Journal of Machine Learning Research 20 (2019) 1-27 Submitted 07/17; Revised 01/18; Published 02/18

Approximation Hardness for A Class of Sparse Optimization Problems

Yichen Chen YICHENC@PRINCETON.EDU
Department of Computer Science, Princeton University
Princeton, NJ 08544, USA

Yinyu Ye YINYU-YE@STANFORD.EDU
Department of Management Science and Engineering, Stanford University
Stanford, CA 94304, USA

Mengdi Wang MENGDIW@PRINCETON.EDU
Department of Operations Research and Financial Engineering, Princeton University
Princeton, NJ 08544, USA

Editor: David Wipf

Abstract

In this paper, we consider three typical optimization problems with a convex loss function and a nonconvex sparse penalty or constraint. For the sparse penalized problem, we prove that finding an $\mathcal{O}(n^{c_1}d^{c_2})$-optimal solution to an $n \times d$ problem is strongly NP-hard for any $c_1, c_2 \in [0, 1)$ such that $c_1 + c_2 < 1$. For two constrained versions of the sparse optimization problem, we show that it is intractable to approximately compute a solution path associated with increasing values of some tuning parameter. The hardness results apply to a broad class of loss functions and sparse penalties. They suggest that one cannot even approximately solve these three problems in polynomial time, unless P = NP.

Keywords: nonconvex optimization, computational complexity, variable selection, NP-hardness, sparsity

1. Introduction

Sparsity is a prominent modeling tool for extracting useful information from high-dimensional data. A practical goal is to minimize the empirical loss using as few variables/features as possible. In this paper, we consider three typical optimization problems arising from sparse machine learning. The first problem takes the form of empirical risk minimization with an additive sparse penalty.

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 also consider two constrained versions of sparse optimization, which are given by Problems 2 and 3. Such problems arise from sparse estimation (Shen et al., 2012) and sparse recovery (Natarajan, 1995; Bruckstein et al., 2009).

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

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

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

Problem 3 Given the loss function $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$, penalty function $p : \mathbb{R} \mapsto \mathbb{R}^+$, consider the problem

$$\min_{x \in \mathbb{R}^d} \sum_{j=1}^{d} p\left(|x_j|\right) \quad \text{s.t.} \sum_{i=1}^{n} \ell\left(a_i^T x, b_i\right) \leq \eta,$$

where $A = (a_1, \ldots, a_n)^T \in \mathbb{R}^{n \times d}$, $b = (b_1, \ldots, b_n)^T \in \mathbb{R}^n$ and the error tolerance parameter $\eta \geq 0$ are input data.

For a given sparsity level K, the optimal solution to Problem 2 is the best K-sparse solution that fits the data set (A, b). 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. Similarly for Problem 3, one often needs to compute the solution path that is associated with a sequence of values of η.

We are interested in the computational complexity of Problems 1, 2 and 3 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 nonconvex function with a unique minimizer at 0. 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 (Huo and 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. The technical note by Chen and Wang (2016) established the hardness of approximately solving Problems 1, 2 for the special case where p is the L_0 norm.

In this paper, we prove that it is strongly NP-hard to approximately solve Problems 1, 2 and 3 within certain levels of suboptimality. For Problem 1, we show that there exists a worst-case lower bound on the suboptimality error that can be achieved by any tractable deterministic algorithm. For Problems 2 and 3, we show that there does not exist any pseudo polynomial-time algorithm that can approximately compute a solution path where K or η increases at a certain speed. 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 is the first work that gives the proof of the approximation hardness for sparse optimization under general conditions on ℓ and p. A preliminary conference version of this paper (Chen et al.,

2017) focused solely on Problem 1. The current journal version extends the analysis to sparse constrained optimization and establishes hardness results for Problems 2 and 3.

Our results on optimization complexity provide new insights into the complexity of sparse feature selection (Zhang et al., 2014; Foster et al., 2015). 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 (although our lower bound is worst-case). 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 when increasing the sparsity level. These observations provide strong evidences for the hardness of variable selection.

Our main contributions are four-folded.

1. We prove the strong NP-hardness for Problems 1, 2 and 3 with a general loss function $\ell(\cdot)$, which are no longer limited to L_2 or L_p functions. These are the first results that apply to the broad class of problems including but not limited to: least square regression, linear model with Laplacian noise, robust regression, Poisson regression, logistic regression, inverse Gaussian models and the generalized linear model under the exponential distributions.

2. We present a general condition on the penalty function $p(\cdot)$ such that Problems 1, 2 and 3 are strongly NP-hard. Our condition is a slightly weaker version of strict concavity. It only requires the penalty function be concave while ruling out the possibility of linear penalty function (i.e., the LASSO) which is concave but also convex. It is satisfied by typical penalty functions such as the L_p norm ($p \in [0, 1)$), clipped L_1 norm, smoothly clipped absolute deviation, etc. To the best of our knowledge, this is the most general 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. Our proof provides a first unified analysis that deals with a broad class of problems taking the form of Problem 1.

4. We prove that it is strongly NP-hard to distinguish the optimal values of instances of Problem 2 (or Problem 3) that are associated with increasing values of the sparsity parameter K (or the error tolerance parameter η). This is the first hardness result for sparsity-constrained optimization with general loss and penalty functions. It implies that it is hard to approximately compute a solution path for the purpose of parameter tuning.

Section 2 reviews the background of sparse optimization and related literatures in the complexity theory. 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. The full proofs are given in Section 6.

2. Background and Related Works

Sparse optimization problems are common in machine learning, estimation, and signal processing. The sparse penalty p plays the important role of variable/feature selection. A common technique of imposing sparsity is to penalize the objective with a penalty function, leading to Problem 1. A well known example is the LASSO where p is the L_1 norm penalty and ℓ is the regression objective (Tibshirani, 1996). Nonconvex choices of p have been extensively studied in order to provide stronger statistical guarantee to the optimal soluion. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty which forces the solution of Problem 1 to be unbiased, sparse and stable in certain statistical sense. Frank and Friedman (1993) proposed the bridge estimator which use the L_p norm $(0 < p < 1)$ as its penalty function. Other related works include exact reconstruction of sparse signals by Candes et al. (2008) and Chartrand (2007), high-dimensional variable selection by Fan and Lv (2010), sparse Ising model by Xue et al. (2012) and regularized M-estimators by Loh and Wainwright (2013), etc.

Problem 2 finds applications in sparse estimation and feature selection. The work by Shen et al. (2012) proposed a statistically optimal estimator as the solution to the maximimum likelihood problem with L_0 sparsity constraint

$$\min_{x\in\mathbb{R}^d} -\sum_{i=1}^{n} \log g(x; a_i, b_i) \quad \text{s.t.} \sum_{i=1}^{d} \|x_i\|_0 \leq K.$$

This is a special case of Problem 2. The work by Fang et al. (2015) proposed an L_0 constrained optimization problem for sparse estimation of large-scale graphical models, which is also a special case of Problem 2. Another related problem is sparse recovery, which is to find the sparsest solution to a system of equations within an error tolerance. For example, Natarajan (1995) considered the problem

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{d} \|x_i\|_0 \quad \text{s.t.} \sum_{i=1}^{n} (a_i^T x - b_i)^2 \leq \delta,$$

which is a special case of Problem 3. See the work by Bruckstein et al. (2009) for more examples of Problem 3.

Within the mathematical programming community, the complexity of Problem 1 has been considered in a few works. Huo and Chen (2010) first proved the hardness result for problems with a relaxed family of penalty functions $\min_{x\in\mathbb{R}^d} \|Ax-b\|_2^2+\lambda \sum_{i=1}^{d} p(|x_i|)$. They show that for the penalties in ℓ_0, hard-thresholded (Antoniadis and Fan, 2001) and SCAD (Fan and Li, 2001), the above optimization problem is NP-hard. Our result (Theorem 1) requires weaker conditions on $p(\cdot)$ than theirs. In particular, our results applies to the $L_p(0 < p < 1)$ penalization and the clipped L_1 penalty function specified in Section 3.1 which do not satisfy the conditions in their paper. Moreover, our result applies to a broad class of ℓ functions and obtains *strong NP-hardness*. 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 (Vazirani, 2001). It is a stronger notion than NP-hardness where NP-hard problems might still be fast to solve in practice using pseudo-polynomial algorithms if the coding size is small (Garey and Johnson, 1978). On the contrary, a

strongly NP-hard problem doesn't have such a pseudo-polynomial algorithm. Chen et al. (2014) showed that the L_2-L_p minimization $\min_{x\in\mathbb{R}^d} \|Ax-b\|_2^2 + \lambda\sum_{i=1}^d |x_i|^p$ is strongly NP-hard when $p \in (0,1)$. At the same time, Bian and Chen (2014) proved the strong NP-hardness for another class of penalty functions. Their result requires $p(t)$ to be at least locally strictly concave, while ours does not. In particular, among the examples listed in Section 3.1, their results do not apply to L_0 penalization and clipped L_1 penalty function. To the best of our knowledge, our results are the most general ones up to today, which contains as special cases a broad class of penalty functions including ℓ_0, hard-thresholded, SCAD, L_p penalization ($p \in (0,1)$), folded concave penalty family (Fan et al., 2014) etc.

Within the theoretical computer science community, there have been several early works on the complexity of sparse recovery, beginning with the work by Arora et al. (1993). Amaldi and 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 linear 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 the estimation of linear models and impose distributional assumptions regarding the input data. For comparison with our results, theirs are stronger in the sense that they exclude the existence of any tractable randomized algorithm that succeds with high probability, while ours apply to deterministic algorithms. In the mean time, their results are less general than ours in the sense that they assume specific data distributions and specific loss functions, while ours are concerned with a much more general setting. In short, existing complexity results on sparse recovery are different in nature with our results on sparse computational optimization.

There remain several open questions. First, existing results do not apply to general loss functions ℓ or sparse penalties p. Existing analyses rely on specific properties of the L_q loss functions, such as the linear shift property $\|ax\|^q = a^q\|x\|^q$ and the property that L_q has sufficiently large second-order derivative around its minimum. However, these nice properties are lost in a majority of estimation problems, such as logistic regression and poisson regression. Second, the existing results from mathematical programming community apply only to the unconstrained Problem 1. The computational complexity of Problems 2 and 3 remain under-investigated in the community of optimization. Third, the results from computer science community apply to Problem 2 when the penalty function is L_0. These results work for specific loss functions and some of them impose distributional assumption about the input (Foster et al., 2015). In this paper, we focus on the worst-case complexity without making any distributional assumption regarding the input data. In this setting, the complexity of Problems 2 and 3 with penalty functions other than L_0 is yet to be established.

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 On The 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 following 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 and Friedman, 1993).

3. To obtain a hard-thresholding estimator, Antoniadis and 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 Assumption 1. Examples include the SCAD (Fan and 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 Assumption 1 are also satisfied by the clipped L_1 penalty function (Antoniadis and 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 and 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 On The 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 distance to 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 and 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 sparse optimization problem. We will provide formal statements and proof of these results in subsequent sections.

4. Main Results

In this paper, we aim to clarify the complexity for a broader class of sparse optimization problems taking the form of Problems 1, 2 and 3. 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 $f(\bar{x}) \le \inf_{x\in X} f(x) + \epsilon$.

Theorem 1 (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 1 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 1, 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.

Next we study the complexity of the sparsity-constrained Problem 2. We denote by $\ell_n(x)$ the normalized loss function:

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

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

$$x_K^* \in \operatorname{argmin} \left\{ \ell_n(x) \middle| \sum_{j=1}^{d} p(|x_j|) \leq K \right\}.$$

We obtain the following result.

Theorem 2 (Strong NP-Hardness of Problem 2) *Let Assumptions 1 and 2 hold, and let $c_1, c_2 \in [0, 1)$ be arbitrary such that $c_1 + c_2 < 1$. Let $\hat{x}_K$ be the approximate solution to Problem 2 with sparsity parameter K. Then there does not exist a pseudo polynomial-time algorithm that takes the input of Problem 2 and outputs a sequence of approximate solutions satisfying*

$$\ell_n(\hat{x}_{K+\kappa n^{c_1} d^{c_2}}) \leq \ell_n(x_K^*),$$

for all $K = 0, \kappa n^{c_1} d^{c_2}, 2\kappa n^{c_1} d^{c_2}, \ldots,$ unless P=NP, where $\kappa = \min_{t \in [\tau/2, \tau]} \{\frac{2p(t/2) - p(t)}{t}\}$.

Let us interpret the results of Theorem 2 in a practical setting. Suppose that we want to solve a sequence of sparsity-constrained problems with different values of the sparsity parameter K. The aim is to compare the corresponding empirical losses $\{\ell_n(x_K^*)\}$ and tune the parameter K.

Theorem 2 suggests that we cannot decide whether and how much the objective value will change by increasing the sparsity level from K to $K + \kappa n^{c_1} d^{c_2}$. Even if $\ell_n(x_K^*)$ is known as a benchmark, we can not find a better approximation of $\ell_n(x_{K+\kappa n^{c_1} d^{c_2}}^*)$ in polynomial time. In short, Theorem 2 tells us that it is computationally intractable to differentiate the minimal empirical losses that correspond to different values of K, unless P=NP. This implies that tuning the parameter K is computationally intractable.

Our last result concerns the error-constrained Problem 3.

Theorem 3 *Let Assumptions 1 and 2 hold, and let $c \in [0,1)$ be arbitrary. Let $\hat{x}_\eta$ be the approximate solution to Problem 3 with error tolerance η and let x^*_η be the corresponding optimal solution. There does not exist a pseudo polynomial-time algorithm that takes the input of Problem 3 and outputs a sequence of approximate solutions satisfying*

$$\sum_{j=1}^{d} p\left((\hat{x}_{\eta+\kappa n^{c_1} d^{c_2}})_j\right) \leq \sum_{j=1}^{d} p\left((x^*_\eta)_j\right),$$

for all $\eta = 0, \kappa n^{c_1} d^{c_2}, 2\kappa n^{c_1} d^{c_2}, \ldots,$ unless P=NP. Here, $(x)_j$ is the j-th component of vector x and $\kappa = \min_{t\in[\tau/2,\tau]}\{\frac{2p(t/2)-p(t)}{t}\}$.

Theorems 1, 2 and 3 are closely related to one another. Recall that the goal of sparse optimization is to make both the loss function and sparsity level small. Theorem 2 and Theorem 3 suggest that it is not possible to approximate the solution path, where either the loss tolerance or the sparsity level varies, in polynomial time. In contrast, Theorem 1 proves the approximation hardness for the sum between the loss tolerance and the sparsity level, when a fixed λ is used.

Theorems 1, 2 and 3 validate the long-lasting belief that optimization involving nonconvex penalty is hard. They provide worst-case 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. Implications of The Hardness Results

In this section, we interpret the strong NP-hardness results in the contexts of linear regression with SCAD penalty (which is a special case of Problem 1) and sparsity parameter tuning (which is related to Problem 2). 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 the work by Fan and Li (2001) and obtain a special case of Problem 1, given by

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

where $\gamma = \sqrt{\log d/n}$. Fan and Li (2001) showed that the optimal solution x^* of problem (1) 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 (1) 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 (1) 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 (1), 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 (1) within the required statistical precision $\sqrt{n \log d}$, where there is no distributional assumption on the data.

This gap is due to that the positive statistical properties of SCAD rely on strong distributional assumptions, while our hardness result does not. This illustrates a sharp contrast between the desirable statistical properties of sparse optimization under distributional assumptions and the worst-case computational complexity. In short, there does not exist a general-purpose polynomial algorithm.

5.2. Hardness of Tuning the Sparsity Level with L_0 Penalty

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 $L_n(x) = \frac{1}{n}\sum_{i=1}^n \ell(a_i^T x, b_i)$. 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

$$x_K^* \in \operatorname{argmin} \{L_n(x) \mid \|x\|_0 \leq K\},$$

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 1 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 $L_n(x_K^*)$ and $L_n(x_{K+1}^*)$.

Now let us interpret the results of Theorem 3 in the setting of tuning parameter K. Theorem 3 can be translated as follows. There 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 $(K+1)$-sparse problem is at least

$$L_n(x_K^*) - L_n(x_{K+1}^*) > 0.$$

The minimal empirical loss using K features is the best possible approximation to the minimal loss using $K+1$ features. In other words, we cannot decide whether and how much the objective value will change by increasing the sparsity level from K to $K+1$. Even if $L_n(x_K^*)$ is known as a benchmark, we can not find a better approximation of $L_n(x_{K+1}^*)$ in polynomial-time. In summary, Theorem 2 tells us that it is computationally intractable to differentiate between the sparsity levels K and $K+1$, unless P=NP. This implies that selection of the sparsity level is computationally intractable.

5.3. Remarks on the NP-Hardness Results

As illustrated by the preceding analysis, the non-approximibility of Problems 1, 2 and 3 suggests that computing the sparse estimator and tuning the sparsity parameter are hard *in the worst case*. 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 and 3 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. It is possible that the worst-case instances are highly unlikely to occur in practical situations.

2. Our results do not exclude the possibility that, under more stringent modeling and distributional assumptions, the problem would be tractable with high probability or on average.

In short, the sparse optimization Problems 1, 2 and 3 are 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. Technical Proofs

In this section, we prove the hardness of approximation of Problem 1, 2 and 3 for general loss function ℓ and penalty function p. We develop the reduction proof through a series of preliminary lemmas.

6.1. Preliminary Lemmas

Our first lemma gives us a key fact about the nonconvex penalty function p. We use $B(\theta, \delta)$ to denote the interval $(\theta - \delta, \theta + \delta)$.

Lemma 4 *For any penalty function p that satisfies Assumption 1, we have*

(i) *$p(t)$ is continuous on $(0, \tau]$.*

(ii) *For any $t_1, ..., t_l \geq 0$, if $\sum_{i=1}^n t_i \leq \tau$, then $\sum_{i=1}^l p(t_i) \geq p(\sum_{i=1}^l t_i)$.*

(iii) *There exists $a \in [1/2, 1)$ such that when $\sum_{i=1}^l t_i \in [a\tau, \tau]$, the above inequality holds as equality if and only if $t_i = t^*$ for some i while $t_j = 0$ for $j \neq i$.*

(iv) *Denote $\kappa = \min_{t \in [a\tau, \tau]} \{\frac{2p(t/2) - p(t)}{t}\}$. For the constant a given in (iii), we have that $\forall \delta > 0, t_1, \cdots, t_l \in \mathbb{R}, \forall \epsilon \leq \kappa\delta$: if $\sum_{i=1}^l t_i = t^* \in [a\tau, \tau]$ and $p(\sum_{i=1}^l t_i) + \epsilon \geq \sum_{i=1}^l p(t_i)$, then there is at most one i such that $t_i \notin B(0, \delta)$.*

Proof As (i), (ii) and (iii) are proved by Ge et al. (2015), we prove (iv) here. We first prove the lemma when $t_1, \cdots, t_l \geq 0$. We start by proving the case when $l = 2$. By (iii), there exists a such that when $t^* \in [a\tau, \tau]$ and $p(t^*) \geq p(t_1) + p(t_2)$, we have $t_1 = 0$ or $t_2 = 0$. It follow that when $t_1 \neq 0, t_2 \neq 0$ and $t^* \in [a\tau, \tau]$, we have $p(t_1 + t_2) < p(t_1) + p(t_2)$. Without loss of genearlity, we assume that $t_1 \leq t_2$. Then, we have

$$\frac{p(t^*) - p(t^* - t_1)}{t_1} < \frac{p(t_1)}{t_1}.$$

Notice that the right term is non-increasing with the increment of t_1 as p is a concave function and the left term is non-decreasing with the increment of t_1 when t^* is fixed. As

$t_1 \leq t^*/2$, we have $\frac{p(t_1)}{t_1} \geq k_1(t^*) \coloneqq \frac{p(t^*/2)}{t^*/2}$ and $\frac{p(t^*)-p(t^*-t_1)}{t_1} \leq k_2(t^*) \coloneqq \frac{p(t^*)-p(t^*/2)}{t^*/2}$. As p is not linear on $[0, t^*]$, we have $k_1(t^*) > k_2(t^*)$.

On the other hand, we can see that when $p(t_1 + t_2) + \epsilon \geq p(t_1) + p(t_2)$,

$$\frac{p(t_1 + t_2) - p(t_2)}{t_1} + \frac{\epsilon}{t_1} \geq \frac{p(t_1)}{t_1}.$$

Assume $t_1 < t_2$, we have $k_2(t^*) + \epsilon/t_1 \geq k_1(t^*)$ [1]. As a result $t_1 \leq \frac{\epsilon}{k_1(t^*)-k_2(t^*)}$. Note that k_1 and k_2 are defined on a closed interval $[a\tau, \tau]$ by (iii), giving us that $\min_{t\in[a\tau,\tau]}(k_1(t) - k_2(t)) > 0$. Therefore, $\exists a \in (0, 1), \forall \delta > 0, \exists \epsilon_0 = \min_{t\in[a\tau,\tau]}(k_1(t) - k_2(t)) \cdot \delta, \forall \epsilon < \epsilon_0$, if $t_1 + t_2 = t^* \in [a\tau, \tau]$ and $p(t_1 + t_2) + \epsilon \geq p(t_1) + p(t_2)$, then $t_1 \leq \frac{\epsilon}{k_1(t^*)-k_2(t^*)} \leq \delta$. Therefore, there is at most one i such that $t_i \notin B(0, \delta)$.

Now consider the case when $l > 2$ and $t_1, \ldots, t_l \geq 0$. If there are more than one i such that $t_i \notin B(0, \delta)$, assume t_1 and t_2 are two of them. By (ii), we have

$$\sum_{i=1}^{l} p(t_i) \geq p(t_1) + p\left(\sum_{i=2}^{l} t_i\right).$$

If $t_1 + \sum_{i=2}^{n} t_i \in [a\tau, \tau]$ and $p(t_1 + \sum_{i=2}^{l} t_i) + \epsilon \geq \sum_{i=1}^{l} p(t_i) \geq p(t_1) + p(\sum_{i=2}^{l} t_i)$, either t_1 or $\sum_{i=2}^{n} t_i$ should be inside $B(0, \delta)$. This is contradictory to our assumption that both t_1 and t_2 are outside $B(0, \delta)$. To this point, we prove (iv) when $t_1, \cdots, t_l \geq 0$.

Next, we prove the lemma when $t_1, \cdots, t_l$ could be smaller than 0. Suppose $t^* = \sum_{i=1}^{l} t_i \in [a\tau, \tau]$ and $p(t^*) + \epsilon \geq \sum_{i=1}^{l} p(t_i)$. We consider two cases separately. In the first case, assume that there is one $t_i \leq -\delta$. Without loss of generality, we assume that $t^* > 0$. Then we can choose $\alpha = \delta, \beta = t^* - \alpha$ and get

$$p(\alpha + \beta) + \epsilon = p(t^*) + \epsilon \geq \sum_{i\in\{j:t_j<0\}} p(t_i) + \sum_{i\in\{j:t_j>0\}} p(t_i) \geq p(\alpha) + p(\beta),$$

which is a contradiction to the previous proof that only one of α, β could be outside of $B(0, \delta)$ as δ is smaller than $t^*/2$ by our choice and $\sum_{i\in\{j:t_j>0\}} t_i > t^* > t^* - \alpha$. We then proceed to the case when there is one $t_i \geq \delta$ and one $t_j \geq \delta$. Suppose that $\alpha = t_i \geq t_j = \beta$. If $\alpha + \beta > t^*$, we set $\alpha' = \delta + \frac{t^*-2\delta}{\alpha+\beta-2\delta} \cdot (\alpha - \delta)$ and $\beta' = \delta + \frac{t^*-2\delta}{\alpha+\beta-2\delta} \cdot (\beta - \delta)$. It is easy to verify that

$$p(\alpha' + \beta') + \epsilon = p(t^*) + \epsilon \geq \sum_{i=1}^{l} p(t_i) \geq p(\alpha) + p(\beta) \geq p(\alpha') + p(\beta'),$$

which is a contradiction. If $\alpha + \beta < t^*$, we can verify that

$$p(\alpha + \beta + t^* - \alpha - \beta) + \epsilon = p(t^*) + \epsilon \geq \sum_{i=1}^{l} p(t_i) \geq p(\alpha) + p(\beta) + p(t^* - \alpha - \beta),$$

which is also a contradiction. To this point, we prove the case that $t_1, \cdots, t_l$ could be smaller than 0, which completes the proof of the lemma.

1. For the case when $t_1 = 0$, (iv) holds trivially.

Remark. In the proof of (iv), our choice of ϵ is linear to δ given δ. However, in the case of L_0, ϵ could be any constant smaller than 1 no matter what δ is. This property of L_0 has wide applications in statistical problems. Actually, suppose that penalty function is indexed by δ and p_δ satisfies

$$p_\delta(\delta) - p_\delta(a\tau) + p_\delta(a\tau - \delta) \geq C$$

for some constant C, then $\forall \delta > 0$ and $\epsilon \leq C$, the proposition stated in (iv) holds. To prove this, just note that if $p(t_1+t_2)-p(t_2)+\epsilon > p(t_1)$ and $t_1 > \delta$, then $p(t_1)-p(t_1+t_2)+p(t_2) > p(\delta)-p(a\tau)+p(a\tau-\delta) \geq C$ which is a contradiction to that ϵ should be smaller than C. ■

Lemma 4 states the key properties of the penalty function p. Property (iv) is of special interest. It indicates that if we can manipulate the sum of non-negative variables to let it lie within $[a\tau, \tau]$ while minimizing the penalty function, we can be sure that only one variable has large absolute value.

Our second lemma explores the relationship between the penalty function p and the loss function ℓ.

Lemma 5 *Let Assumption 1 hold. Let $f(\cdot)$ be a convex function with a unique minimizer $\hat{\tau} \in (a\tau, \tau)$ and $\frac{f(\hat{\tau}\pm x)-f(\hat{\tau})}{x^N} \geq C (0 < x < \bar{\delta})$ for some $N \in \mathbb{Z}^+, \bar{\delta} \in \mathbb{R}^+, C \in \mathbb{R}^+$. Define*

$$g_\mu(t) = p(|t|) + \mu \cdot f(t),$$

where $\mu > 0$. Let $h(\mu)$ be the minimum value of $g_\mu(\cdot)$. We have $\forall \delta < \bar{\delta}, \mu_\delta > \frac{p(|\hat{\tau}|)2^N}{C\delta^N}, \exists \epsilon_0 = \mu_\delta \cdot C \cdot \left(\frac{\delta}{2}\right)^N - p(|\hat{\tau}|)$: if t satisfies $h(\mu_\delta) + \epsilon_0 \geq g_{\mu_\delta}(t) \geq h(\mu_\delta)$, then $t \in [\hat{\tau} - \delta/2, \hat{\tau} + \delta/2]$.

Proof First, we can see that when $t > \hat{\tau} + \delta/2$, we have

$$\begin{aligned} g_{\mu_\delta}(t) &\geq p(|\hat{\tau}|) + \mu_\delta \cdot f(t) > p(|\hat{\tau}|) + \mu_\delta \cdot f(\hat{\tau} + \delta/2) \geq p(|\hat{\tau}|) + \mu_\delta \cdot f(\hat{\tau}) + \mu_\delta \cdot C \cdot \left(\frac{\delta}{2}\right)^N \\ &= g_{\mu_\delta}(\hat{\tau}) + \mu_\delta \cdot C \cdot \left(\frac{\delta}{2}\right)^N \geq h(\mu_\delta) + \mu_\delta \cdot C \cdot \left(\frac{\delta}{2}\right)^N \geq h(\mu_\delta) + \epsilon_0, \end{aligned}$$

by the definition of $f(\cdot)$. When $t < \hat{\tau} - \delta/2$, we have

$$\begin{aligned} g_{\mu_\delta}(t) &\geq \mu_\delta \cdot f(t) > \mu_\delta \cdot f(\hat{\tau} - \delta/2) \geq \mu_\delta \cdot f(\hat{\tau}) + \mu_\delta \cdot C \cdot \left(\frac{\delta}{2}\right)^N \\ &= \mu_\delta \cdot f(\hat{\tau}) + \frac{p(|\hat{\tau}|)2^N}{C\delta^N} \cdot C \cdot \left(\frac{\delta}{2}\right)^N + \left(\mu_\delta - \frac{p(|\hat{\tau}|)2^N}{C\delta^N}\right) \cdot C \cdot \left(\frac{\delta}{2}\right)^N \\ &\geq h(\mu_\delta) + \mu_\delta \cdot C \cdot \left(\frac{\delta}{2}\right)^N - p(|\hat{\tau}|). \end{aligned}$$

Therefore, when we choose $\epsilon_0 = \mu_\delta \cdot C \cdot \left(\frac{\delta}{2}\right)^N - p(|\hat{\tau}|)$, point t satisfying $h(\mu_\delta) + \epsilon_0 \geq g_{\mu_\delta}(t) \geq h(\mu_\delta)$ must lie in $[\hat{\tau} - \delta/2, \hat{\tau} + \delta/2]$. ■

Lemma 6 *Let Assumption 1 hold and let $f(\cdot)$ be a convex function with a unique minimizer $\hat{\tau} \in (a\tau, \tau)$ and $\frac{f(\hat{\tau}\pm x)-f(\hat{\tau})}{x^N} \geq C_1 (0 < x < \bar{\delta})$ for some $N \in \mathbb{Z}^+, \bar{\delta} \in \mathbb{R}^+, C_1 \in \mathbb{R}^+$. Let $h(\mu)$ be the minimum value of $g_\mu(x) = p(|x|) + \mu \cdot f(x)$, then we have*

(i) $\forall \mu \in \mathbb{Z}^+, t_1, ..., t_n \in \mathbb{R} : \sum_{j=1}^n p(|t_j|) + \mu \cdot f\left(\sum_{j=1}^n t_j\right) \geq h(\mu)$.

(ii) $\exists \kappa = \min_{t\in[a\tau,\tau]}\{\frac{2p(t/2)-p(t)}{t}\}, \forall \delta \leq \min\{\bar{\delta}, 4\tau - 4\hat{\tau}, 4\hat{\tau} - 4a\tau\}, \exists \mu = \frac{p(|\hat{\tau}|)4^{N+1}}{C_1\delta^N}, \epsilon_0 = \kappa \cdot \frac{\delta}{n}, \forall \theta \in [\hat{\tau} - \delta/4, \hat{\tau} + \delta/4]$: *if $t_1, ..., t_n \in \mathbb{R}$ satisfy*

$$h(\mu) + \epsilon_0 \geq \sum_{j=1}^n p(|t_j|) + \mu \cdot f\left(\sum_{j=1}^n t_j\right) \geq h(\mu), \tag{2}$$

then $t_i \in B(\theta, \delta)$ for one i and $t_j \in B(0, \delta)$ for all $j \neq i$.

Proof

We first prove (i). We consider two cases separately. In the first case, we suppose that $|\sum_{j=1}^n t_j| > \tau$. Then we have

$$\sum_{j=1}^n p(|t_j|) \geq \sum_{j=1}^n p\left(\frac{\tau}{\sum_{k=1}^n |t_k|} \cdot |t_j|\right) \geq p\left(\sum_{j=1}^n \frac{\tau}{\sum_{k=1}^n |t_k|} \cdot |t_j|\right) \geq p(\tau),$$

where the first inequality is inferred by the monotonicity of p and the second inequality is due to (ii) of Lemma 4. Thus, we have

$$\sum_{j=1}^n p(|t_j|) + \mu \cdot f\left(\sum_{j=1}^n t_j\right) > \min\{p(\tau) + \mu \cdot f(\tau), p(\tau) + \mu \cdot f(-\tau)\} \geq h(\mu).$$

As a result, we can see that (i) holds when $|\sum_{j=1}^n t_j| > \tau$. In the second case, we suppose $|\sum_{j=1}^n t_j| \leq \tau$ and obtain

$$\sum_{j=1}^n p(|t_j|) \geq \sum_{j=1}^n p\left(\frac{|\sum_{k=1}^n t_k|}{\sum_{k=1}^n |t_k|} |t_j|\right) \geq p\left(\sum_{j=1}^n \frac{|\sum_{k=1}^n t_k|}{\sum_{k=1}^n |t_k|} |t_j|\right) \geq p\left(\left|\sum_{j=1}^n t_j\right|\right),$$

where the second inequality is due to (ii) of Lemma 4. It follows that

$$\sum_{j=1}^n p(|t_j|) + \mu \cdot f\left(\sum_{j=1}^n t_j\right) \geq p\left(\left|\sum_{j=1}^n t_j\right|\right) + \mu \cdot f\left(\sum_{j=1}^n t_j\right) = g_\mu\left(\sum_{j=1}^n t_j\right) \geq h(\mu). \tag{3}$$

which completes our proof of (i).

We then prove (ii). Assume equation (2) holds. If $\sum_{j=1}^n t_j > \tau$, we can see that by choosing $\epsilon_0 \leq g_\mu(\tau) - g_\mu(\hat{\tau})$, we have

$$\sum_{j=1}^n p(|t_j|) + \mu \cdot f\left(\sum_{j=1}^n t_j\right) > g_\mu(\tau) = g_\mu(\hat{\tau}) + g_\mu(\tau) - g_\mu(\hat{\tau}) \geq h(\mu) + \epsilon_0.$$

We will show later that our choice of ϵ_0 is indeed smaller than $g_\mu(\tau) - g_\mu(\hat{\tau})$. We will also show later that equation (2) cannot hold when $\sum_{j=1}^n t_j < -\tau$ under our choice of parameters. Thus, if equation (2) holds, then $|\sum_{j=1}^n t_j| \leq \tau$, which implies that

$$p\left(\left|\sum_{j=1}^n t_j\right|\right) + \mu \cdot f\left(\sum_{j=1}^n t_j\right) \leq h(\mu) + \epsilon_0, \tag{4}$$

by equation (2) and the first inequality of (3), and

$$\sum_{j=1}^n p(|t_j|) \leq p\left(\left|\sum_{j=1}^n t_j\right|\right) + \epsilon_0, \tag{5}$$

due to equation (2) and equation (3). Note that we just need to prove the case when δ is sufficiently small. Thus, we assume in the following paper that δ is smaller than $\bar{\delta}, 4\tau - 4\hat{\tau}, 4\hat{\tau} - 4a\tau$.

Consider the case when equation (4) holds. By Lemma 6, if we choose $\mu = \frac{p(|\hat{\tau}|)4^{N+1}}{C\delta^N}$ and $\epsilon_1 = 3p(|\hat{\tau}|)$, then all of the points t such that $h(\mu)+\epsilon_1 \geq g_\mu(t) \geq h(\mu)$ lie in $[\hat{\tau}-\delta/4, \hat{\tau}+\delta/4]$. Thus, we have $\sum_{j=1}^n t_j \in [a\tau, \tau]$ and $\sum_{j=1}^n t_j \in B(\theta, \frac{\delta}{2})$ for all $\theta \in [\hat{\tau} - \delta/4, \hat{\tau} + \delta/4]$. Note that $g_\mu(t)$ is non-increasing when $t < 0$, meaning that equation (2) cannot hold under our choice of ϵ_1 when $\sum_{j=1}^n t_j \leq -\tau$.

On the other hand, if equation (2) holds, equation (5) should also hold. By (iv) of Lemma 4, for the same δ, $\exists \epsilon_2 = \min_{t\in[a\tau,\tau]}(k_1(t) - k_2(t)) \cdot \frac{\delta}{2n-2}$, there is at most one i such that $t_i \notin B(0, \frac{\delta}{2n-2})$. As $\sum_{j=1}^n t_j \in B(\theta, \frac{\delta}{2})$, we have $t_i \in B(\theta, \delta)$ for all $i = 1, \cdots, n$. Observe that $g_\mu(\tau) - g_\mu(\hat{\tau})$ is always larger than ϵ_1. Also, $\epsilon_1 > \epsilon_2$ if δ is sufficiently small. Therefore, $\exists \kappa = \min_{t\in[a\tau,\tau]}(k_1(t) - k_2(t))/2$, $\forall \delta \leq \min\{\bar{\delta}, 4\tau - 4\hat{\tau}, 4\hat{\tau} - 4a\tau\}, \exists \mu = \frac{p(|\hat{\tau}|)4^{N+1}}{C\delta^N}, \epsilon = \kappa \cdot \frac{\delta}{n}, \forall \theta \in [\hat{\tau} - \delta/4, \hat{\tau} + \delta/4]$: if $h(\mu) + \epsilon \geq g_\mu(\sum_{j=1}^n t_i)$, then $t_i \in B(\theta, \delta)$ for some i while $t_j \in B(0, \delta)$ for all $j \neq i$. ■

6.2. Proof of Theorem 1

Now we are ready to prove Theorem 1.

Proof Suppose that we are given the input to the 3-partition problem, i.e., $3m$ positive integers $s_1, ..., s_{3m}$. Assume *without loss of generality* that all s_i's are upper bounded by some polynomial function $M(m)$. 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. We 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. For the interval $[a\tau, \tau]$, we choose $\{b_{1i}\}_{i=1}^{k_1}$ such that $\ell_1(y) = \frac{1}{\lambda}\sum_{i=1}^{k_1} \ell(y, b_{1i})$ satisfies Assumption 1 with constants $C, N, \bar{\delta}$ and has a unique minimizer $\hat{\tau}$ inside the interval $(a\tau, \tau)$. Let $\kappa = \min_{t\in[a\tau,\tau]}\{\frac{2p(t/2)-p(t)}{t}\}$. Let $\delta \leq \{\frac{a\tau}{9m\cdot M(m)}, \bar{\delta}, 4\tau - 4\hat{\tau}, 4\hat{\tau} - 4a\tau\}$,

$\mu \geq \frac{p(|\hat{\tau}|)4^{N+1}}{C_1\delta^N}$ and $\epsilon = \kappa \cdot \frac{\delta}{3m}$ such that Lemma 6 is satisfied. Note that $\epsilon \geq \frac{C_3}{m^2 \cdot M(m)}$ for some constant C_3 by our construction.

2. For the μ and ϵ chosen in the previous step, all the minimizers of $g_\mu(x) = p(|x|) + \mu \cdot \ell_1(x)$ lie in $[\hat{\tau}-\delta/4, \hat{\tau}+\delta/4]$ by Lemma 6. By the Lipschitz continuity of $p(|x|)$, $f(x)$ and thus $g_\mu(x)$ on $[a\tau, \tau]$, there exists $\delta_\epsilon = \frac{\epsilon}{6mK}$ (K is the Lipschitz constant) such that we can find in polynomial time an interval $[\theta_1, \theta_2]$ where $\theta_2 - \theta_1 = \delta_\epsilon$ and $g_\mu(x) - g_\mu(t^*) < \frac{\epsilon}{6m}$ for $x \in [\theta_1, \theta_2]$. This interval can be find in polynomial time as $g_\mu(x)$ is Lipschitz continuous.

3. By Assumption 2, for the interval $[\theta_1, \theta_2]$, we choose $\{b_{2i}\}_{i=1}^{k_2}$ to construct a loss function $\ell_2 : \mathbb{R} \mapsto \mathbb{R}$ in polynomial time with regard to $1/\delta_\epsilon$ such that $\ell_2(y) = \frac{1}{\lambda}\sum_{i=1}^{k_2} \ell(y, b_{2i})$ has a unique minimizer at $\tilde{t} \in [\theta_1, \theta_2]$. We choose

$$\nu = \left\lceil \epsilon / \max\left(\ell_2(\tilde{t} + 2\delta m) - \ell_2(\tilde{t}), \ell_2(\tilde{t} - 2\delta m) - \ell_2(\tilde{t})\right) \right\rceil + 1,$$

and construct function $f : \mathbb{R}^{3m \times m} \mapsto \mathbb{R}$ where

$$f(x) = \lambda \cdot \sum_{i=1}^{3m}\sum_{j=1}^{m} p\left(|x_{ij}|\right) + \lambda\mu \cdot \sum_{i=1}^{3m} \ell_1\left(\sum_{j=1}^{m} x_{ij}\right) + \lambda\nu \cdot \sum_{j=1}^{m} \ell_2\left(\sum_{i}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij}\right). \tag{6}$$

Note that by (iii) of Assumption 2, ν is polynomial in $\max(\lceil \frac{1}{\delta_\epsilon} \rceil, \lceil \theta_2 \rceil)$. In the rest of the paper, we ignore the $\lceil \theta_2 \rceil$ term in the bound as it can be upperbounded by τ, which can be taken as a constant in the reduction.

4. Let $\Phi_1 = 3m \cdot p(|\tilde{t}|) + \mu \cdot 3m \cdot \ell_1(\tilde{t}) - \frac{\epsilon}{2}$ and $\Phi_2 = \nu \cdot m \cdot \ell_2(\tilde{t})$. We claim that

   (i) If there exists z such that

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

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

   (ii) If the 3-partition problem has a solution, we have $\frac{1}{\lambda}\min_x f(x) \leq \Phi_1 + \Phi_2 + \frac{\epsilon}{2}$.

5. Choose $r = \left\lceil \left(\frac{2(3m\cdot\lambda\cdot\mu\cdot k_1 + m\cdot\lambda\cdot\nu\cdot k_2)^{c_1}(3m^2)^{c_2}}{\epsilon/\kappa}\right)^{1/(1-c_1-c_2)} \right\rceil$ where c_1 and c_2 are two arbitrary constants that $c_1 + c_2 < 1$. Construct the following instance of Problem 1:

$$\begin{aligned} \min_{x^{(1)},\cdots,x^{(r)} \in \mathbb{R}^{3m\times m}} \sum_{q=1}^{r} f(x^{(q)}) = \min_{x^{(1)},\cdots,x^{(r)} \in \mathbb{R}^{3m\times m}} \lambda \cdot \sum_{q=1}^{r}\sum_{i=1}^{3m}\sum_{j=1}^{m} p(|x_{ij}^{(q)}|) + \\ \lambda\mu \sum_{q=1}^{r}\sum_{i=1}^{3m}\sum_{t=1}^{k_1} \ell\left(\sum_{j=1}^{m} x_{ij}^{(q)}, b_{1t}\right) + \lambda\nu \sum_{q=1}^{r}\sum_{j=1}^{m}\sum_{t=1}^{k_2} \ell\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij}^{(q)}, b_{2t}\right), \end{aligned} \tag{7}$$

where the input data are coefficients of x and the values $b_{11}, \ldots, b_{1t}, b_{21}, \ldots, b_{2t}$. The variable dimension d is $r \cdot 3m^2$ and the sample size n is $\lambda \cdot \mu \cdot r \cdot 3m \cdot k_1 + \lambda \cdot \nu \cdot r \cdot m \cdot k_2$. The input size is polynomial with respect to m. Our choice of r is the solution to $\epsilon r = 2\kappa n^{c_1} d^{c_2}$ where $\kappa = \min_{t \in [a\tau, \tau]} \{\frac{2p(t/2) - p(t)}{t}\}$.

The parameters μ, ν, δ, r, d are bounded by polynomial functions of m. Computing their values also takes polynomial time. The parameter k_1 and k_2 is a constant determined by the loss function ℓ and is not related to m. As a result, the reduction is polynomial.

6. Let $z^{(1)}, \cdots z^{(r)} \in \mathbb{R}^{3m \times m}$ be a $\lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$-optimal solution to problem (13) such that $\sum_{i=1}^r f(z^{(i)}) \leq \min_{x^{(1)}, \cdots, x^{(r)}} \sum_{i=1}^r f(x^{(i)}) + \lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$. We claim that

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

   $$\frac{1}{\lambda} \sum_{i=1}^r f(z^{(i)}) \leq r\Phi_1 + r\Phi_2 + 2\kappa n^{c_1} d^{c_2}, \tag{8}$$

   we can choose one $z^{(i)}$ such that $\Phi_1 + \Phi_2 + \epsilon \geq \frac{1}{\lambda} f(z^{(i)}) \geq \Phi_1 + \Phi_2$ and obtain a feasible assignment: If $z_{ij}^{(i)} \in B(\tilde{t}, \delta)$, we assign number i to subset j. If the $\lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$-optimal solution $z^{(1)}, \cdots z^{(r)}$ does not satisfy (8), the 3-partition problem has no feasible solution.

We have constructed a polynomial reduction from the 3-partition problem to finding an $\lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$-optimal solution to problem (13). 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 6(i), we can see that for all $x \in \mathbb{R}^{3m \times m}$,

$$\sum_{i=1}^{3m} \sum_{j=1}^{m} p(|x_{ij}|) + \mu \cdot \sum_{i=1}^{3m} \ell_1 \left(\sum_{j=1}^{m} x_{ij} \right) \geq 3m \cdot p(|t^*|) + \mu \cdot 3m \cdot \ell_1(t^*) \geq \Phi_1,$$

where the last inequality is due to that $g_\mu(\tilde{t}) - g_\mu(t^*) < \frac{\epsilon}{6m}$. By the fact $\tilde{t} = \operatorname{argmin}_t \ell_2(t)$, we have for all $x \in \mathbb{R}^{3m \times m}$ that

$$\nu \cdot \sum_{j=1}^{m} h \left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij} \right) \geq \nu \cdot m \cdot \ell_2(\tilde{t}) = \Phi_2.$$

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

$$\Phi_1 + \epsilon \geq \sum_{i=1}^{3m} \sum_{j=1}^{m} p(|z_{ij}|) + \mu \cdot \sum_{i=1}^{3m} h \left(\sum_{j=1}^{m} z_{ij} \right) \geq \Phi_1, \tag{9}$$

and

$$\Phi_2 + \epsilon \geq \nu \cdot \sum_{j=1}^{m} h \left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij} \right) \geq \Phi_2. \tag{10}$$

In order for equation (9) to hold, we have that for all i,

$$p(|\tilde{t}|) + \mu \cdot \ell_1(\tilde{t}) + \frac{\epsilon}{2} \geq \sum_{j=1}^{m} p(|z_{ij}|) + \mu \cdot \ell_1 \left(\sum_{j=1}^{m} z_{ij} \right) \geq p(|t^*|) + \mu \cdot \ell_1(t^*).$$

Consider an arbitrary i. By Lemma 6(ii) and $g_\mu(\tilde{t}) - g_\mu(t^*) < \frac{\epsilon}{6m}$, we have $z_{ij} \in B(\tilde{t}, \delta)$ for one j while $z_{ik} = 0$ for all $k \neq j$. If $z_{ij} \in B(\tilde{t}, \delta)$, we assign number i to subset j. As $\delta < a\tau/2 \leq \tilde{t}/2$, $B(\tilde{t}, \delta)$ and $B(0, \delta)$ are not overlapping. Thus 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 $I_j = \{i \mid z_{ij} \in B(\tilde{t}, \delta)\}$. Thus, $\sum_{i \in I_j} s_i \geq \sum_{i=1}^{3m} s_i/m + 1$. As a result, we have

$$\begin{aligned} \sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij} &\geq \sum_{i \in I_1} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} (\tilde{t} - \delta) + \sum_{i \in I_2} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} (-\delta) \\ &\geq \frac{\sum_{i=1}^{3m} s_i/m + 1}{\sum_{i=1}^{3m} s_i/m} \tilde{t} - \delta m = \tilde{t} + \frac{\tilde{t}}{\sum_{i=1}^{3m} s_i/m} - \delta m. \end{aligned}$$

Because $s_i \leq M(m)$ for all i and $\delta = \frac{a\tau}{9m \cdot M(m)}$, we have

$$\frac{\tilde{t}}{\sum_{i=1}^{3m} s_i/m} - \delta m \geq \frac{a\tau}{3m \cdot M(n)} m - \delta m = 2\delta m > 0.$$

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

$$\ell_2 \left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij} \right) \geq \ell_2(\tilde{t} + 2\delta m).$$

By our construction of ν and Assumption 1(iii), we further have

$$\nu \cdot \left(\ell_2 \left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij} \right) - \ell_2(\tilde{t}) \right) \geq \nu \cdot \big(\ell_2(\tilde{t} + 2\delta m) - \ell_2(\tilde{t}) \big) > \epsilon. \tag{11}$$

However, in order for equation (10) to hold, we have that for all j,

$$\nu \cdot \ell_2(\tilde{t}) + \epsilon \geq \nu \cdot \ell_2 \left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij} \right) \geq \nu \cdot \ell_2(\tilde{t}),$$

yielding a contradiction to (11). 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 z to the optimization problem such that $f(z) \leq \Phi_1 + \Phi_2 + \frac{\epsilon}{2}$. For all

$1 \le i \le 3m$, if number i is assigned to subset j, let $z_{ij} = \tilde{t}$ and $z_{ik} = 0$ for all $k \neq j$. We can easily verify that

$$\sum_{i=1}^{3m}\sum_{j=1}^{m} p\left(|z_{ij}|\right) + \mu \cdot \sum_{i=1}^{3m} \ell_1\left(\sum_{j=1}^{m} z_{ij}\right) = 3m \cdot \left(p(\tilde{t}) + \mu \cdot \ell_1(\tilde{t})\right) = \Phi_1 + \frac{\epsilon}{2},$$

Also, we have

$$\nu \cdot \sum_{j=1}^{m} \ell_2\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} z_{ij}\right) = \nu \cdot m \cdot \ell_2(\tilde{t}) = \Phi_2.$$

Therefore,

$$\frac{1}{\lambda} f(z) \le \Phi_1 + \Phi_2 + \frac{\epsilon}{2}. \tag{12}$$

which completes the proof of (ii).

Step 4: Proof of Claim (iii). Suppose that the $\lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$-optimal solution satisfies (8), i.e., $\frac{1}{\lambda}\sum_{i=1}^{r} f(z^{(i)}) \le r\Phi_1 + r\Phi_2 + 2\kappa n^{c_1} d^{c_2}$. It follows that there exists at least one term $z^{(i)}$ such that

$$\frac{1}{\lambda} f(z^{(i)}) \le \Phi_1 + \Phi_2 + \frac{2\kappa n^{c_1} d^{c_2}}{r} \le \Phi_1 + \Phi_2 + \epsilon.$$

where the second inequality equality uses $\epsilon r = 2\kappa n^{c_1} d^{c_2}$. Therefore, by claim (ii), we can find a solution to the 3-partition problem.

Suppose that the 3-partition problem has a solution. By claim (ii), there exists z such that $\frac{1}{\lambda} f(z) \le \Phi_1 + \Phi_2 + \frac{\epsilon}{2}$. Thus we have

$$\min_{x^{(1)},\cdots,x^{(r)}} \frac{1}{\lambda}\sum_{i=1}^{r} f(x^{(i)}) \le \frac{r}{\lambda} f(z) \le r\Phi_1 + r\Phi_2 + \kappa n^{c_1} d^{c_2}.$$

Thus if $z^{(1)}, \cdots z^{(r)}$ is a $\lambda \cdot \kappa \cdot n^{c_1} d^{c_2}$-optimal solution to (13), we have

$$\frac{1}{\lambda}\sum_{i=1}^{r} f(z^{(i)}) \le \min_{x^{(1)},\cdots,x^{(r)}} \frac{1}{\lambda}\sum_{i=1}^{r} f(x^{(i)}) + \kappa n^{c_1} d^{c_2} \le r\Phi_1 + r\Phi_2 + 2\kappa n^{c_1} d^{c_2}$$

implying that the relation (8) must hold. If (8) is not satisfied, the 3-partition problem has no solution. ■

Remark. When the loss function is L_2 loss, we can move $\lambda\mu$ and $\lambda\nu$ of equation (13) into the loss. Specifically, we have

$$\begin{aligned}\min_{x^{(1)},\cdots,x^{(r)} \in \mathbb{R}^{3m\times m}} \sum_{q=1}^{r} f(x^{(q)}) = \min_{x^{(1)},\cdots,x^{(r)} \in \mathbb{R}^{3m\times m}} \lambda \cdot \sum_{q=1}^{r}\sum_{i=1}^{3m}\sum_{j=1}^{m} p(|x_{ij}^{(q)}|) + \\ \sum_{q=1}^{r}\sum_{i=1}^{3m}\left(\sum_{j=1}^{m}\sqrt{\lambda\mu}x_{ij}^{(q)} - \sqrt{\lambda\mu}b_1\right)^2 + \sum_{q=1}^{r}\sum_{j=1}^{m}\left(\sum_{i=1}^{3m}\frac{\sqrt{\lambda\nu}s_i}{\sum_{i'=1}^{3m} s_{i'}/m}x_{ij}^{(q)} - \sqrt{\lambda\nu}b_2\right)^2,\end{aligned} \tag{13}$$

where μ, ν is chosen such that $\sqrt{\lambda\mu}, \sqrt{\lambda\nu}$ are rational numbers. In this case, the variable dimension is $r \cdot 3m^2$ and the sample size n is $4r \cdot m$. Our choice of r is the solution to $\epsilon r = 2\kappa n^{c_1} d^{c_2}$ which is $r = \left\lceil \left(\frac{2(4m)^{c_1}(3m^2)^{c_2}}{\epsilon/\kappa} \right)^{1/(1-c_1-c_2)} \right\rceil$. The value of r doesn't depend on λ and p, which means that we can plug in any λ, p and the reduction is still polynomial in m. It means that for any choice of λ and p, it is strongly NP hard to find a $\lambda\kappa n^{c_1} d^{c_2}$-optimal solution.

6.3. Proof of Theorem 2

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

Proof We will use a reduction from the 3-partition problem to prove the theorem. The reduction is developed through the following steps. We first constructed a polynomial reduction from the 3-partition problem to finding an approximate solution to Problem 2. We then prove that the reduction works.

1. Given the input to the 3-partition problem, we conduct the first three steps of the reduction in Theorem 1 to compute $\mu, \nu, \tilde{t}$ and ϵ with $\lambda = 1$. The nuance is that we pick $\delta_\epsilon = \frac{\epsilon}{12mK}$ in step 2 so that $g_\mu(\tilde{t}) - g_\mu(t^*) \le \frac{\epsilon}{12m}$ where $g_\mu(x) = p(|x|) + \mu \cdot \ell_1(x)$. Denote $f(x) = \mu \cdot \sum_{i=1}^{3m} \sum_{t=1}^{k_1} \ell\left(\sum_{j=1}^{m} x_{ij}, b_{1t} \right) + \nu \cdot \sum_{j=1}^{m} \sum_{t=1}^{k_2} \ell\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij}, b_{2t} \right)$ and $q(x) = \sum_{i=1}^{3m} \sum_{j=1}^{m} p\left(|x_{ij}|\right)$.

2. Choose $r = \left\lceil \left(\frac{4(3m\cdot\mu\cdot k_1 + m\cdot\nu\cdot k_2)^{c_1}(3m^2)^{c_2}}{\epsilon/\kappa} \right)^{1/(1-c_1-c_2)} \right\rceil$ where c_1 and c_2 are two arbitrary constants that $c_1 + c_2 < 1$. Note that $\kappa n^{c_1} d^{c_2} = \frac{\epsilon r}{4}$ by our choice of r. Construct the following instance of Problem 2:

$$\min_{x^{(1)}, x^{(2)}, \ldots, x^{(r)} \in \mathbb{R}^{3m \times m}} \sum_{i=1}^{r} f(x^{(i)}) \quad \text{s.t.} \quad \sum_{i=1}^{r} q(x^{(i)}) \le \bar{K}, \tag{14}$$

   where $\bar{K} \in [3m \cdot r \cdot p(\tilde{t}), 3m \cdot r \cdot p(\tilde{t}) + \epsilon r/4)$. The coding size of $\bar{K}$ is bounded by a polynomial function of m because $\epsilon r/4$ and $3m \cdot r \cdot p(\tilde{t})$ are both bounded by a polynomial functions of n. Denote the minimizer of the minimization problem (14) to be $x_{\bar{K}}^{(1)}, \cdots x_{\bar{K}}^{(r)}$.

3. Let $\Phi_1 = 3m \cdot p(|\tilde{t}|) + \mu \cdot 3m \cdot \ell_1(\tilde{t}) - \frac{\epsilon}{4}$ and $\Phi_2 = \nu \cdot m \cdot \ell_2(\tilde{t})$. We claim that if the 3-partition problem has a solution, then

   (i) $\sum_{i=1}^{r} f(x_{\bar{K}}^{(i)}) + \sum_{i=1}^{r} q(x_{\bar{K}}^{(i)}) \le r\Phi_1 + r\Phi_2 + \frac{\epsilon r}{2}$.

   (ii) $\sum_{i=1}^{r} q(x_{\bar{K}}^{(i)}) \ge 3m \cdot r \cdot p(\tilde{t}) - \epsilon r/4$.

4. Suppose we have approximate solutions satisfying $\sum_{i=1}^{r} f(\hat{x}_{\bar{K}+\kappa n^{c_1} d^{c_2}}^{(i)}) \le \sum_{i=1}^{r} f(x_{\bar{K}}^{(i)})$, we claim that

(iii) If the approximate solutions satisfy

$$\sum_{i=1}^{r} f(\hat{x}^{(i)}_{\bar{K}+\kappa n^{c_1} d^{c_2}}) + \sum_{i=1}^{r} q(\hat{x}^{(i)}_{\bar{K}+\kappa n^{c_1} d^{c_2}}) \leq r\Phi_1 + r\Phi_2 + 4\kappa n^{c_1} d^{c_2}, \tag{15}$$

we can choose one index k such that $\Phi_1 + \Phi_2 + \epsilon \geq f(\hat{x}^{(k)}_{\bar{K}+\kappa n^{c_1} d^{c_2}}) \geq \Phi_1 + \Phi_2$ and obtain a feasible assignment: If $\left(\hat{x}^{(k)}_{\bar{K}+\kappa n^{c_1} d^{c_2}}\right)_{ij} \in B(\tilde{t}, \delta)$, we assign number i to subset j. If the approximate solutions do not satisfy (15), the 3-partition problem has no feasible solution.

We begin with the proof of (i). By the condition that the 3-partition problem has a solution, we construct $x^* \in \mathbb{R}^{3m \times m}$ as follows. If number i is assigned to subset j, let $x^*_{ij} = \tilde{t}$ and $x^*_{ik} = 0$ otherwise. We can see that $x^{(1)} = \cdots = x^{(r)} = x^*$ satisfy the constraint of (14) with sparsity level $\bar{K}$ and $\sum_{i=1}^{r} q(x^{(i)}_{\bar{K}}) \leq 3m \cdot r \cdot p(\tilde{t}) + \epsilon r/4$. Thus,

$$\begin{aligned} \sum_{i=1}^{r} f(x^{(i)}_{\bar{K}}) + \sum_{i=1}^{r} q(x^{(i)}_{\bar{K}}) &\leq r \cdot f(x^*) + 3m \cdot r \cdot p(\tilde{t}) + \frac{\epsilon r}{4} = r \cdot 3m \cdot g_\mu(\tilde{t}) + r\Phi_2 + \frac{\epsilon r}{4} \\ &= r \cdot 3m \cdot \left(g_\mu(\tilde{t}) - \frac{\epsilon}{12m}\right) + r\Phi_2 + \frac{\epsilon r}{2} = r\Phi_1 + r\Phi_2 + \frac{\epsilon r}{2}, \end{aligned} \tag{16}$$

where $g_\mu(x) = p(|x|) + \mu \cdot \ell_1(x)$. To prove (ii), we just need to notice that if $\sum_{i=1}^{r} q(x^{(i)}_{\bar{K}}) < 3m \cdot r \cdot p(\tilde{t}) - \epsilon r/4$, we would have $\sum_{i=1}^{r} f(x^{(i)}_{\bar{K}}) + \sum_{i=1}^{r} q(x^{(i)}_{\bar{K}}) < r\Phi_1 + r\Phi_2$ by the same reasoning of equation (16), yielding a contradiction as $\sum_{i=1}^{r} f(x^{(i)}_{\bar{K}}) + \sum_{i=1}^{r} q(x^{(i)}_{\bar{K}})$ will always be greater than or equal to $r\Phi_1 + r\Phi_2$.

Now we prove (iii). To prove the first half of the claim, we only need to use $4\kappa n^{c_1} d^{c_2} = r\epsilon$ and apply the proof of Theorem 1 to get the result. To prove the second half of the claim, assume that we have an algorithm that outputs $\hat{x}_{\bar{K}+\kappa n^{c_1} d^{c_2}}$ satisfying $\sum_{i=1}^{r} f(\hat{x}^{(i)}_{\bar{K}+\kappa n^{c_1} d^{c_2}}) \leq \sum_{i=1}^{r} f(x^{(i)}_{\bar{K}})$. Suppose that the 3-partition problem has a solution. Then we have

$$\sum_{i=1}^{r} f(\hat{x}^{(i)}_{\bar{K}+\kappa n^{c_1} d^{c_2}}) + \sum_{i=1}^{r} q(\hat{x}^{(i)}_{\bar{K}+\kappa n^{c_1} d^{c_2}}) \leq \sum_{i=1}^{r} f(x^{(i)}_{\bar{K}}) + \sum_{i=1}^{r} q(x^{(i)}_{\bar{K}}) + \frac{\epsilon r}{2} \leq r\Phi_1 + r\Phi_2 + \epsilon r,$$

where the first inequality is due to (ii) and the second inequality is due to (i). It means that the approximate solutions satisfy (15). To this point, we have finished the proof of Theorem 2. ■

Remark. Note that K is the input of Problem 2, which means that for all ℓ and p, we only need to find a K that makes Problem 2 hard to solve. A natural question to ask is whether or not K could be the parameter of Problem 2 such that the hardness result still holds. Unfortunately, we have the following counterexample. Assume p is L_0 norm and $K = 1$. Then the constraint is $\sum_{j=1}^{d} L_0(x_j) \leq 1$, which means that there is at most one component of x that is not equal to 0. In this case, we could solve the optimization problem in polynomial time by searching for the nonzero component of x.

6.4. Proof of Theorem 3

Using a similar argument, we can prove the last part of our main result.

Proof The proof is analogous to the proof of Theorem 2. We will use a reduction from the 3-partition problem to prove the theorem. The reduction is developed through the following steps.

1. Given the input to the 3-partition problem, we conduct the first reduction step of Theorem 2 to compute $\mu, \nu, \tilde{t}$ and ϵ with $\lambda = 1$. Let $f(x) = \mu \cdot \sum_{i=1}^{3m} \sum_{t=1}^{k_1} \ell\left(\sum_{j=1}^{m} x_{ij}, b_{1t}\right) + \nu \cdot \sum_{j=1}^{m} \sum_{t=1}^{k_2} \ell\left(\sum_{i=1}^{3m} \frac{s_i}{\sum_{i'=1}^{3m} s_{i'}/m} x_{ij}, b_{2t}\right)$ and $q(x) = \sum_{i=1}^{3m} \sum_{j=1}^{m} p\left(|x_{ij}|\right)$.

2. Choose $r = \left\lceil \left(\frac{4(3m \cdot \mu \cdot k_1 + m \cdot \nu \cdot k_2)^{c_1} (3m^2)^{c_2}}{\epsilon/\kappa} \right)^{1/(1-c_1-c_2)} \right\rceil$ where c_1 and c_2 are two arbitrary constants that $c_1 + c_2 < 1$. Note that $\kappa n^{c_1} d^{c_2} = \frac{\epsilon r}{4}$ by our choice of r. Construct the following instance of Problem 3:

$$\min_{x^{(1)}, x^{(2)}, \ldots, x^{(r)} \in \mathbb{R}^{3m \times m}} \sum_{i=1}^{r} q(x^{(i)}) \quad \text{s.t.} \quad \sum_{i=1}^{r} f(x^{(i)}) \leq \bar{\eta}, \tag{17}$$

   where $\bar{\eta} \in [\mu \cdot 3m \cdot \ell_1(\tilde{t}) + \nu \cdot m \cdot \ell_2(\tilde{t}), \mu \cdot 3m \cdot \ell_1(\tilde{t}) + \nu \cdot m \cdot \ell_2(\tilde{t}) + \epsilon r/4)$. Note that the parameters $\mu, \nu, \delta, m, r, d$ and $\bar{\eta}$ are bounded by polynomial functions of n. Computing their values also takes polynomial time. Given the sparsity level $\bar{\eta}$, denote the minimizer of (17) to be $x_{\bar{\eta}}^{(1)}, \cdots x_{\bar{\eta}}^{(r)}$.

3. Let $\Phi_1 = 3m \cdot p(|\tilde{t}|) + \mu \cdot 3m \cdot \ell_1(\tilde{t}) - \frac{\epsilon}{4}$ and $\Phi_2 = \nu \cdot m \cdot \ell_2(\tilde{t})$. We claim that if the 3-partition problem has a solution, then

   (i) $\sum_{i=1}^{r} f(x_{\bar{\eta}}^{(i)}) + \sum_{i=1}^{r} q(x_{\bar{\eta}}^{(i)}) \leq r\Phi_1 + r\Phi_2 + \frac{\epsilon r}{2}$.

   (ii) $\sum_{i=1}^{r} f(x_{\bar{\eta}}^{(i)}) \geq \mu \cdot 3m \cdot \ell_1(\tilde{t}) + \nu \cdot m \cdot \ell_2(\tilde{t}) - \epsilon r/4$.

4. Suppose we have an approximate solution satisfying $\sum_{i=1}^{r} f(\hat{x}_{\bar{\eta}+\kappa n^{c_1} d^{c_2}}^{(i)}) \leq \sum_{i=1}^{r} f(x_{\bar{\eta}}^{(i)})$, we claim that

   (iii) If the approximate solution satisfies

$$\sum_{i=1}^{r} f(\hat{x}_{\bar{\eta}+\kappa n^{c_1} d^{c_2}}^{(i)}) + \sum_{i=1}^{r} q(\hat{x}_{\bar{\eta}+\kappa n^{c_1} d^{c_2}}^{(i)}) \leq r\Phi_1 + r\Phi_2 + 4\kappa n^{c_1} d^{c_2}, \tag{18}$$

   we can choose the index k such that $\Phi_1 + \Phi_2 + \epsilon \geq f(\hat{x}_{\bar{\eta}+\kappa n^{c_1} d^{c_2}}^{(k)}) \geq \Phi_1 + \Phi_2$ and obtain a feasible assignment: If $\left(\hat{x}_{\bar{\eta}+\kappa n^{c_1} d^{c_2}}^{(k)}\right)_{ij} \in B(\tilde{t}, \delta)$, we assign number i to subset j. If the approximate solutions do not satisfy (18), the 3-partition problem has no feasible solution.

We have constructed a polynomial reduction from the 3-partition problem to finding an approximate solution to Problem 3. We then prove that the reduction works. We begin

with the proof of (i). By the condition that the 3-partition problem has a solution, we construct $x^* \in \mathbb{R}^{3m\times m}$ as follows. If number i is assigned to subset j, let $x^*_{ij} = \tilde{t}$ and $x^*_{ik} = 0$ otherwise. We can see that $x^{(1)} = \cdots = x^{(r)} = x^*$ satisfy the constraint of (14) with error tolerance $\bar{\eta}$ and $\sum_{i=1}^r f(x^{(i)}_{\bar{\eta}}) \le \mu \cdot 3m \cdot \ell_1(\tilde{t}) + \nu \cdot m \cdot \ell_2(\tilde{t}) + \epsilon r/4$. Thus,

$$\begin{aligned}\sum_{i=1}^r f(x^{(i)}_{\bar{\eta}}) + \sum_{i=1}^r q(x^{(i)}_{\bar{\eta}}) &\le r \cdot q(x^*) + \mu \cdot 3m \cdot \ell_1(\tilde{t}) + \nu \cdot m \cdot \ell_2(\tilde{t}) + \epsilon r/4 \\ &\le r \cdot 3m \cdot \Big(g_\mu(t^*) - \frac{\epsilon}{12m}\Big) + r\Phi_2 + \frac{\epsilon r}{2} = r\Phi_1 + r\Phi_2 + \frac{\epsilon r}{2},\end{aligned} \tag{19}$$

where $g_\mu(x) = p(|x|) + \mu \cdot \ell_1(x)$ and the last inequality is due to the choice of $\tilde{t}$ in step 1 of the reduction.

To prove (ii), we just need to notice that if $\sum_{i=1}^r f(x^{(i)}_{\bar{\eta}}) \le \mu\cdot 3m\cdot\ell_1(\tilde{t})+\nu\cdot m\cdot\ell_2(\tilde{t})-\epsilon r/4$, we would have $\sum_{i=1}^r f(x^{(i)}_{\bar{\eta}}) + \sum_{i=1}^r q(x^{(i)}_{\bar{\eta}}) < r\Phi_1 + r\Phi_2$ by the same reasoning of equation (16), yielding a contradiction as $\sum_{i=1}^r f(x^{(i)}_{\bar{\eta}}) + \sum_{i=1}^r q(x^{(i)}_{\bar{\eta}})$ will always be greater than or equal to $r\Phi_1 + r\Phi_2$.

Now we prove (iii). To prove the first half of the claim, we only need to use $4\kappa n^{c_1} d^{c_2} = r\epsilon$ and apply the proof of Theorem 1 to get the result. To prove the second half of the claim, assume that we have an algorithm that outputs $\hat{x}_\eta$ satisfying $\sum_{i=1}^r f(\hat{x}^{(i)}_{\eta+\kappa n^{c_1} d^{c_2}}) \le \sum_{i=1}^r f(x^{(i)}_\eta)$. Suppose that the 3-partition problem has a solution. Replacing η by $\bar{\eta}$ gives us

$$\sum_{i=1}^r f(\hat{x}^{(i)}_{\bar{\eta}+\kappa n^{c_1} d^{c_2}}) + \sum_{i=1}^r q(\hat{x}^{(i)}_{\bar{\eta}+\kappa n^{c_1} d^{c_2}}) \le \sum_{i=1}^r f(x^{(i)}_{\bar{\eta}}) + \sum_{i=1}^r q(x^{(i)}_{\bar{\eta}}) + \frac{\epsilon r}{2} \le r\Phi_1 + r\Phi_2 + \epsilon r,$$

where the first inequality is due to (ii) and the second inequality is due to (i). It means that the approximate solutions satisfy (18). To this point, we have finished the proof of Theorem 3. ■

References

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

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

S. Arora, L. Babai, J. 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.

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

A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

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

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

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

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

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

Y. Chen, D. Ge, M. Wang, Z. Wang, Y. Ye, and H. Yin. Strong NP-hardness for sparse optimization with concave penalty functions. *ICML*, 2017.

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

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

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

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

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

E. X. Fang, H. Liu, and M. Wang. Blessing of massive scale: Spatial graphical model inference with a total cardinality constraint. *working paper*, 2015.

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

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

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

D. Ge, Z. Wang, Y. Ye, and H. Yin. Strong NP-hardness result for regularized L_q-minimization problems with concave penalty functions. *arXiv preprint arXiv:1501.00622*, 2015.

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

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

P.-L. Loh and M. 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.

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

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

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

X. Shen, W. Pan, and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

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

V. V. Vazirani. *Approximation Algorithms*. Springer Science & Business Media, 2001.

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

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

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

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