

Hardness on Numerical Realization of Some Penalized Likelihood Estimators

Xiaoming Huo

*School of Industrial & Systems Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332-0205, USA
e-mail: xiaoming@isye.gatech.edu*

and

Jie Chen

*School of Industrial & Systems Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332-0205, USA
e-mail: jchen@isye.gatech.edu*

Georgia Institute of Technology

Abstract: We show that with a class of penalty functions, numerical problems associated with the implementation of the penalized least square estimators are equivalent to the exact cover by 3-sets problem, which belongs to a class of NP-hard problems. We then extend this NP-hardness result to the cases of penalized least absolute deviation regression and penalized support vector machines. We discuss the practical implication of our results. In particular, we emphasize that the oracle property of a penalized likelihood estimator requires a local extremum, instead of a global extremum. Hence the penalized likelihood estimators are still favorable; however one should not attempt to find its global extremum(a)!

AMS 2000 subject classifications: Primary 62J99; secondary 62F99.

Keywords and phrases: Penalized least square, NP-hardness, least absolute deviation regression, SCAD, penalized support vector machine.

1. Introduction

Penalized least square estimators are well presented and advocated in the statistics literature, see, e.g., [1, 4, 5]. We show that for several existing types of penalty functions, the corresponding penalized least square problems are equivalent to the exact cover by 3-sets problem, which belongs to an NP-hard class [8]. No polynomial-time numerical solution is available by now. We then extend the NP-hardness results to penalized least absolute

deviation regression and a problem that is derived from penalized support vector machines. The proof of NP-hardness is preceded by [12] and some recent works [11, 13]. However, the proofs in this paper require much more involved techniques.

Our NP-hardness result does not oppose the principle of penalized likelihood estimators. Instead, our results forewarn a misuse of penalized likelihood estimators: i.e., one should not attempt to find the *global* extremum(a) in numerical implementations. It was proven by Fan and Li [4] that the oracle property of a penalized likelihood estimator just requires a *local* extremum. The correct way to utilizing penalized likelihood estimator is the following: starting with a consistent estimator (e.g., the maximum likelihood estimator), then modifying it via optimizing the penalized likelihood function locally.

The rest of the paper is organized as follows. Section 2 presents a formulation of the penalized least square estimation. Section 3 describes some well-adopted penalty functions and some known NP-hardness results. Section 4 establishes the NP-hardness for penalized least square estimators in a general way. Specific results regarding each class of penalty functions are given in corollaries. Section 5 extends the NP-hardness to regression with least absolute deviations. Section 6 extends the NP-hardness to penalized support vector machines. Section 7 discuss other possible penalized likelihood estimators and conjecture on their NP-hardness. Section 8 gives some literature points on the oracle property of penalized likelihood estimators, emphasizing their requirement of local extremum (instead of a global extremum). We discuss the practical implication of the NP-hardness of the penalized likelihood estimators in Section 8.

2. Problem Formulation

Consider linear regression model, $y = \Phi x + \varepsilon$, where we have $y \in \mathbb{R}^m$, $\Phi \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$. Vectors y , x , and ε are called responses, coefficients, and random errors respectively. Matrix Φ is called the model matrix. The penalized least square estimator is the solution to the following optimization problem:

$$(PLS) \quad \min_x \quad \|y - \Phi x\|_2^2 + \lambda_0 \sum_{i=1}^n p(|x_i|),$$

where term $\|\cdot\|_2^2$ corresponds to the residual sum of squares, λ_0 is a prescribed algorithmic parameter, penalty function $p(\cdot)$ maps nonnegative value

to nonnegative value ($p : \mathbb{R}^+ \rightarrow \mathbb{R}^+$), and x_i is the i th entry of the coefficient vector x .

3. Penalty Functions and Known NP-hardness Results

Some choices of p are listed below. We always assume $x \geq 0$.

- ℓ_0 penalty: $p(x) = I(x \neq 0)$, where $I(\cdot)$ is the indicator function.
- ℓ_1 penalty: $p(x) = |x|$; Lasso [15] and its variants utilize this penalty function.
- Ridge regression: $p(x) = x^2$.
- More generally, for $0 < c < 2$, bridge regression [7] takes $p(x) = x^c$.
- Hard-threshold penalty [3]: $p(x; \lambda) = \lambda^2 - [(\lambda - |x|)_+]^2$, where λ ($\lambda > 0$) is another algorithmic parameter. This penalty function is smoother than the ℓ_0 penalty function.
- Nikolova penalty [14]: $p(x) = \frac{x}{1+x}$.
- Finally, the smoothly clipped absolute deviation (SCAD) penalty [4]: for $\lambda > 0, a > 1$,

$$p(x) = \begin{cases} \lambda x, & \text{if } 0 \leq x < \lambda; \\ -(x^2 - 2a\lambda x + \lambda^2)/[2(a-1)], & \text{if } \lambda \leq x < a\lambda; \\ (a+1)\lambda^2/2, & \text{if } x \geq a\lambda. \end{cases}$$

As one can see, penalized least square covers many problems in model selection and estimation. It is well-known that when the model matrix Φ is orthogonal, the solutions to the above problems are trivial: just apply some univariate operators. It is shown in [11, 13] that for generic model matrix Φ , when the ℓ_0 penalty is chosen, the problem is NP-hard. The proof of Huo and Ni [11] utilizes the result of Natarajan [12], which states that sparse approximate solutions (SAS) to a linear system are equivalent to the exact cover by 3-sets (X3C) problem, which is known to be NP-hard [8]. Huo and Ni [11] applied the principle of Lagrange multiplier to establish the relation between the ℓ_0 penalized PLS and SAS; then they proved the NP-hardness. It remained open whether a more general PLS is NP-hard.

4. General NP-Hardness for PLS Estimators

In this paper, we establish the following theorem.

Theorem 4.1 (NP-Hardness of PLS). *For general model matrix Φ , problem (PLS) is NP-hard if the penalty function $p(\cdot)$ ($p : \mathbb{R}^+ \rightarrow \mathbb{R}^+$) satisfies the following four conditions.*

C1. $p(0) = 0$ and function $p(x), x \geq 0$, is monotone increasing: $\forall 0 \leq x_1 < x_2, p(x_1) \leq p(x_2)$.

C2. There exists $\tau_0 > 0$ and a constant $c > 0$, such that $\forall 0 \leq x < \tau_0$, we have

$$p(x) \geq p(\tau_0) - c(\tau_0 - x)^2.$$

C3. For the aforementioned τ_0 , if $x_1, x_2 < \tau_0$, then $p(x_1) + p(x_2) \geq p(x_1 + x_2)$.

C4. $\forall 0 \leq x < \tau_0, p(x) + p(\tau_0 - x) > p(\tau_0)$.

A proof is given in Appendix A. Note in most cases of PLS, we have $p(0) = 0$ and function $p(\cdot)$ is monotone increasing. Hence C1 is satisfied. Condition C3 is satisfied if function $p(x)$ is concave in $[0, 2\tau_0]$. Condition C4 holds if function $p(x)$ is strictly concave for point 0 and point τ_0 . See Appendix B for a brief justification.

For the penalty functions in ℓ_0 penalty, bridge regression with $0 < c < 1$, Hard-threshold, Nikolova penalty, and SCAD, one can easily see that C3 and C4 hold.

Condition C2 is less intuitive. However, it is important to ensure the NP-hardness. Recall that in Fan and Li [4], $p'(|x|) = 0$ for large $|x|$ is a sufficient condition for the unbiasedness of a PLS estimator. On the other hand, if $p'(|x|) = 0$ for $|x|$ larger than a positive value, it is possible to find a quadratic function, $y = p(\tau_0) - c(\tau_0 - x)^2$, which is below penalty function $y = p(x)$ for $0 \leq x < \tau_0$ with positive constants τ_0 and c . For ℓ_0 , hard-threshold, and SCAD penalties, one can verify C2 by taking the following values for τ_0 and c .

- For the ℓ_0 penalty, one takes $\tau_0 = 1$ and $c = 1$.
- For the hard-threshold penalty, one may take $\tau_0 = \lambda$ and $c = 1$.
- For the SCAD penalty, one takes $\tau_0 = a\lambda$ and $c = \frac{1}{2(a-1)}$.

From the above, we immediately have the following.

Corollary 4.2. *For the penalties in ℓ_0 , hard-threshold, and SCAD, the implementation of PLS lead to NP-hard problems.*

Note that the result of NP-hardness in Huo and Ni [11] becomes a special case.

It is well known that the ℓ_1 penalty leads to linear programming problems. It is also well known that ridge regression and bridge regression with $c \geq 1$ lead to convex optimization problems, hence they have polynomial time solutions.

Solely based on Theorem 4.1, we can not establish the NP-hardness for the PLS problem with Nikolova penalty function or bridge regression with $0 < c < 1$. One can not establish C2 for the Nikolova penalty; neither can we for the bridge regression with $0 < c < 1$. The derivatives of both penalty functions converge to zero as the variable goes to the positive infinity. In Appendix C, it is shown that if C2 holds, then $p'(\tau_0) = 0$.

For bridge regression with $0 < c < 1$, we conjecture that the related PLS problem is NP-hard. However, we do not establish a proof in the present paper. In part, it is found that $p'(x) = c \cdot x^{c-1} \rightarrow +\infty$, as $x \rightarrow 0$. In all the cases that we have proved so far, we have $p'(x)$ upper bounded in interval $[0, \tau_0)$. At the same time, the gradient of the objective function in (PLS) becomes instable (going to infinities) as some elements of x converge to 0. Hence we believe it is hard to numerically realize a bridge regression with $0 < c < 1$. Furthermore, it is shown in [1] and [4] that such a bridge regression is *not* continuous.

The following theorem will be used to establish the NP-hardness related to the Nikolova penalty. A proof is given in Appendix D.

Theorem 4.3. *Assume the model matrix Φ is of full row rank. For continuous penalty function $p(x)$ that satisfies condition C1 and is strictly concave within interval $(0, \infty)$. Suppose penalty function $p(x)$ satisfies the Lipschitz condition: there exists a constant $C_1 > 0$ such that $|p(x_1) - p(x_2)| \leq C_1|x_1 - x_2|$ for any $0 < x_1, x_2 < \infty$. Then the corresponding PLS problem is NP-hard.*

For the PLS estimators with Nikolova penalty, it is observed that

$$p'(x) = (1+x)^{-2} \Rightarrow 0 < p'(x) \leq 1, \forall x \in [0, +\infty).$$

Evidently, the Lipschitz condition holds. We immediately have the following.

Corollary 4.4. *Assume the model matrix Φ is of full row rank. For a PLS estimator with Nikolova penalty, the corresponding optimization problem is NP-hard.*

5. Least Absolute Deviation Regression

It is interesting to know that when the quadratic term $\|y - \Phi x\|_2^2$ in (PLS) is replaced by a sum of the absolute values of the residuals (i.e., $\|y - \Phi x\|_1$), for several penalty functions, the corresponding optimization problems are NP-hard. The proof of NP-hardness is nearly identical with the proof of Theorem 4.1. Recall that these problems are associated with the least absolute deviations (LAD) regression [2, 9].

We consider

$$(PLAD) \quad \min_x \|y - \Phi x\|_1 + \lambda_0 \sum_{i=1}^n p(|x_i|),$$

where all the notations are predefined. We have the following theorem.

Theorem 5.1 (NP-hardness for Penalized LAD). *For general model matrix Φ , problem (PLAD) is NP-hard if the penalty function $p(\cdot)$ ($p : \mathbb{R}^+ \rightarrow \mathbb{R}^+$) satisfies the following three conditions:*

- D1. $p(0) = 0$ and function $p(x), x \geq 0$, is monotone increasing: $\forall 0 \leq x_1 < x_2, p(x_1) \leq p(x_2)$.
- D2. There exists a constant $\tau_0 > 0$, such that function $p(x)$ is concave in the interval $[0, 2\tau_0]$.
- D3. $\forall 0 \leq x < \tau_0, p(x) + p(\tau_0 - x) > p(\tau_0)$.

It is easy to verify that for the function $p(x)$ in ℓ_0 penalty, bridge regression with $0 < c < 1$, hard-threshold, Nikolova penalty, and SCAD, the conditions in the above theorem are satisfied. Hence we immediately have the following.

Corollary 5.2. *If the penalty function is chosen according to the ℓ_0 penalty, bridge regression with $0 < c < 1$, hard-threshold, Nikolova penalty, or SCAD, the resulting problem as in (PLAD) is NP-hard.*

We explain why the proof of Theorem 5.1 will be an easy extension from the proof of Theorem 4.1. First of all, note that conditions D1 and D3 in Theorem 5.1 are identical with the conditions C1 and C4 in Theorem 4.1. Moreover, given D2, it is easy to verify that a condition like C3 is satisfied, referring to the discussion in Appendix B. Finally, given D2, for $0 < x < \tau_0$, we have

$$\begin{aligned} p(x) &\geq \frac{x}{\tau_0}p(\tau_0) + \frac{\tau_0 - x}{\tau_0}p(0) \\ &= \frac{x}{\tau_0}p(\tau_0) \\ &= p(\tau_0) - \frac{p(\tau_0)}{\tau_0}(\tau_0 - x); \end{aligned}$$

i.e., a condition like C2 is satisfied. As an exercise, readers can verify that the proof of Theorem 4.1 in Appendix A can be modified to prove the Theorem 5.1.

6. A Problem Related to Machine Learning and Data Mining

The following problem is rooted in machine learning and data mining [5, Section 6.4]. We consider

$$(PSVM) \quad \min_{\beta} \sum_{i=1}^n [1 - y_i(\mathbf{x}_i^T \beta)]_+ + \lambda_0 \sum_{j=1}^d p(|\beta_j|),$$

where (\mathbf{x}_i, y_i) , $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$, $i = 1, 2, \dots, n$, are training data; coefficient vector $\beta \in \mathbb{R}^d$ has elements β_j , $j = 1, 2, \dots, d$; function $[\cdot]_+$ corresponds to the hinge loss and only takes nonnegative value:

$$[x]_+ = \begin{cases} x, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0; \end{cases}$$

constant λ_0 is an algorithmic parameter; function $p(\cdot)$ is the aforementioned penalty function. In 1-norm support vector machine ([16] and references therein), we have $p(\beta) = |\beta|$; while in ordinary support vector machine, we have $p(\beta) = \beta^2$.

We will show that for a class of penalty function $p(\cdot)$, the problem (PSVM) is NP-hard. The proof of this NP-hardness result bears strong similarity with the proof of Theorem 4.1. However, it is not a direct extension. In the proof, several steps require slightly different treatments. We provide a complete proof in Appendix E.

Theorem 6.1 (Penalized Support Vector Machines). *For general training data (\mathbf{x}_i, y_i) , $i = 1, 2, \dots, n$, the problem (PSVM) is NP-hard if there exists a constant $\tau_0 > 0$, such that*

$$\lambda_0 \leq 3/p(\tau_0) \tag{1}$$

and (for this τ_0) the penalty function $p(\cdot)$ ($p: \mathbb{R}^+ \rightarrow \mathbb{R}^+$) satisfies the three conditions (D1-D3) in Theorem 5.1.

We now discuss when we can apply the above theorem. Recall the penalty functions that are described in Section 3.

- For the ℓ_0 penalty, due to the concave condition D3, we must choose $\tau_0 > 0$. Hence $p(\tau_0) \equiv 1$. Hence the above theorem only applies when $\lambda_0 \leq 3$.
- For the bridge regression with $0 < c < 1$, we can choose τ_0 to be any value in interval $(0, \infty)$. Hence $p(\tau_0)$ takes any value between 0 and $+\infty$. Hence Theorem 6.1 applies for any $\lambda_0 > 0$.

- For hard-threshold, one can choose $\tau_0 > 0$. Because $p(\tau_0)$ takes any value in interval $(0, \lambda^2)$, the above theorem applies for any $\lambda_0 > 0$.
- For the Nikolova penalty, one can choose $\tau_0 > 0$. The possible values of $p(\tau_0)$ form interval $(0, 1)$. Hence Theorem 6.1 applies for any $\lambda_0 > 0$.
- For SCAD, we must choose $\tau_0 > \lambda$. We have $\lambda^2 < p(\tau_0) \leq \frac{a+1}{2}\lambda^2$. Hence Theorem 6.1 applies when $\lambda_0 < 3/\lambda^2$.

Corollary 6.2. *For a penalty function and its corresponding domain of the parameter λ_0 that is specified in the foregoing list, the problem (PSVM) is NP-hard.*

7. Other Penalized Likelihood Estimators

One can see that when the penalty functions, $p_{\lambda_{t,j}}(\cdot)$, are identical, the sparse covariance matrix estimation problem [5, Equation (6.7)] is a PLS problem. Hence the NP-hardness result of PLS applies.

We have considered a subset of penalized likelihood estimators. There are other penalized likelihood estimators, e.g., a penalized likelihood estimator in logistic regression [5, Example 1] and a penalized likelihood estimator in Poisson log-linear regression [5, Example 2]. We conjecture that the corresponding optimization problems are NP-hard. However, this paper does not provide a proof.

8. Oracle Property and Local Minimizers

It would be mistaken to think that the NP-hardness results in this paper are against the use of penalized likelihood estimators. In fact, in [4], the authors have pointed out that the corresponding optimization problems are hard. Moreover, authors of [4] showed that a *local* extremum of the penalized likelihood possess the oracle property. The oracle property is one of the most interesting theoretical properties of a penalized likelihood estimator. The NP-hardness results of this paper simply demonstrate that it will be a misinterpretation of the penalized likelihood principle if someone tries to find the global extremum(a). The computational issue of penalized likelihood estimators are further discussed in [10], in which the authors correctly pointed out the difficulty of optimizing the penalized likelihood function.

The account of oracle property (for a local minimizer) in [4] is inspiring. Further technical analysis is reported in [6]. Another recent discovery on the oracle property of Lasso (in fact, a modified version) is reported in [17]. They have a common flavor: starting from a consistent estimator, adjust the

estimator by either optimizing the penalized likelihood function locally or modifying the weights in the penalty function of Lasso. It will be interesting to exploit the connection between these approaches. In particular, can we provide a unified condition on when such a local adjustment will lead to oracle property? We leave it as a topic of future research.

Appendix A: Proof of Theorem 4.1

It is known that the exact cover by 3-sets (X3C) is NP-hard [8]. Let S denote a set with m elements. Let C denote a collection of 3-element subsets of S . The X3C is [12]: Does C contain an exact cover for S ; i.e., a subcollection \hat{C} of C such that every element of S occurs exactly once in \hat{C} . Without loss of generality, we assume that m is divisible by 3; otherwise X3C can never be done.

Let $f(x)$ denote the objective function in (PLS); i.e., $f(x) = \|y - \Phi x\|_2^2 + \lambda_0 \sum_{i=1}^n p(|x_i|)$. We will show that for a pair of (Φ, y) , there exists a constant M , such that $f(x) \leq M$ if and only if there is a solution to X3C. Hence if (PLS) is not NP-hard, then X3C is not either; such a contradiction leads to the NP-hardness of (PLS).

We now construct Φ and y . The number of columns in Φ is equal to the number of subsets in C . Let ϕ_j ($1 \leq j \leq |C|$) denote the j th column of the matrix Φ . We assign: for $1 \leq i \leq m$, $(\phi_j)_i = \sqrt{c(\lambda_0 + 1)}/3$ if the i th element of S appears in the j th 3-subset in C ; and $(\phi_j)_i = 0$, otherwise. Here $(\phi_j)_i$ is the i th entry of vector ϕ_j . Apparently, we have $n = |C|$, the size of C . Let $y = \tau_0 \sqrt{c(\lambda_0 + 1)}/3 \cdot \mathbf{1}_{m \times 1}$, where $\mathbf{1}_{m \times 1}$ is an all-one vector.

Suppose X3C has a solution. We create vector x^* as: $(x^*)_i = \tau_0$, if the i th 3-element subset in C is used in the solution to X3C; and $(x^*)_i = 0$, otherwise. Here $(x^*)_i$ denotes the i th entry of vector x^* . One can easily verify that $y = \Phi x^*$. Hence we have $f(x^*) = \frac{m}{3} \lambda_0 p(\tau_0)$.

Now assign $M = \frac{m}{3} \lambda_0 p(\tau_0)$. We show that if there exists x' satisfying $f(x') \leq M$, then we must have x' to be a solution of the X3C problem. Recalling that x^* corresponds to a solution to X3C as described above, if the solution to the X3C problem is not unique, it is possible that x' and x^* are not identical.

For $1 \leq k \leq m$, Let Ω_k denote a set of indices (of C) corresponding to the nonzero entries in the k th row of matrix Φ . Given $y = \Phi x^*$, it is evident that there is exactly one $j \in \Omega_k$, such that $x_j^* = \tau_0$; while for other $j \in \Omega_k$, we should have $x_j^* = 0$. We will need the following lemma.

Lemma A.1. *Suppose $p(\cdot)$ satisfies condition C1-C4. For $1 \leq k \leq m$, the following strict inequality holds if at least one side of it is not equal to zero:*

$$\frac{1}{3} \sum_{i \in \Omega_k} [p(|x_i^*|) - p(|x'_i|)] < \lambda_0^{-1} \cdot \frac{(\lambda_0 + 1)c}{3} \left[\sum_{i \in \Omega_k} (x_i^* - x'_i) \right]^2. \quad (2)$$

Before giving the proof of the above lemma, we first introduce the following lemma which will be used in the proof of Lemma A.1 and the proofs of other theorems in the paper.

Lemma A.2. *If $\tau_0 \leq \sum_{i \in \Omega_k} |x'_i|$ and more than one x'_i are not equal to 0 for $i \in \Omega_k$, we have $p(\tau_0) < \sum_{i \in \Omega_k} p(|x'_i|)$.*

Proof of Lemma A.2. Let $c_1 = \sum_{i \in \Omega_k} |x'_i| \geq \tau_0$, we have

$$p(\tau_0) = p\left(\frac{\tau_0}{c_1} \sum_{i \in \Omega_k} |x'_i|\right) \stackrel{C3, C4}{<} \sum_{i \in \Omega_k} p\left(\frac{\tau_0}{c_1} |x'_i|\right) \leq \sum_{i \in \Omega_k} p(|x'_i|).$$

Hence, Lemma A.2 holds. \square

Proof of Lemma A.1. We prove this lemma in two cases.

In the first case, we suppose that the right hand side of (2) is nonzero. Hence the right hand side is always positive. The (2) holds trivially if the left hand side is nonpositive. If the left hand side of (2) is positive, we have

$$\begin{aligned} \frac{\frac{1}{3} \sum_{i \in \Omega_k} [p(|x_i^*|) - p(|x'_i|)]}{\frac{(\lambda_0 + 1)c}{3} [\sum_{i \in \Omega_k} (x_i^* - x'_i)]^2} &= \frac{p(\tau_0) - \sum_{i \in \Omega_k} p(|x'_i|)}{(\lambda_0 + 1)c(\tau_0 - \sum_{i \in \Omega_k} x'_i)^2} \\ &\stackrel{C3}{\leq} \frac{p(\tau_0) - p(\sum_{i \in \Omega_k} |x'_i|)}{(\lambda_0 + 1)c(\tau_0 - \sum_{i \in \Omega_k} x'_i)^2} \\ &\leq \frac{p(\tau_0) - p(\sum_{i \in \Omega_k} |x'_i|)}{(\lambda_0 + 1)c(\tau_0 - \sum_{i \in \Omega_k} |x'_i|)^2} \\ &\stackrel{C2}{\leq} (1 + \lambda_0)^{-1} < \lambda_0^{-1}. \end{aligned}$$

Note in all the three inequalities, we implicitly utilizes

$$p(\tau_0) > \sum_{i \in \Omega_k} p(|x'_i|). \quad (3)$$

By using Lemma A.2, we have $\tau_0 > \sum_{i \in \Omega_k} |x'_i|$. Then the first equality holds by condition C3, and the last inequality holds by condition C2.

In the second case, we assume that the right hand side of (2) is zero, we have $\tau_0 = \sum_{i \in \Omega_k} x'_i$. From the assumption of our lemma, the left hand side

can *not* be zero simultaneously. Condition C1 and Lemma A.2 demonstrate that the left hand side can only be negative: first, we have

$$p(\tau_0) = p(|\sum_{i \in \Omega_k} x'_i|) \stackrel{C1}{\leq} p(\sum_{i \in \Omega_k} |x'_i|);$$

Thus, $\tau_0 \leq \sum_{i \in \Omega_k} |x'_i|$. By Lemma A.2, it is easy to see (2) holds. \square

Given the definition of Ω_k , it is not hard to see that

$$\sum_{i=1}^n [p(|x_i^*|) - p(|x'_i|)] = \sum_{k=1}^m \frac{1}{3} \sum_{i \in \Omega_k} [p(|x_i^*|) - p(|x'_i|)]. \quad (4)$$

From the construction of Φ , it is not hard to verify the following:

$$\begin{aligned} \|y - \Phi x'\|_2^2 &= \|\Phi x^* - \Phi x'\|_2^2 \\ &= \|\Phi(x^* - x')\|_2^2 \\ &= \sum_{k=1}^m \frac{(\lambda_0 + 1)c}{3} [\sum_{i \in \Omega_k} (x_i^* - x'_i)]^2. \end{aligned} \quad (5)$$

Combine Lemma A.1 with (4) and (5), we have

$$\sum_{i=1}^n p(|x_i^*|) - \sum_{i=1}^n p(|x'_i|) \leq \lambda_0^{-1} \|y - \Phi x'\|_2^2;$$

and the equality holds if and only if the two sides of (2) are equal to zero for every k , $1 \leq k \leq m$. Note the above is equivalent to $M = f(x^*) \leq f(x')$. Recall $f(x') \leq M$, we must have $f(x') = M$ and $\forall k, p(\tau_0) = \sum_{i \in \Omega_k} p(|x'_i|)$ and $\tau_0 = \sum_{i \in \Omega_k} x'_i$. Utilizing Lemma A.2, one can show that within set $\{x'_i : i \in \Omega_k\}$, we must have exactly one element that is equal to τ_0 and the rest are zeros. Given the design of x^* , it is not hard to see that x' corresponds to another solution to X3C. (Note the solutions to X3C is not necessarily unique.)

From all the above, the theorem is proved.

Appendix B: Justifications Related to C3 and C4

Recall function $p(x)$ is concave in $[0, 2\tau_0]$. Hence for $x_1, x_2 < \tau_0$, we have, for $0 \leq \lambda \leq 1$,

$$p[\lambda x_1 + (1 - \lambda)x_2] \geq \lambda p(x_1) + (1 - \lambda)p(x_2),$$

Therefore, we have

$$p(x_i) \geq \frac{x_i}{x_1 + x_2} p(x_1 + x_2) + \frac{x_{\{3-i\}}}{x_1 + x_2} p(0), \quad i = 1, 2.$$

Adding the above two and using $p(0) = 0$, we have

$$p(x_1) + p(x_2) \geq p(x_1 + x_2).$$

The above is condition C3. The justification regarding C4 is nearly identical.

Appendix C: Proof of “First Derivative is Zero”

Recall $p(x)$, $x \geq 0$, is nondecreasing; hence $p'(x) \geq 0$. On the other hand, it is obvious that when C2 holds, we have

$$f(x) \triangleq p(x) - p(\tau_0) + c(\tau_0 - x)^2 \geq 0, \quad \text{for } x \geq \tau_0,$$

and $f(\tau_0) = 0$; hence $f'(\tau_0) \leq 0$, which leads to $p'(\tau_0) \leq 0$. From all the above, we have that $p'(\tau_0) = 0$.

Appendix D: Proof of Theorem 4.3

Let

$$f(x) = \|y - \Phi x\|_2^2 + \lambda_0 \sum_{i=1}^n p(|x_i|),$$

where $p(x)$ is the penalty function as in (PLS). Consider a truncated version of the penalty function:

$$p_t(x; N) = \begin{cases} p(x), & 0 \leq x \leq N, \\ p(N), & x > N, \end{cases}$$

where N is a positive constant. Correspondingly, we define

$$f_t(x; N) = \|y - \Phi x\|_2^2 + \lambda_0 \sum_{i=1}^n p_t(|x_i|; N).$$

Minimizing the $f(x)$ in \mathbb{R}^n is the original PLS optimization problem; while minimizing $f(x; N)$ with a prefixed N is its truncated version. Applying Theorem 4.1, it is not hard to see that the latter is NP-hard. We omit some obvious details here.

If we can show that for a constant N' that is large enough, the two problems have identical solutions, then we prove that the PLS problem with the original penalty is NP-hard.

We will show below that the solutions to the aforementioned problems are upper bounded by a constant; hence by choosing N' as this upper bound, the two problems are identical.

Let x_0 be the minimizer of either objective $f(x)$ or $f_t(x; N)$. Without loss of generality, assume x_0 is the minimizer of $f_t(x; N)$. Note the same argument will apply to objective $f(x)$ as well. We have $\forall a \in \mathbb{R}^n$,

$$f_t(x_0 + a; N) \geq f_t(x_0; N).$$

From the definition of $f_t(\cdot; N)$, we have

$$\|y - \Phi(x_0 + a)\|_2^2 + \lambda_0 \sum_{i=1}^n p_t[(x_0 + a)_i; N] \geq \|y - \Phi x_0\|_2^2 + \lambda_0 \sum_{i=1}^n p_t[(x_0)_i; N],$$

where $(\cdot)_i$ denotes the i th element of a vector. Simplifying the above, we have

$$\begin{aligned} & a^T \Phi^T \Phi a + 2(x_0^T \Phi^T \Phi - y^T \Phi) a \\ & + \lambda_0 \sum_{i=1}^n \{p_t[(x_0 + a)_i; N] - p_t[(x_0)_i; N]\} \geq 0. \end{aligned} \quad (6)$$

We will need the following inequality, which is presented in a lemma.

Lemma D.1. *For $1 \leq i \leq n$, we have*

$$|(\Phi^T \Phi x_0 - \Phi^T y)_i| \leq \frac{1}{2} \lambda_0 C_1,$$

where C_1 is the Lipschitz constant that is given in the theorem statement.

Proof. For $1 \leq i \leq n$, within vector a , we set every entry except a_i to be equal to 0. From (6), we have $\forall a_i$,

$$T_1 a_i^2 + 2T_2 a_i + \lambda_0 [p_t(|T_3 + a_i|; N) - p_t(|T_3|; N)] \geq 0, \quad (7)$$

where $T_1 = (\Phi^T \Phi)_{ii}$, $T_2 = (\Phi^T \Phi x_0 - \Phi^T y)_i$, and $T_3 = (x_0)_i$. Without loss of generality, in the following argument, we assume that $a_i > 0$. Readers can verify that a trivially modified argument holds when $a_i < 0$. Replacing a_i with $-a_i$ in (7), we have

$$T_1 a_i^2 - 2T_2 a_i + \lambda_0 [p_t(|T_3 - a_i|; N) - p_t(|T_3|; N)] \geq 0. \quad (8)$$

Given the definition of $p_t(\cdot, N)$ and the Lipschitz property of $p(\cdot)$, it is easy to see that function $p_t(\cdot; N)$ also satisfies the Lipschitz condition: for $a_i \neq 0$,

$$\left| \frac{p_t(|\alpha + a_i|; N) - p_t(|\alpha|; N)}{a_i} \right| < C_1,$$

where α is an arbitrary real number. From (7), we have

$$\begin{aligned} T_2 &\geq -\frac{1}{2}T_1a_i - \frac{1}{2}\lambda_0 \frac{p_t(|T_3 + a_i|; N) - p_t(|T_3|; N)}{a_i} \\ &\geq -\frac{1}{2}T_1a_i - \frac{1}{2}\lambda_0 C_1. \end{aligned} \quad (9)$$

Similarly from (8), we have

$$\begin{aligned} T_2 &\leq \frac{1}{2}T_1a_i + \frac{1}{2}\lambda_0 \frac{p_t(|T_3 - a_i|; N) - p_t(|T_3|; N)}{a_i} \\ &\leq \frac{1}{2}T_1a_i + \frac{1}{2}\lambda_0 C_1. \end{aligned} \quad (10)$$

Letting $a_i \rightarrow 0$, combining (9) and (10), we have $|T_2| \leq \frac{1}{2}\lambda_0 C_1$. \square

Let $v = \Phi^T \Phi x_0 - \Phi^T y$, the above leads to the following:

$$\begin{aligned} \|x_0\|_\infty &\leq \|x_0\|_2 \\ &\leq \|(\Phi^T \Phi)^{-1} \Phi^T y\|_2 + \sup_{\|v\|_\infty < \frac{1}{2}\lambda_0 C_1} \|(\Phi^T \Phi)^{-1} v\|_2 \\ &\leq \|(\Phi^T \Phi)^{-1} \Phi^T y\|_2 + \frac{\sqrt{n} \frac{1}{2} \lambda_0 C_1}{\mu_{\min}(\Phi^T \Phi)}. \end{aligned}$$

where $\mu_{\min}(\cdot)$ is the smallest eigenvalue of the matrix. Note the last term is a constant, which is determined by Φ , y , λ_0 , and C_1 . By taking N' to be the above constant, the above establishes the equivalence between the PLS problem with the original penalty function and the PLS problem with the truncated penalty function. Because the PLS problem with the truncated penalty function is NP-hard, we conclude that the PLS problem with the original penalty function is NP-hard as well. The theorem is proved.

Appendix E: Proof of Theorem 6.1

Similar to the proof in Appendix A, we will show that if problem (PSVM) is *not* NP-hard, neither is the exact cover by 3-sets problem. (Recall that the exact cover by 3-sets problem is denoted by X3C.) Because we know that X3C is NP-hard, so is the (PSVM). The proof is again constructive.

Let matrix $\Phi = \text{diag}(y_1, y_2, \dots, y_n)\mathbf{X}$, where $\text{diag}(y_1, y_2, \dots, y_n)$ is a diagonal matrix with diagonal entries y_1, y_2, \dots, y_n and that

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}.$$

Note $\Phi \in \mathbb{R}^{n \times d}$. For any given matrix Φ , one can find a (non-unique) set of data $(\mathbf{x}_i, y_i), i = 1, 2, \dots, n$, such that the above holds.

Let $f(\beta) = \|\mathbf{1}_n - \Phi\beta\|_+ + \lambda_0 \sum_{j=1}^d p(|\beta_j|)$, where $\mathbf{1}_n \in \mathbb{R}^{n \times 1}$ is an all-one vector, and for vector $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$, we have $\|\mathbf{x}\|_+ = \sum_{i=1}^n (x_i)_+$. It is evident that problem (PSVM) is equivalent to

$$\min_{\beta} f(\beta). \quad (11)$$

We now construct a matrix Φ . Let S be a set with n elements. Without loss of generality, we assume that n is divisible by 3. Let C denotes a collection of 3-element subsets of S . (Recall that the S and C are used in Appendix A.) Each column of matrix Φ (denoted by $\phi_k, 1 \leq k \leq |C|$) corresponds to a subset in collection C . Moreover, we have $(\phi_k)_i = 1/\tau_0$ if and only if the i th element of S is in the k th subset of C . Assume $\widehat{C} \subset C$ corresponds to an exact cover of S by 3-sets. For vector $\beta^* \in \mathbb{R}^{|C|}$, we have $\beta_k^* = \tau_0$ if and only if $k \in \widehat{C}$; the rest of β_k^* 's are equal to zero. Evidently, we have $\mathbf{1}_n = \Phi\beta^*$ and $f(\beta^*) = \lambda_0 \sum_{j=1}^{|C|} p(|\beta_j^*|) = \lambda_0 \frac{n}{3} p(\tau_0)$. We will show that $f(\beta^*)$ is a global minimum of $f(\beta)$. Moreover, any global solution of (11) corresponds to an exact cover by 3-sets. Hence the NP-hardness of X3C will lead to the NP-hardness of (11), and then (PSVM).

Let $\Omega_k, 1 \leq k \leq n$, to be the same subset of indices of C that is defined in Appendix A (right before Lemma A.1). For any other $\beta' \in \mathbb{R}^{|C|}$, we establish the following lemma.

Lemma E.1. *For any $k, 1 \leq k \leq n$, we have*

$$\frac{1}{3} \sum_{j \in \Omega_k} \lambda_0 [p(|\beta_j^*|) - p(|\beta_j'|)] \leq \frac{1}{\tau_0} \left\| \sum_{j \in \Omega_k} (\beta_j^* - \beta_j') \right\|_+; \quad (12)$$

and the inequality is strict unless both sides of the inequality are equal to zero.

Proof. We consider two cases:

- *Case 1*: when the right hand side of (12) is positive, and
- *Case 2*: when the right hand side of (12) is equal to zero.

Note the right hand side of (12) is nonnegative, the above two cases cover all possibilities.

In *case 1*, we have

$$\|\tau_0 - \sum_{j \in \Omega_k} \beta'_j\|_+ > 0 \Rightarrow \tau_0 > \sum_{j \in \Omega_k} \beta'_j.$$

We only need to consider the situation when for any $j \in \Omega_k$, we have $|\beta'_j| < \tau_0$; otherwise, the left hand side of (12) is nonpositive and the lemma trivially holds.

Using Lemma A.2, we can show that we only need to consider the case when any partial sum of quantities $|\beta'_j|$, $j \in \Omega_k$, must be less than τ_0 . Because otherwise, the left hand side of (12) is no more than zero; hence the lemma holds.

Note condition D3 is identical with condition C4. We will show that the following is true: $\forall 0 < x < \tau_0$,

$$p(x) > p(\tau_0) - \frac{p(\tau_0)}{\tau_0}(\tau_0 - x) = \frac{p(\tau_0)}{\tau_0}x. \quad (13)$$

To see the above, recall that at the end of Section 5, we have proved that when condition D2 holds, we have

$$p(x) \geq p(\tau_0) - \frac{p(\tau_0)}{\tau_0}(\tau_0 - x).$$

Without loss of generality, let us assume that $x < \frac{1}{2}\tau_0$. Recall condition D3 ($p(x) + p(\tau_0 - x) > p(\tau_0)$). If $p(x) = p(\tau_0) - \frac{p(\tau_0)}{\tau_0}(\tau_0 - x) = \frac{p(\tau_0)}{\tau_0}x$, we have

$$p(\tau_0 - x) > p(\tau_0) - \frac{p(\tau_0)}{\tau_0}x = \frac{p(\tau_0)}{\tau_0}(\tau_0 - x).$$

The three points $0 < x < \tau_0 - x$ form a counterexample of the concavity condition. Similarly, when $x > \frac{1}{2}\tau_0$, a counterexample will be found (with three points $\tau_0 - x < x < \tau_0$). The counterexample demonstrates that (13) holds.

Utilizing the above results, we have

$$\begin{aligned}
\frac{\text{left hand side of (12)}}{\text{right hand side of (12)}} &= \frac{\lambda_0 \tau_0 p(\tau_0) - \sum_{j \in \Omega_k} p(|\beta'_j|)}{3 \|\tau_0 - \sum_{j \in \Omega_k} \beta'_j\|_+} \\
&\stackrel{\text{C3}}{\leq} \frac{\lambda_0 \tau_0 p(\tau_0) - p(\sum_{j \in \Omega_k} |\beta'_j|)}{3 \tau_0 - \sum_{j \in \Omega_k} |\beta'_j|} \\
&\stackrel{(13)}{<} \frac{\lambda_0 \tau_0 p(\tau_0)}{3 \tau_0} \\
&\stackrel{(1)}{\leq} 1.
\end{aligned}$$

Hence (12) holds with strict inequality.

For *case 2*, we have $\tau_0 - \sum_{j \in \Omega_k} \beta'_j \leq 0$. Hence we have

$$\tau_0 \leq \sum_{j \in \Omega_k} \beta'_j \leq \sum_{j \in \Omega_k} |\beta'_j|.$$

Using Lemma A.2, we have

$$p(\tau_0) \leq \sum_{j \in \Omega_k} p(|\beta'_j|).$$

The above indicates that the left hand side of (12) is no more than zero. Hence (12) holds. \square

We show that β^* minimizes the function $f(\beta)$. Similar to the argument in Appendix A, we add up inequalities (12) for all $k, 1 \leq k \leq n$. We have

$$f(\beta^*) \leq f(\beta').$$

The above is true for every β' ; hence β^* is a minimizer.

Now we show that the minimum of function $f(\beta)$ is achieved when β corresponds to an exact cover by 3-sets. Suppose we have $f(\beta^*) = f(\beta')$. Based on Lemma E.1, both sides of inequality (12) must be equal to zero for any $k, 1 \leq k \leq n$. That is, $\forall 1 \leq k \leq n$, we have

$$p(\tau_0) = \sum_{j \in \Omega_k} p(|\beta'_j|), \quad (14)$$

and

$$\tau_0 \leq \sum_{j \in \Omega_k} \beta'_j. \quad (15)$$

From (15), we have $\tau_0 \leq \sum_{j \in \Omega_k} |\beta'_j|$. By Lemma A.2, we have $p(\tau_0) < \sum_{j \in \Omega_k} p(|\beta'_j|)$ if more than one β'_j are not equal to 0 for $j \in \Omega_k$. Thus, the

equality in (14) holds if and only if for each $k, 1 \leq k \leq n$, there is exactly one $j, j \in \Omega_k$, such that $\beta'_j \geq \tau_0$ and the rest of $\beta'_{j'}$'s ($j' \in \Omega_k$) are equal to zero. Evidently, the positions of nonzero entries of β' corresponds to another solution to X3C.

From all the above, we prove the theorem.

References

- [1] ANTONIADIS, A. AND FAN, J. (2001). Regularization of wavelets approximations (with discussion). *J. Amer. Statist. Assoc.* **96**, 455 (September), 939–967.
- [2] BLOOMFIELD, P. AND STEIGER, W. L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms*. Birkhäuser, Boston.
- [3] DONOHO, D. AND JOHNSTONE, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 3, 425–455.
- [4] FAN, J. AND LI, R. (2001). Variable selection via nonconvex penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 456 (December), 1348–1360.
- [5] FAN, J. AND LI, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the Madrid International Congress of Mathematicians*. To appear.
- [6] FAN, J. AND PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 3 (June), 928–961.
- [7] FRANK, I. E. AND FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 2, 109–148.
- [8] GAREY, M. R. AND JOHNSON, D. S. (1979). *Computers and intractability. A guide to the theory of NP-completeness*. W. H. Freeman and Co., San Francisco, CA, USA.
- [9] GENTLE, J. E. (1977). Least absolute values estimation: an introduction. *Comm. Statist. B* **6**, 313–328.
- [10] HUNTER, D. R. AND LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 4 (August), 1617–1642.
- [11] HUO, X. AND NI, X. S. (2005). When do stepwise algorithms meet subset selection criteria? ISyE Statistics Technical Report, URL = <http://www.isye.gatech.edu/statistics/papers/>.
- [12] NATARAJAN, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing* **24**, 2, 227–234.
- [13] NI, X. S. (2006). New results in detection, estimation, and model selection. Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA. <http://etd.gatech.edu/>.

- [14] NIKOLOVA, M. (2000). Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.* **61**, 2, 633–658.
- [15] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 1, 267–288.
- [16] ZHU, J., ROSSET, S., HASTIE, T., AND TIBSHIRANI, R. (2003). 1-norm support vector machines. In *Neural Information Processing Systems* Vol. **16**. .
- [17] ZOU, H. (2005). The adaptive Lasso and its oracle properties. Unpublished manuscript.