question: whether one can approximately solve Problem 1 within certain constant error.

Now we show that it is not even possible to efficiently approximate the global optimal solution of Problem 1, unless $P = NP$. 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 $f(\bar{x}) \leq \inf_{x \in X} f(x) + \epsilon$.

Theorem 2 (Strong NP-Hardness of Problem 1)**.** *Let Assumptions 1 and 2 hold, and let $c_1, c_2 \in [0, 1)$ be arbitrary such that $c_1 + c_2 < 1$. Then it is strongly NP-hard to find a $\lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$-optimal solution of Problem 1, where d is the dimension of variable space and $\kappa = \min_{t \in [\tau/2, \tau]} \{\frac{2p(t/2) - p(t)}{t}\}$.*

The non-approximable error in Theorem 2 involves the constant κ which is determined by the sparse penalty function p. In the case where p is the L_0 norm function, we can take $\kappa = 1$. In the case of piecewise linear L_1 penalty, we have $\kappa = (k_1 - k_2)/4$. In the case of SCAD penalty, we have $\kappa = \Theta(\gamma^2)$.

According to Theorem 2, the non-approximable error $\lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$ is determined by three factors: (i) properties of the regularization penalty $\lambda \cdot \kappa$; (ii) data size n; and (iii) dimension or number of variables d. This result illustrates a fundamental gap that can not be closed by any polynomial-time deterministic algorithm. This gap scales up when either the data size or the number of variables increases. In Section 5.1, we will see that this gap is substantially larger than the desired estimation precision in a special case of sparse linear regression.

Theorems 1 and 2 validate the long-lasting belief that optimization involving nonconvex penalty is hard. More importantly, Theorem 2 provide lower bounds for 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.

5 An Application and Remarks

In this section, we analyze the strong NP-hardness results in the special case of linear regression with SCAD penalty (Problem 1). We give a few remarks on the implication of our hardness results.

5.1 Hardness of Regression with SCAD Penalty

Let us try to understand how significant is the non-approximable error of Problem 1. We consider the special case of linear models with SCAD penalty. 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 follow (Fan & Li, 2001) and obtain a special case of Problem 1, given by

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + n \sum_{j=1}^{d} p_\gamma(|x_j|), \tag{2}$$

where $\gamma = \sqrt{\log d/n}$. (Fan & Li, 2001) showed that the optimal solution x^* of problem (2) has a small statistical error, i.e., $\|\bar{x} - x^*\|_2^2 = \mathcal{O}\left(n^{-1/2} + a_n\right)$, where $a_n = \max\{p'_\lambda(|x_j^*|) : x_j^* \neq 0\}$. (Fan et al., 2015) further showed that we only need to find a $\sqrt{n \log d}$-optimal solution to (2) to achieve such a small estimation error.

However, Theorem 2 tells us that it is not possible to compute an $\epsilon_{d,n}$-optimal solution for problem (2) in polynomial time, where $\epsilon_{d,n} = \lambda\kappa n^{1/2} d^{1/3}$ (by letting $c_1 = 1/2, c_2 = 1/3$). In the special case of problem (2), we can verify that $\lambda = n$ and $\kappa = \Omega(\gamma^2) = \Omega(\log d/n)$. As a result, we see that

$$\epsilon_{d,n} = \Omega(n^{1/2} d^{1/3}) \gg \sqrt{n \log d},$$

for high values of the dimension d. According to Theorem 2, it is strongly NP-hard to approximately solve problem (2) within the required statistical precision $\sqrt{n \log d}$. This result illustrates a sharp contrast between statistical properties of sparse estimation and the worst-case computational complexity.

5.2 Remarks on the NP-Hardness Results

As illustrated by the preceding analysis, the non-approximibility of Problem 1 suggests that computing the sparse estimator is 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 make the following remarks:

1. Theorems 1, 2 are *worst-case* complexity results. They suggest that one cannot find a tractable solution to the sparse optimization problems, without making any additional assumption to rule out the worst-case instances.

2. Our results do not exclude the possibility that, under more stringent modeling and distributional assump-

tions, the problem would be tractable with high probability or on average.

In short, the sparse optimization Problem 1 is fundamentally hard from a purely computational perspective. This paper together with the prior related works provide a complete answer to the computational complexity of sparse optimization.

6 Proof of Theorem 1

In this section, we prove Theorem 1. The proof of Theorems 2 is deferred to the appendix which is based on the idea of the proof in this section. We construct a polynomial-time reduction from the *3-partition problem* (Garey & Johnson, 1978) 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 (Garey & Johnson, 1978). The main proof idea bears a similar spirit as the works by Huo & Chen (2010), Chen et al. (2014) and Chen & Wang (2016). The proofs of all the lemmas can be found in the appendix.

We first illustrate several properties of the penalty function if it satisfies the conditions in Theorem 1.

Lemma 3. *If $p(t)$ satisfies the conditions in Theorem 1, then for any $l \geq 2$, and any $t_1, t_2, \dots, t_l \in \mathbb{R}$, we have $p(|t_1|) + \cdots + p(|t_l|) \geq \min\{p(|t_1 + \cdots + t_l|), p(\tau)\}$.*

Lemma 4. *If $p(t)$ satisfies the conditions in Theorem 1, then there exists $\tau_0 \in (0, \tau)$ such that $p(\cdot)$ is concave but not linear on $[0, \tau_0]$ and is twice continuously differentiable on $[\tau_0, \tau]$. Furthermore, for any $\tilde{t} \in (\tau_0, \tau)$, let $\bar{\delta} = \min\{\tau_0/3, \tilde{t} - \tau_0, \tau - \tilde{t}\}$. Then for any $\delta \in (0, \bar{\delta})$ $l \geq 2$, and any $t_1, t_2, \dots, t_l$ such that $t_1 + \cdots + t_l = \tilde{t}$, we have*

$$p(|t_1|) + \cdots + p(|t_l|) < p(\tilde{t}) + C_1\delta$$

only if $|t_i - \tilde{t}| < \delta$ for some i while $|t_j| < \delta$ for all $j \neq i$, where $C_1 = \frac{p(\tau_0/3) + p(2\tau_0/3) - p(\tau_0)}{\tau_0/3} > 0$.

In our proof of Theorem 1, we will consider the following function

$$g_{\theta,\mu}(t) := p(|t|) + \theta \cdot |t|^q + \mu \cdot |t - \hat{\tau}|^q$$

with $\theta, \mu > 0$, where $\hat{\tau}$ is an arbitrary fixed rational number in (τ_0, τ). We have the following lemma about $g_{\theta,\mu}(t)$.

Lemma 5. *If $p(t)$ satisfies the conditions in Theorem 1, $q > 1$, and τ_0 satisfies the properties in Lemma 4, then there exist $\underline{\theta} > 0$ and $\underline{\mu} > 0$ such that for any $\theta \geq \underline{\theta}$ and $\mu \geq \underline{\mu} \cdot \theta$, the following properties are satisfied:*

1. *$g''_{\theta,\mu}(t) \geq 1$ for any $t \in [\tau_0, \tau]$;*
2. *$g_{\theta,\mu}(t)$ has a unique global minimizer $t^*(\theta, \mu) \in (\tau_0, \tau)$;*
3. *Let $\bar{\delta} = \min\{t^*(\theta, \mu) - \tau_0, \tau - t^*(\theta, \mu), 1\}$, then for any $\delta \in (0, \bar{\delta})$, we have $g_{\theta,\mu}(t) < h(\theta, \mu) + \delta^2$ only if $|t - t^*(\theta, \mu)| < \delta$, where $h(\theta, \mu)$ is the minimal value of $g_{\theta,\mu}(t)$.*

Lemma 6. *If $p(t)$ satisfies the conditions in Theorem 1, $q = 1$, and τ_0 satisfies the properties in Lemma 4, then there exist $\underline{\mu} > 0$ such that for any $\mu \geq \underline{\mu}$, the following properties are satisfied:*

1. *$g'_{0,\mu}(t) < -1$ for any $t \in [\tau_0, \hat{\tau})$ and $g'_{0,\mu}(t) > 1$ for any $t \in (\hat{\tau}, \tau]$;*
2. *$g_{0,\mu}(t)$ has a unique global minimizer $t^*(0, \mu) = \hat{\tau} \in (\tau_0, \tau)$;*
3. *Let $\bar{\delta} = \min\{\hat{\tau} - \tau_0, \tau - \hat{\tau}, 1\}$, then for any $\delta \in (0, \bar{\delta})$, we have $g_{0,\mu}(t) < h(0, \mu) + \delta^2$ only if $|t - \hat{\tau}| < \delta$.*

By combining the above results, we have the following lemma, which is useful in our proof of Theorem 1.

Lemma 7. *Suppose $p(t)$ satisfies the conditions in Theorem 1 and τ_0 satisfies the properties in Lemma 4. Let $h(\theta, \mu)$ and $t^*(\theta, \mu)$ be as defined in Lemma 5 and Lemma 6 respectively for the case $q > 1$ and $q = 1$. Then we can find θ and μ such that for any $l \geq 2$, $t_1, \dots, t_l \in \mathbb{R}$,*

$$\sum_{j=1}^{l} p(|t_j|) + \theta \cdot \left|\sum_{j=1}^{l} t_j\right|^q + \mu \cdot \left|\sum_{j=1}^{l} t_j - \hat{\tau}\right|^q \geq h(\theta, \mu).$$

Moreover, let $\bar{\delta} = \min\left\{\frac{\tau_0}{3}, \frac{t^*(\theta,\mu) - \tau_0}{2}, \frac{\tau - t^*(\theta,\mu)}{2}, 1, C_1\right\}$ where C_1 is defined in Lemma 4, then for any $\delta \in (0, \bar{\delta})$, we have

$$\sum_{j=1}^{l} p(|t_j|) + \theta \cdot \left|\sum_{j=1}^{l} t_j\right|^q + \mu \cdot \left|\sum_{j=1}^{l} t_j - \hat{\tau}\right|^q < h(\theta, \mu) + \delta^2 \tag{3}$$

holds only if $|t_i - t^*(\theta, \mu)| < 2\delta$ for some i while $|t_j| \leq \delta$ for all $j \neq i$.

Proof of Theorem 1. We present a polynomial time reduction to problem (1) from the 3-partition problem. For any given instance of the 3-partition problem with $b = (b_1, \dots, b_{3m})$, we consider the minimization problem $\min_x f(x)$ in the form of (1) with $x = \{x_{ij}\}, 1 \leq i \leq$

$3m, 1 \leq j \leq m$, where

$$f(x) := \sum_{j=2}^{m} \left| \sum_{i=1}^{3m} b_i x_{ij} - \sum_{i=1}^{3m} b_i x_{i1} \right|^q + \sum_{i=1}^{3m} \left| (\lambda\theta)^{\frac{1}{q}} \sum_{j=1}^{m} x_{ij} \right|^q$$
$$+ \sum_{i=1}^{3m} \left| (\lambda\mu)^{\frac{1}{q}} \left(\sum_{j=1}^{m} x_{ij} - \hat{\tau} \right) \right|^q + \lambda \sum_{i=1}^{3m} \sum_{j=1}^{m} p(|x_{ij}|).$$

Note that the lower bounds $\underline{\theta}$, $\underline{\mu}$, and $\hat{\underline{\mu}}$ only depend on the penalty function $p(\cdot)$, we can choose $\theta \geq \underline{\theta}$ and $\mu \geq \underline{\mu}\theta$ if $q > 1$, or $\theta = 0$ and $\mu \geq \hat{\underline{\mu}}$ if $q = 1$, such that $(\lambda\theta)^{1/q}$ and $(\lambda\mu)^{1/q}$ are both rational numbers. Since $\hat{\tau}$ is also rational, all the coefficients of $f(x)$ are of finite size and independent of the input size of the given 3-partition instance. Therefore, the minimization problem $\min_x f(x)$ has polynomial size with respect to the given 3-partition instance.

For any x, by Lemma 7,

$$f(x) \geq 0 + \lambda \cdot \sum_{i=1}^{3m} \left\{ \sum_{j=1}^{m} p(|x_{ij}|) + \theta \cdot \left| \sum_{j=1}^{m} x_{ij} \right|^q + \mu \cdot \left| \sum_{j=1}^{m} x_{ij} - \hat{\tau} \right|^q \right\} \geq 3m\lambda \cdot h(\theta, \mu). \tag{4}$$

Now we claim that there exists an equitable partition to the 3-partition problem if and only if the optimal value of $f(x)$ is smaller than $3m\lambda \cdot h(\theta, \mu) + \epsilon$ where ϵ is specified later. On one hand, if S can be equally partitioned into m subsets, then we define

$$x_{ij} = \begin{cases} t^*(\theta, \mu) & \text{if } b_i \text{ belongs to the } j\text{th subset;} \\ 0 & \text{otherwise.} \end{cases}$$

It can be easily verified that these x_{ij}'s satisfy $f(x) = 3m\lambda \cdot h(\theta, \mu)$. Then due to (4), we know that these x_{ij}'s provide an optimal solution to $f(x)$ with optimal value $3m\lambda \cdot h(\theta, \mu)$.

On the other hand, suppose the optimal value of $f(x)$ is $3m\lambda \cdot h(\theta, \mu)$, and there is a polynomial-time algorithm that solves (1). Then for

$$\delta = \min \left\{ \frac{\tau_0}{8 \sum_{i=1}^{3m} b_i}, \bar{\delta} \right\} \quad \text{and} \quad \epsilon = \min\{\lambda\delta^2, (\tau_0/2)^q\}$$

where

$$\bar{\delta} = \min \left\{ \frac{\tau_0}{3}, \frac{t^*(\theta, \mu) - \tau_0}{2}, \frac{\tau - t^*(\theta, \mu)}{2}, \frac{p(\tau_0/3) + p(2\tau_0/3) - p(\tau_0)}{\tau_0/3}, 1 \right\},$$

we are able to find a near-optimal solution x such that $f(x) < 3m\lambda \cdot h(\theta, \mu) + \epsilon$ within a polynomial time of $\log(1/\epsilon)$ and the size of $f(x)$, which is polynomial with respect to the size of the given 3-partition instance. Now we show that we can find an equitable partition based on this near-optimal solution. By the definition of ϵ, $f(x) < 3m\lambda \cdot h(\theta, \mu) + \epsilon$ implies

$$\sum_{j=1}^{m} p(|x_{ij}|) + \theta \left| \sum_{j=1}^{m} x_{ij} \right|^q + \mu \cdot \left| \sum_{j=1}^{m} x_{ij} - \tau \right|^q < h(\theta, \mu) + \delta^2, \quad \forall i = 1, \ldots, 3m. \tag{5}$$

According to Lemma 7, for each $i = 1, \ldots, 3m$, (5) implies that there exists k such that $|x_{ik} - t^*(\theta, \mu)| < 2\delta$ and $|x_{ij}| < \delta$ for any $j \neq k$. Now let

$$y_{ij} = \begin{cases} t^*(\theta, \mu) & \text{if } |x_{ik} - t^*(\theta, \mu)| < 2\delta \\ 0 & \text{if } |x_{ij}| < \delta \end{cases}.$$

We define a partition by assigning b_i to the jth subset S_j if $y_{ij} = t^*(\theta, \mu)$. Note that this partition is well-defined since for each i, by the definition of δ, there exists one and only one $y_{ik} = t^*(\theta, \mu)$ while the others equal 0. Now we show that this is an equitable partition.

Note that for any $j = 1, \ldots, m$, the difference between the sum of the j-th subset and the first subset is

$$\left| \sum_{S_j} b_i - \sum_{S_1} b_i \right| = \left| \sum_{i=1}^{3m} \frac{y_{ij}}{t^*(\theta, \mu)} \cdot b_i - \sum_{i=1}^{3m} \frac{y_{i1}}{t^*(\theta, \mu)} \cdot b_i \right|$$
$$= \frac{1}{t^*(\theta, \mu)} \left| \sum_{i=1}^{3m} b_i y_{ij} - \sum_{i=1}^{3m} b_i y_{i1} \right|.$$

By triangle inequality, we have

$$\left| \sum_{S_j} b_i - \sum_{S_1} b_i \right| \leq \frac{1}{t^*(\theta, \mu)} \left(\sum_{i=1}^{3m} b_i \cdot |y_{ij} - x_{ij}| + \sum_{i=1}^{3m} b_i \cdot |y_{i1} - x_{i1}| + \left| \sum_{i=1}^{3m} b_i x_{ij} - \sum_{i=1}^{3m} b_i x_{i1} \right| \right).$$

By the definition of y_{ij}, we have $|y_{ij} - x_{ij}| < 2\delta$ for any i, j. for the last term, since $f(x) < 3m\lambda \cdot h(\theta, \mu) + \epsilon$, we know that

$$\left| \sum_{i=1}^{n} b_i x_{ij} - \sum_{i=1}^{n} b_i x_{i1} \right| < \epsilon^{1/q} \leq \tau_0/2.$$

Therefore, we have

$$\left| \sum_{S_j} b_i - \sum_{S_1} b_i \right| < \frac{1}{t^*(\theta, \mu)} \left(4\delta \sum_{i=1}^{n} b_i + \frac{\tau_0}{2} \right) \leq 1.$$

Now since b_i's are all integers, we must have $\sum_{S_j} b_i = \sum_{S_1} b_i$, which means that the partition is equitable. □

References

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

Antoniadis, A. and Fan, J. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001.

Arora, S., Babai, L., Stern, J., and Sweedy, Z. The hardness of approximate optima in lattices, codes, and systems of linear equations. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pp. 724–733. IEEE, 1993.

Bian, W. and Chen, X. Optimality conditions and complexity for non-lipschitz constrained optimization problems. *Preprint*, 2014.

Cameron, A. C. and Trivedi, P. K. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

Candes, E., Wakin, M., and Boyd, S. Enhancing sparsity by reweighted L_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.

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

Chen, X., Ge, D., Wang, Z., and Ye, Y. Complexity of unconstrained $L_2 - L_p$ minimization. *Mathematical Programming*, 143(1-2):371–383, 2014.

Chen, Y. and Wang, M. Hardness of approximation for sparse optimization with L_0 norm. *Technical Report*, 2016.

Davis, G., Mallat, S., and Avellaneda, M. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Fan, J. and Lv, J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20 (1):101–148, 2010.

Fan, J., Xue, L., and Zou, H. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819–849, 2014.

Fan, J., Liu, H., Sun, Q., and Zhang, T. TAC for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *arXiv preprint arXiv:1507.01037*, 2015.

Foster, D., Karloff, H., and Thaler, J. Variable selection is hard. In *COLT*, pp. 696–709, 2015.

Frank, L. E. and Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2): 109–135, 1993.

Garey, M. R. and Johnson, D. S. "Strong"NP-completeness results: Motivation, examples, and implications. *Journal of the ACM (JACM)*, 25(3):499–508, 1978.

Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Huo, X. and Chen, J. Complexity of penalized likelihood estimation. *Journal of Statistical Computation and Simulation*, 80(7):747–759, 2010.

Loh, P.-L. and Wainwright, M. J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pp. 476–484, 2013.

Lv, J. and Fan, Y. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37(6A):3498–3528, 2009.

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

Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Wang, Z., Liu, H., and Zhang, T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Xue, L., Zou, H., Cai, T., et al. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429, 2012.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38 (2):894–942, 2010a.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010b.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *COLT*, 2014.