Journal of Global Optimization

A Unified Smooth Framework for Nonconvex Penalized Least Squares Problems

--Manuscript Draft--

Manuscript Number:	JOGO-D-20-00039
Full Title:	A Unified Smooth Framework for Nonconvex Penalized Least Squares Problems
Article Type:	Manuscript
Keywords:	Sparse optimization; nonconvex penalties; classical smooth methods; proximity operator; Moreau envelope
Corresponding Author:	Yongchao Yu Applied Mathematic Hennan, CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Applied Mathematic
Corresponding Author's Secondary Institution:	
First Author:	Yongchao Yu
First Author Secondary Information:	
Order of Authors:	Yongchao Yu
	Jigen Peng
Order of Authors Secondary Information:	
Funding Information:	the National Natural Science Foundation of China (61901404) — Dr. Yongchao Yu
	the National Natural Science Foundation of China (11771347) — Prof. Jigen Peng
	the Nanhu Scholars Program for Young Scholars of XYNU (the Nanhu Scholars Program for Young Scholars of XYNU) — Dr. Yongchao Yu
Abstract:	Due to the non-convexity and the non-smoothness of popular sparsity-promoting penalties, solving nonconvex penalized least squares problems is much more challenging. In particular, the non-smoothness of penalties prevents us from applying classical smooth methods to solve these models. In this paper, we introduce a unified smooth framework for nonconvex penalized least squares problems. We first show that most of popular penalties can be decomposed as the sum of twice-continuously differentiable concave functions and simply convex functions in the sense that their proximity operators have closed-form solutions, and then also propose a new penalty function which is a modified version of three well-known penalties. By utilizing special decomposition properties of penalties and the Moreau envelope technique, we prove that most of nonconvex penalized least squares problems can be equivalent to corresponding smooth unconstrained optimization problems in the sense that sets of globally optimal solutions, and optimal values of the original and the smooth problems are equal, respectively. Our approach is also extended to address other sparse models such as nonconvex sparse logistic regression models and nonconvex penalized matrix least squares models.

A Unified Smooth Framework for Nonconvex Penalized Least Squares Problems

Yongchao Yu* Jigen Peng†

January 12, 2020

Abstract

Due to the non-convexity and the non-smoothness of popular sparsity-promoting penalties, solving nonconvex penalized least squares problems is much more challenging. In particular, the non-smoothness of penalties prevents us from applying classical smooth methods to solve these models. In this paper, we introduce a unified smooth framework for nonconvex penalized least squares problems. We first show that most of popular penalties can be decomposed as the sum of twice-continuously differentiable concave functions and simply convex functions in the sense that their proximity operators have closed-form solutions, and then also propose a new penalty function which is a modified version of three well-known penalties. By utilizing special decomposition properties of penalties and the Moreau envelope technique, we prove that most of nonconvex penalized least squares problems can be equivalent to corresponding smooth unconstrained optimization problems in the sense that sets of globally optimal solutions, and optimal values of the original and the smooth problems are equal, respectively. Our approach is also extended to address other sparse models such as nonconvex sparse logistic regression models and nonconvex penalized matrix least squares models.

Keywords: Sparse optimization; nonconvex penalties; classical smooth methods; proximity operator; Moreau envelope

1 Introduction

Consider the linear model

$$b = Ax + e, \tag{1}$$

where $b \in \mathbb{R}^m$ is a vector, $A \in \mathbb{R}^{m\times n}$ is a matrix, $x \in \mathbb{R}^n$ is a vector of underlying parameters, coefficients or an unknown signal vector, and $e \in \mathbb{R}^m$ is a vector. This model (1) arises in many scientific research fields [1, 4, 5]. For instance, in statistical regression and machine learning [1, 2], b is the output data, A is often called a design or predictor matrix which collects the input data, x is a vector of regression parameters, and e is a random noise term. In the context of signal representation [5], b is a signal of interest, the matrix A represents an over-complete dictionary of elementary signals or atoms, the vector x contains representation coefficients of the signal b, and e denotes an approximation error. Moreover, in compressive sensing [4], the vector b collects the measured data, A is a measurement or sensing matrix, x represents a signal of interest, and e denotes a random noise vector. In the last decade, many researchers [1, 3, 4, 5, 6] from these fields have focused on the case in which n is much larger than m and the model is sparse in the sense that only a relatively small number of non-zero components of x are important and meaningful. Indeed, in statistical regression [1], determining a sparse parameter vector corresponds to

*School of Mathematics and Statistics, Xinyang Normal University, Xinyang, 464000, China. Email: sparseelad@126.com
†School of Mathematics and Information Sciences, Guangzhou University, 230 GuangZhou University City Outer Ring Road, GuangZhou, 510006, China. Email: jgpengxjtu@126.com

selecting a few relevant explanatory features; in signal representation field [5], one hopes that a signal can be represented as a sparse linear combination of atoms from a given over-complete dictionary; in compressive sensing [4], one desires to recover a sparse signal from a small number of linear and noisy measurements.

Therefore, sparse estimation has become a core issue in recent years. One popular technique to estimate a sparse solution to model (1) is by sparse penalization. Since the most natural measure of sparsity is the ℓ_0-'norm', one can consider the ℓ_0-penalized least squares problem [7]

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_0, \tag{2}$$

where $\lambda>0$ is a penalization parameter, $\|\cdot\|$ denotes the ℓ_2-norm (Euclidean norm), $\|x\|_0$ is the ℓ_0-norm of x and denotes the number of non-zero components of x. In the context of variable selection [7], solving problem (2) is related to selection of the best-subset, which is in general NP-hard and statistically troublesome because of the $\|\cdot\|_0$ penalty function being discrete and nonconvex. One widely used alternative to problem (2) is the ℓ_1-penalized least squares problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1, \tag{3}$$

where $\|x\|_1=\sum_{i=1}^n |x_i|$ denotes the ℓ_1-norm of x. Problem (2) is the well-known Lasso introduced in [8] in statistics and it is also called the Basis Pursuit Denoising problem [9] in signal and image processing.

A large number of theoretical works [2, 10] have shown that the Lasso enjoys attractive statistical properties and obtains great performance in prediction. More importantly, the Lasso is a convex optimization problem and thus can be easily solved by many efficient approaches such as, coordinate-descent method [5], quadratic programming approach [11], and proximal gradient method [12] or forward-backward splitting method [13]. However, it is known that Lasso requires rather stringent conditions on the design matrix to estimate underlying sparse parameters. Moreover, the ℓ_1 penalty [14, 15] related to Lasso tends to produce biased estimates for large parameters.

To circumvent the drawbacks of the ℓ_1 penalty, Fan and Li suggested to replace the ℓ_1 penalty with other penalty functions which lead to sparse and unbiased models. To this end, penalty functions should be singular at the origin for obtaining sparsity and their derivatives vanish for large values [14]. The first nonconvex penalty named the Smoothly Clipped Absolute Deviation (SCAD) penalty [14] satisfying above two conditions was then proposed by Fan and Li. Another two popular nonconvex penalties in statistical regression are the Minimax Concave Plus (MCP) penalty [16] and Capped-ℓ_1 [17] which also have interesting theoretical properties. Besides the three widely-studied penalties, there exist other nonconvex penalties in statistical regression and machine learning, such as the Exponential-Type Penalty (ETP) [15]. From an approximation point of view, nonconvex penalty functions with some parameters typically yield the tighter approximation to the ℓ_0-norm than the ℓ_1-norm. This fact also leads to more nonconvex surrogates in compressive sensing such as the ℓ_p pseudo-norm with $0<p<1$ [18, 19, 20, 21], the smoothed-ℓ_p [22, 23], the Transformed-ℓ_1 function [24], the fraction function [25], the logarithmic function [26]. These parameterized nonconvex penalties mentioned above have been empirically demonstrated to have better practical performance on various sparse estimation tasks.

The resulting least squares problems with nonconvex penalties have the general form

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2+\sum_{i=1}^n p_{\lambda,\tau}(|x_i|), \tag{4}$$

where $p_{\lambda,\tau}(|\cdot|)$ denotes a nonconvex penalty function which depends on the penalization parameter λ and the approximation parameter τ. Compared with the ℓ_1-penalized least squares problem (3), model (4) is computationally harder to solve due to nondifferentiability and nonconvexity of the penalty function.

We briefly review some methods available in the literature for solving model (4). We begin with some related works for addressing the SCAD penalty. Fan and Li [14] proposed a local quadratic approximation for the nonconvex penalty function and then used a single Newton step to solve the resulting objective function at each iteration. Zou and Li [27] in more general cases suggested to replace the local quadratic with a local linear approximation (LLA) for the penalty function. This leads to the linear local approximation estimator which can be viewed as a reweighted Lasso problem at each iteration. Since the method at each iteration needs to solve a convex problem, it may be often slow in practice. Another related technique for solving (4) is the iteratively reweighted-type algorithms such as the iteratively reweighted least squares algorithm (IRLS) [5] and the iteratively reweighted ℓ_1 algorithm (IRL1) [22]. The LLA method is a special case of IRL1. Moreover, Zhao and Li [22] also thoroughly studied the convergence of IRL1 for (4) with more penalties. However, the iteratively reweighted-type algorithms may be not very efficient because they also have to solve convex subproblems at each iteration.

The proximal gradient method has recently received considerable attention in sparse estimation because of its simplicity and efficiency. The popular method [12] was originally designed to minimize the sum of a convex smooth function f with Lipschitz continuous gradient and a simple (possibly nonsmooth) convex function g in the sense that its proximity operator is computationally inexpensive. The key building block of this method is to compute the proximity operator of g. The proximity operator of a general convex function is not easy to compute, however, for the simple ℓ_1-norm, its proximity operator is the well-known soft-thresholding operator. Applying the proximal gradient method to Lasso problem (3) yields the iterative shrinkage-thresholding algorithm. The method has also been directly extended to address the nonconvex penalized problem (4). For most of nonconvex penalties, their proximity operators also have closed-form solutions which are slightly more complex than the soft-thresholding operator. However, for some penalty functions such as the ℓ_p pseudo-norm ($p \neq 0, 1/2, 2/3, 1$) [28], the ETP, the smoothed-ℓ_p, closed-form solutions of proximity operators do not exist, which makes the proximal gradient method applied to (4) inefficient in practice. In addition, the proximal gradient method is a first-order method and thus it is usually slow. Furthermore, the accelerated proximal gradient method [12] for convex problems can not be directly applied to the nonconvex problem (4).

Recently, Lu *et al.* [29] have presented the proximal iteratively reweighted ℓ_1 algorithm with convergence guarantee for solving the general model (4) with all the aforementioned nonconvex penalties. The novel algorithm avoids solving a subproblem at each iteration as in IRL1 and computing the proximal operator of a nonconvex penalty as in the proximal gradient method. The algorithm is based on the key fact that all the existing nonconvex penalties are concave and monotonically increasing in the nonnegative real line, and thus their derivatives (or so-called supergradients of concave functions at nonsmooth points) are nonnegative and monotonically decreasing. They not only majorizes the nonconvex penalty in model (4) by a linear function in each iteration, but also majorizes the least squares part by a quadratic function. This algorithm obtains the next iterate by minimizing the sum of these two surrogate functions, which leads to a closed-form solution.

Another general approach [30] to address the nonconvex problem (4) is to express the nonconvex penalty as a Difference of two Convex functions, reformulate (4) as a DC programming and then apply DC Algorithms (DCA) to solve it. However, the DCA at each iteration also needs to solve a convex optimization problem, and thus the DCA may be slow in general. The work [31] has studied the application of the general method to problem (4) with some of nonconvex penalties. It is worth mentioning that in [30], DC approximation approach has recently been thoroughly investigated for sparse optimization. In particular, a new DC decomposition framework for penalty functions was proposed in [30], which plays a critical role in this paper.

When these above-mentioned methods are applied, how to choose a step size and to speed up them are two practical problems. However, for an unconstrained smooth optimization problem, many fast, effective classical methods with proper step size strategies have been well developed such as various types of nonlinear conjugate gradient methods and trust region methods. These smooth optimization methods are rarely used to solve (4) because of the non-smoothness of non-convex penalties. For the ℓ_1-penalized

least squares problem (2), that is, the Lasso problem, the work [32] has constructed a smooth function by using the Moreau envelope of the ℓ_1-norm, and has proved that the Lasso problem is equivalent to the problem of minimizing the new smooth function in the sense that the two problems have same sets of optimal solutions and optimal values. We therefore can apply classical smooth solvers to the equivalent smooth problem. A natural question is put forward: *Can nonconvex penalized least squares problems (4) with most of penalties be equivalent to corresponding smooth unconstrained optimization problems in the sense that sets of globally optimal solutions, and optimal values of the original problem and the corresponding smooth optimization problem are equal, respectively?* In this paper, by utilizing special decomposition properties of nonconvex penalties and the Moreau envelope technique, we give an affirmative answer to the above question. This is also main theoretical contribution of this paper.

The remainder of this paper is organized as follows. In Sect. 2, we study decomposition properties of most of used widely nonconvex penalties. We modify the SCAD, MCP and Capped-ℓ_1 penalties to form a new penalty call the TOP penalty in Sect. 3. In Sect. 4, a unified smooth optimization framework based on the Moreau envelope technique is further analyzed and then most of nonconvex penalized least squares problems is proved to be equivalent to corresponding smooth unconstrained optimization problems. Some extensions and concluding remarks are given in Sect. 5.

2 Decomposition properties of nonconvex penalties

This section presents important decomposition properties of most of nonconvex penalties which will discussed later. We first give these penalties in Table 1.

As pointed out in [30], some of penalties do not directly approximate ℓ_0-norm, however, if they are multiplied by an appropriate factor which can be incorporated into the parameter λ, and are added an appropriate term which does not affect original optimization problems, these modified penalties can become approximations of ℓ_0-norm. A key observation is that, except for the Capped-ℓ_1 and ℓ_p pseudo-norm, each of other popular penalties can be decomposed as the sum of a twice-continuously differentiable nonconvex function and a simple convex function whose proximity operator has a closed-form solution. To this end, We consider two types of decomposition for these penalties.

For convenience, we denote the penalty function by

$$P(x) = \sum_{i=1}^{n} p_{\lambda,\tau}(|x_i|), \quad x \in \mathbb{R}^n.$$

where $p_{\lambda,\tau}(|\cdot|)$ is the scalar penalty function with parameters $\lambda > 0, \tau > 0$, and $p_{\lambda,\tau}$ defined in $\mathbb{R}_+ = [0, +\infty)$ is a nonnegative function. To present the first decomposition form, we assume that the nonnegative function $p_{\lambda,\tau}$ satisfies the following conditions:

(i) $p_{\lambda,\tau}$ is concave and increasing in $\mathbb{R}_+$;

(ii) $p_{\lambda,\tau}(|\cdot|)$ is continuous in $\mathbb{R}$, and twice differentiable in $\mathbb{R}\backslash\{0\}$;

(iii) $p'_{\lambda,\tau}(0^+) > 0$, $p'_{\lambda,\tau}(z) > 0, \ \forall z > 0$;

(iv) $p''_{\lambda,\tau}(0^+) \leq 0$, $p''_{\lambda,\tau}(z) \leq 0, \ \forall z > 0$;

(v) $|p''_{\lambda,\tau}(0^+)| \leq L$, $|p''_{\lambda,\tau}(z)| \leq L, \ \forall z > 0$.

Inspired by the work in [30], we further study the new DC decomposition form of $p_{\lambda,\tau}(|\cdot|)$, that is,

$$p_{\lambda,\tau}(|z|) = h(z) + p'_{\lambda,\tau}(0^+)|z|, \quad h(z) = p_{\lambda,\tau}(|z|) - p'_{\lambda,\tau}(0^+)|z|. \tag{5}$$

For the function h defined above, we have the following important arguments.

Penalty functions	$p_{\lambda,\tau}(z)$
Exp	$\lambda(1-\exp(-\frac{z}{\tau}))$.
Smoothed-ℓ_p	$\lambda(z+\tau)^p$, $0<q<1$.
Log	$\lambda\log(\frac{z}{\tau}+1)$.
Arctan	$\frac{2\lambda}{\tau\sqrt{3}}\Big(\tan^{-1}\Big(\frac{1+2\tau z}{\sqrt{3}}\Big)-\frac{\pi}{6}\Big)$.
Fraction	$\frac{\lambda z}{z+\tau}$.
Transformed-ℓ_1	$\frac{\lambda(\tau+1)z}{z+\tau}$.
SCAD	$\begin{cases} \lambda z, & 0\le z<\lambda, \\ \frac{-z^2+2\lambda\tau z-\lambda^2}{2(\tau-1)}, & \lambda\le z<\tau\lambda, \\ \frac{\lambda^2(\tau+1)}{2}, & z\ge\tau\lambda. \end{cases}$
MCP	$\begin{cases} \lambda z-\frac{z^2}{2\tau}, & 0\le z<\lambda\tau, \\ \frac{\lambda^2\tau}{2}, & z\ge\lambda\tau. \end{cases}$
ℓ_p	λz^p, $0<p<1$.
Capped-ℓ_1	$\begin{cases} \lambda z, & 0\le z<\tau\lambda, \\ \lambda^2\tau, & z\ge\tau\lambda. \end{cases}$

Table 1: Nonconvex penalty functions defined in $\mathbb{R}_+$.

Theorem 2.1 The function h defined in (5) is concave, twice-continuously differentiable in $\mathbb{R}$, and its derivative is Lipschitz continuous with constant L.

Proof: By using simple derivation, we can express the derivative function of h as:

$$h'(z) = \begin{cases} p'_{\lambda,\tau}(z) - p'_{\lambda,\tau}(0^+), & z > 0, \\ 0, & z = 0, \\ -p'_{\lambda,\tau}(-z) + p'_{\lambda,\tau}(0^+), & z < 0. \end{cases}$$

Furthermore, we also can obtain

$$h''(x) = \begin{cases} p''_{\lambda,\tau}(z), & z > 0, \\ p''_{\lambda,\tau}(0^+), & z = 0, \\ p''_{\lambda,\tau}(-z), & z < 0. \end{cases}$$

Based on the assumption of the function $p_{\lambda,\tau}$, we have that $h'' \leq 0$, and thus the function h is concave, two differentiable in $\mathbb{R}$. We further state that h' is Lipschitz continuous. To this end, we consider three cases.

Case 1: $z_1 > z_2 \geq 0$, $h'(z_1) = p'_{\lambda,\tau}(z_1) - p'_{\lambda,\tau}(0^+)$, $h'(z_2) = p'_{\lambda,\tau}(z_2) - p'_{\lambda,\tau}(0^+)$. The property (v) of $p_{\lambda,\tau}$ shows that

$$|h'(z_2) - h'(z_2)| = |p'_{\lambda,\tau}(z_1) - p'_{\lambda,\tau}(z_2)| \leq L|z_1 - z_2|.$$

Case 2:, $z_1 < z_2 \leq 0$, $h'(z_1) = -p'_{\lambda,\tau}(-z_1) + p'_{\lambda,\tau}(0^+)$, $h'(z_2) = -p'_{\lambda,\tau}(-z_2) + p'_{\lambda,\tau}(0^+)$. The property (v) of $p_{\lambda,\tau}$ shows that

$$|h'(z_1) - h'(z_2)| = |p'_{\lambda,\tau}(-z_1) - p'_{\lambda,\tau}(-z_2)| \leq L|z_1 - z_2|.$$

Case 3: $z_1 > 0$, $z_2 < 0$, $h'(z_1) = p'_{\lambda,\tau}(z_1) - p'_{\lambda,\tau}(0^+)$, $h'(z_2) = -p'_{\lambda,\tau}(-z_2) + p'_{\lambda,\tau}(0^+)$. Applying the property (v) of $p_{\lambda,\tau}$ can obtain that

$$\begin{aligned} |h'(z_1) - h'(z_2)| &= |p'_{\lambda,\tau}(z_1) - p'_{\lambda,\tau}(0^+) + p'_{\lambda,\tau}(-z_2) - p'_{\lambda,\tau}(0^+)| \\ &\leq |p'_{\lambda,\tau}(z_1) - p'_{\lambda,\tau}(0^+)| + |p'_{\lambda,\tau}(-z_2) - p'_{\lambda,\tau}(0^+)| \\ &\leq Lz_1 - Lz_2 = L|z_1 - z_2|. \end{aligned}$$

Based on the above discussion, we state that $|h'(z_1) - h'(z_2)| \leq L|z_1 - z_2|$, $\forall z_1, z_2 \in \mathbb{R}$. $\sharp$

For $x \in \mathbb{R}^n$, we denote by $H(x) = \sum_{i=1}^n h(x_i)$. Theorem 2.1 can also indicate that the function H is concave, two differentiable in $\mathbb{R}^n$, and more importantly, it has a L-Lipschitz continuous gradient, that is,

$$||\nabla H(x) - \nabla H(y)||_2 \leq L||x - y||_2, \ \forall x, y \in \mathbb{R}^n.$$

Then, the penalty P can be decomposed as the sum of a twice-continuously differentiable concave function H and ℓ_1-norm (multiplied by factor $p'_{\lambda,\tau}(0^+)$), that is,

$$P(x) = H(x) + p'_{\lambda,\tau}(0^+)\|x\|_1. \tag{6}$$

We show below that Theorem 2.1 holds for the first six penalties in Table 1.

The exponential penalty. This function was first proposed to address support sector machine in [33]. Recently, it and its modified version have been introduced into sparse signal recovery [34] and sparse regression [15], respectively. We consider the initial scalar exponential penalty:

$$p_{\lambda,\tau}(|z|) = \lambda(1 - \exp^{-\frac{|z|}{\tau}}).$$

Obviously, the constant $p_{\lambda,\tau}(0^+) = \frac{\lambda}{\tau}$. The function h defined in 5 is

$$h(z) = \lambda(1 - \exp^{-\frac{|z|}{\tau}}) - \frac{\lambda}{\tau}|z|,$$

and its first and second derivatives are

$$h'(z) = \frac{\lambda}{\tau}(\exp^{-\frac{|z|}{\tau}} - 1)\text{sign}(z), \quad h''(z) = -\frac{\lambda}{\tau^2}\exp^{-\frac{|z|}{\tau}}.$$

The second derivative indicates that the function h' is $\frac{\lambda}{\tau^2}$-Lipschitz continuous.

The smoothed-ℓ_p penalty. This function is a modified version of ℓ_p, $1 < p < 1$, and is to avoid singularity at the origin by adding a small τ. Its form is as follows:

$$p_{\lambda,\tau}(|z|) = \lambda(|z| + \tau)^p.$$

The constant $p_{\lambda,\tau}(0^+)$ is $p\tau^{p-1}$ and then the form of h is

$$h(z) = \lambda(|z| + \tau)^p - \lambda p\tau^{p-1}|z|.$$

It is easy to obtain its first and second derivatives, that is,

$$h'(z) = \lambda p((|z| + \tau)^{p-1} - \tau^{p-1})\text{sign}(z), \quad h''(z) = \lambda p(p-1)(|z| + \tau)^{p-2}.$$

Clearly, h' is $\lambda p(1-p)\tau^{p-2}$-Lipschitz continuous.

The logarithmic penalty. There exist some different logarithmic penalties which can be found in [7, 23, 26]. We take the logarithmic function introduced in [26] as an example, that is,

$$p_{\lambda,\tau}(|z|) = \lambda\log(\frac{|z|}{\tau} + 1).$$

This function h is

$$h(z) = \lambda\log(\frac{|z|}{\tau} + 1) - \frac{\lambda}{\tau}|z|,$$

and then its first and second derivatives are

$$h'(z) = \frac{-\lambda z}{\tau(|z| + \tau)}, \quad h''(z) = \frac{-\lambda}{(|z| + \tau)^2}.$$

Thus, h' is $\frac{\lambda}{\tau^2}$-Lipschitz continuous.

The arctangent penalty. Compared with the logarithmic penalty, the arctangent penalty introduced in [23] can generate a so-called threshold function which increases more rapidly toward the identity function. The form of the arctangent penalty is expressed as:

$$p_{\lambda,\tau}(|z|) = \frac{2\lambda}{\tau\sqrt{3}}\Big(\tan^{-1}\Big(\frac{1 + 2\tau|z|}{\sqrt{3}}\Big) - \frac{\pi}{6}\Big).$$

It follows from [23] that

$$p_{\lambda,\tau}(0^+) = \lambda,\ h'(z) = \lambda\Big(\frac{1}{\tau^2 z^2 + \tau|z| + 1} - 1\Big)\mathrm{sign}(z),\ h''(z) = -\frac{\lambda\tau(2|\tau| + 1)}{(\tau^2 z^2 + \tau|z| + 1)^2}.$$

Further, we have that

$$h'''(z) = \frac{6\tau^3(\tau z^2 + z)}{(\tau^2 z^2 + \tau z + 1)^3} \geq 0, \quad \forall z \geq 0,$$

which implies that the function h'' is increasing on $\mathbb{R}_+$. In addition, h'' is an even function and $h''(z) \leq 0$, $\forall z \in \mathbb{R}$. Thus, $|h''(z)| \leq |h''(0)| = \lambda\tau$, and h' is $\lambda\tau$-Lipschitz continuous.

The fraction and transformed-ℓ_1 penalties. Studied recently in [25], the fraction function has the form of

$$p_{\lambda,\tau}(|z|) = \frac{\lambda|z|}{|z| + \tau}.$$

Thus, the function h can be expressed as

$$h(z) = \frac{\lambda|z|}{|z| + \tau} - \frac{\lambda}{\tau}|z|,$$

h' and h'' are

$$h'(z) = \lambda\Big(\frac{\tau}{(|z| + \tau)^2} - \frac{1}{\tau}\Big)\mathrm{sign}(z), \quad h''(z) = -\frac{2\lambda\tau}{(|z| + \tau)^3}.$$

Since $|h''(z)| \leq \frac{2\lambda}{\tau^2}$, $\forall z \in \mathbb{R}$, h' is $\frac{2\lambda}{\tau^2}$-Lipschitz continuous.

The transformed-ℓ_1 function is a variant of the fraction function, and its form is

$$p_{\lambda,\tau}(|z|) = \frac{\lambda(\tau + 1)|z|}{|z| + \tau}.$$

Thus, we have that

$$h(z) = \frac{\lambda(\tau + 1)|z|}{|z| + \tau} - \frac{\lambda(\tau + 1)}{\tau}|z|.$$

It is easy to show that the function h is also two differentiable in $\mathbb{R}$ and h' is $\frac{2\lambda(\tau+1)}{\tau^2}$-Lipschitz continuous.

Compared with the above six functions, the scalar SCAD and MCP functions are not two differentiable on $\mathbb{R}\backslash\{0\}$, and thus they do not possess the special decomposition form in (5) where the function h satisfies Theorem 2.1. However, the SCAD and MCP functions can be weakly convex (or semiconvex). A function ϕ defined in $\mathbb{R}^n$ is weakly convex with constant $\omega \geq 0$ if the function $\phi(x) + \frac{\omega}{2}\|x\|^2$ is convex. Thus, for the scalar SCAD and MCP functions, we can consider the second decomposition form, that is,

$$p_{\lambda,\tau}(|z|) = h(x) + \varphi(z),\ h(z) = -\frac{\omega}{2}z^2,\ \varphi(z) = p_{\lambda,\tau}(|z|) + \frac{\omega}{2}z^2,\ z \in \mathbb{R}. \tag{7}$$

Obviously, the quadratic function h is concave, twice-continuously differentiable in $\mathbb{R}$, and its derivative is ω-Lipschitz continuous.

Theorem 2.2 The scalar SCAD and MCP functions are weakly convex with constants $\frac{1}{\tau-1}$ and $\frac{1}{\tau}$, respectively.

Proof: (1) We set

$$s(x) = p_{\lambda,\tau}(|z|) + \frac{\omega}{2}z^2,$$

where the scalar SCAD function $p_{\lambda,\tau}(|\cdot|)$ is as follows: for $\tau > 2$ and $\lambda > 0$,

$$p_{\lambda,\tau}(|z|) = \begin{cases} \lambda|z|, & |z| < \lambda, \\ \frac{-z^2+2\lambda\tau|z|-\lambda^2}{2(\tau-1)}, & \lambda \le |z| < \tau\lambda, \\ \frac{\lambda^2(\tau+1)}{2}, & |z| \ge \tau\lambda. \end{cases}$$

For $z \in \mathbb{R}_+$, we can obtain

$$s'(z) = \begin{cases} \omega z + \lambda, & 0 \le z < \lambda, \\ (\omega - \frac{1}{\tau-1})z + \frac{\lambda\tau}{\tau-1}, & \lambda \le z < \tau\lambda, \\ \omega z, & z \ge \tau\lambda. \end{cases}$$

If $\omega \ge \frac{1}{\tau-1}$, we have

$$s'(z) \ge 0, \ \forall z \ge 0,$$

which indicates that s is convex in $\mathbb{R}_+$. Obviously, s is even and continuous at the origin. Thus, s is also convex in $\mathbb{R}_- = (-\infty, 0]$. For proving that s is convex in $\mathbb{R}$. we only need to indicate that for any $z_1 \in \mathbb{R}_+$, $z_2 \in \mathbb{R}_-$ and $t \in (0, 1)$, the point

$$(tz_1 + (1-t)z_2, ts(z_1) + (1-t)s(z_2) \in \text{epi}s = \{(z, y)|s(z) \le y\}.$$

In fact, this is equivalent to show that, for $tz_1 + (1-t)z_2 = 0$, the following inequality holds:

$$s(0) \le ts(z_1) + (1-t)s(z_2).$$

Since $s(0) = 0$, $s(z) \ge 0$, $\forall z \in \mathbb{R}$, the above inequality is obvious. The above proof indicates that the scalar SCAD function is weakly convex with constant $\frac{1}{\tau-1}$.

(2) The scalar MCP function is defined as

$$p_{\lambda,\tau}(|z|) = \begin{cases} \lambda|z| - \frac{z^2}{2\tau}, & |z| < \lambda\tau, \\ \frac{\lambda^2\tau}{2}, & |z| \ge \lambda\tau. \end{cases}$$

We also set

$$s(x) = p_{\lambda,\tau}(|z|) + \frac{\omega}{2}z^2.$$

When $z \in \mathbb{R}_+$, $s'(z)$ can be written as

$$s'(z) = \begin{cases} (\omega - \frac{1}{\tau})z + \lambda, & 0 \le z < \lambda\tau, \\ \omega z, & z \ge \lambda\tau, \end{cases}$$

which implies that s is convex in $\mathbb{R}_+$. Like the proof in (1), we can have that, if $\omega \ge \frac{1}{\tau}$, s is convex in $\mathbb{R}$. Thus, the scalar MCP function is weakly convex with constant $\frac{1}{\tau}$. ♯

Theorem 2.2 can also show further that the SCAD and MCP penalties can be decomposed as

$$P(x) = -\frac{\omega}{2}\|x\|^2 + \Phi(x), \ \Phi(x) = \sum_{i=1}^{n} \varphi(x_i), \ x \in \mathbb{R}^n, \tag{8}$$

where the function Φ is convex when ω is an appropriate value.

3 A new penalty function

Among all sparsity-promoting nonconvex penalties, the SCAD, MCP and Capped-ℓ_1 penalties possess the unbiasedness, that is, their derivatives vanish for large values and thus estimators related to these functions do not penalize large values. However, the second decomposition form of the SCAD and MCP penalties compared with the first decomposition form of other nonconvex penalties have a little disadvantage that the proximity opeartor and the Moreau envelope of the convex function Φ is more complex than the proximity opeartor and the Moreau envelope of the ℓ_1-norm. To construct the smooth optimization models of (4), we need to use the proximity opeartors and the Moreau envelopes of Φ and the ℓ_1-norm. Hence, in this section, we propose a new penalty function which not only enjoys the unbiasedness of the SCAD, MCP and Capped-ℓ_1 penalties, but also the first decomposition form.

To this end, we first express the scalar SCAD, MCP and Capped-ℓ_1 functions as the integral forms of their derivatives in $\mathbb{R}_+$, respectively, that is,

$$p_{\lambda,\tau}^{\rm SCAD}(|z|) = \lambda \int_0^{|z|} \min\{1, (\tau - t/\lambda)_+/(\tau+1)\}dt,$$
$$p_{\lambda,\tau}^{\rm MCP}(|z|) = \lambda \int_0^{|z|} \Big(1 - \frac{t}{\lambda\tau}\Big)\mathcal{I}(0 < t < \lambda\tau)dt,$$
$$p_{\lambda,\tau}^{\text{Capped-}\ell_1}(|z|) = \lambda \int_0^{|z|} \mathcal{I}(0 < t < \tau)dt.$$

Here, $(z)_+ = \max\{z, 0\}$, $\mathcal{I}$ is the characteristic function of a set. In fact, the MCP and Capped-ℓ_1 functions are linear and quadratic approximations of the SCAD function, respectively. Their derivatives in $\mathbb{R}_+$ are shown in Figure 1, which indicates that their derivatives are not smooth. This inspires us to consider the following function having a smooth derivative in $\mathbb{R}_+$:

$$p_{\lambda,\tau}(|z|) = \lambda \int_0^{|z|} c(t-\gamma)^2\mathcal{I}(0 < t < \tau)dt, \tag{9}$$

To guarantee the unbiasedness of the penalty function (9), we also need to have $p_{\lambda,\tau}(\tau) = \lambda$, which indicates that the constant c is

$$c = \frac{1}{\int_0^\tau (t-\tau)^2 dt} = \frac{3}{\tau^3}.$$

Thus, the new penalty function can be fully expressed as

$$p_{\lambda,\tau}(|z|) = \begin{cases} \frac{\lambda}{\tau^3}(|z|-\tau)^3 + \lambda, & |z| < \tau, \\ \lambda, & |z| \geq \tau. \end{cases} \tag{10}$$

Since the nonconstant part of the function in $\mathbb{R}_+$ is a Three-Order Polynomial function, we call it the scalar TOP function. The function and its derivative are shown in Figure 2. Obviously, the new penalty function (10) is singular at the origin to achieve sparsity and its derivative vanishes for large values. Furthermore, we have

$$p'_{\lambda,\tau}(z) = \frac{3\lambda}{\tau^3}(z-\tau)^2\mathcal{I}(0 < z < \tau), \quad p''_{\lambda,\tau}(z) = \frac{6\lambda}{\tau^3}(z-\tau)\mathcal{I}(0 < z < \tau),\ z \in \mathbb{R}_+.$$

This means that $p_{\lambda,\tau}$ in $\mathbb{R}_+$ is concave and increasing, $p'_{\lambda,\tau}(0^+) = \frac{3\lambda}{\tau}$, $|p''_{\lambda,\tau}(z)| \leq \frac{6\lambda}{\tau^2}$. Thus, The scalar TOP function has the decomposition form (5). We now define the TOP penalty in $\mathbb{R}^n$ as $P(x) = \sum_{i=1}^n p_{\lambda,\tau}(|x_i|)$ where $p_{\lambda,\tau}(|\cdot|)$ is the scalar TOP function. Clearly, the TOP penalty enjoys the first decomposition form (6).

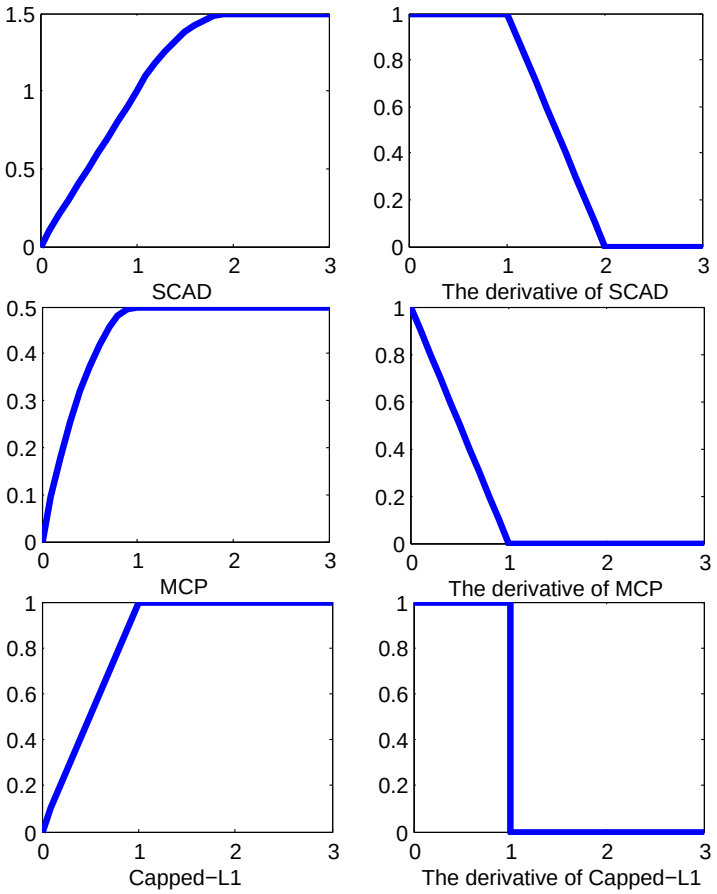

Figure 1: The curves of the SCAD ($\lambda = 1, \tau = 2.2$), MCP ($\lambda = 1, \tau = 1$), Capped-ℓ_1 ($\lambda = 1, \tau = 1$) and these derivatives in $\mathbb{R}_+$.

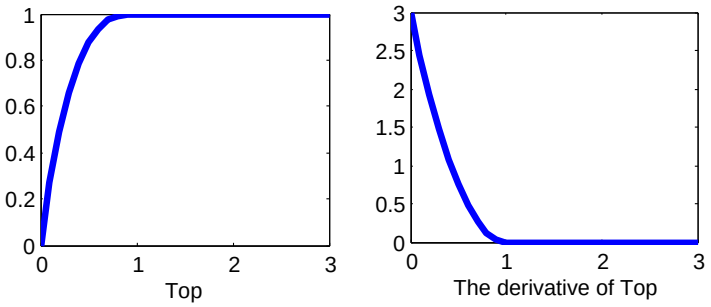

Figure 2: The curves of Top ($\lambda = 1, \tau = 1$) and its derivative in $\mathbb{R}_+$.

4 A unified smooth optimization framework

This section presents our main result that each of nonconvex and nonsmooth problems (4) with first eight penalties in Table 1 and the TOP penalty is equivalent to the problem of minimizing an unconstrained continuously differentiable function.

To this end, we first recall some notations and notions from convex analysis theory. The set of all proper, lower semicontinuous convex extended-value functions defined in $\mathbb{R}^n$ is denotes by $\Gamma_0(\mathbb{R}^n)$. For a function $\psi \in \Gamma_0(\mathbb{R}^n)$, its subdifferential, denoted by $\partial\psi$, is a set-valued mapping from $\mathbb{R}^n$ to $2^{\mathbb{R}^n}$, defined at a given point $x \in \mathbb{R}^n$ by

$$\partial\psi(x) = \{u \in \mathbb{R}^n : \psi(v) \geq \psi(x) + \langle u, v - x\rangle, \ \forall v \in \mathbb{R}^n\}.$$

In particular, if the convex ψ is differentiable at x, we have $\partial\psi(x) = \{\nabla\psi(x)\}$. Moreover, the subdifferential operator $\partial\psi$ is maximally monotone.

The Moreau envelope and proximity operator of a function $\psi \in \Gamma_0(\mathbb{R}^n)$ is two key tools used in the latter analysis. The Moreau envelope of a function $\psi \in \Gamma_0(\mathbb{R}^n)$ with parameter $\beta > 0$ is the function $\psi^\beta : \mathbb{R}^n \to \mathbb{R}$, defined as

$$\psi^\beta(x) = \inf_{u\in\mathbb{R}^n} \left\{ \frac{1}{2\beta}||u - x||^2 + \psi(u) \right\}. \tag{11}$$

The minimizer of the optimization problem (11) yields the proximity operator denoted by $\text{prox}_{\beta\psi}$, that is,

$$\text{prox}_{\beta\psi}(x) = \operatorname*{argmin}_{u\in\mathbb{R}^n} \left\{ \frac{1}{2\beta}||u - x||^2 + \psi(u) \right\}. \tag{12}$$

It is well-known that the proximity operator is single-valued and firmly nonexpansive, and has following property, that is, for $\beta > 0$,

$$x = \text{prox}_{\beta\psi}(x + \beta u) \quad \Leftrightarrow \quad u \in \partial\psi(x). \tag{13}$$

For the Moreau envelope, it is convex, real-valued, continuously differentiable in $\mathbb{R}^n$ and has the β^{-1}-Lipschitz continuous gradient expressed as

$$\nabla\psi^\beta(x) = \beta^{-1}(x - \text{prox}_{\beta\psi}(x)). \tag{14}$$

It follows from [35] that the Moreau envelope ψ^β converges to ψ as $\beta \to 0^+$, and more importantly, the minimum values and sets of minimizers of ψ and ψ^β are equal, respectively. This means that by using the Moreau envelope technique, one can equivalently transform a nonsmooth convex optimization problem into a smooth convex optimization problem.

We now consider general nonsmooth optimization problems of the form

$$\min_{x\in\mathbb{R}^n} F(x) = f(x) + g(x), \tag{15}$$

where f is a nonconvex smooth function and $g \in \Gamma_0(\mathbb{R}^n)$ is possibly nonsmooth. The set of stationary points of problem (15) is denoted by

$$S_F = \{x \in \mathbb{R}^n : 0 \in \nabla f(x) + \partial g(x)\}.$$

It follows from [36] that a point x being a local minimizer of (15) satisfies $x \in S_F$. The property (13) indicates that x is a stationary point of (15) if and only if

$$x = \text{prox}_{\beta g}(x - \beta\nabla f(x)). \tag{16}$$

For simplicity, we introduce two operators T_β, R_β defined as, respectively,

$$T_\beta(x) = \text{prox}_{\beta g}(x - \beta\nabla f(x)), \tag{17}$$

$$R_\beta(x) = \beta^{-1}(x - T_\beta(x)). \tag{18}$$

The operator T_β is also called the proximal gradient operator or the forward-backward mapping in [32]. Like the definition of the proximity operator, T_β can be further expressed as

$$T_\beta(x) = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2\beta}||u - x||^2 + f(x) + \langle \nabla f(x), u - x\rangle + g(u) \right\}. \tag{19}$$

We aim to transform problem (15) into the minimization of an unconstrained continuously differentiable function. Inspired by the definition of the Moreau envelope, the work in [32] considers the value function of the optimization problem (19) and defines the so-called forward-backward envelope.

Definition 4.1 (Forward-backward envelope) Let F be the objective function in problem (15) and $\beta > 0$. The forward-backward envelope of F with parameter β is

$$F_\beta(x) = \min_{u\in\mathbb{R}^n} \left\{ \frac{1}{2\beta}||u - x||^2 + f(x) + \langle \nabla f(x), u - x\rangle + g(u) \right\}. \tag{20}$$

By using (19) and the definition of the Moreau envelope, we can express explicitly F_β as

$$F_\beta(x) = f(x) - \frac{\beta}{2}\|\nabla f(x)\|^2 + g^\beta(x - \beta\nabla f(x)). \tag{21}$$

Similar to the Moreau envelope, the function F_β is real-valued. If the function f in problem (15) has a L_f-Lipschitz continuous gradient, the function F_β enjoys many other favorable properties proved by [32]. We also present these interesting properties. Note that these properties can be derived simply from the convexity of g and the well-known inequality about f, that is,

$$f(u) \le f(x) + \langle \nabla f(x), u - x\rangle + \frac{L_f}{2}\|u - x\|^2,\ \forall u,\ x \in \mathbb{R}^n. \tag{22}$$

Proposition 4.1 Let f in problem (15) have a L_f-Lipschitz continuous gradient, and $g \in \Gamma_0(\mathbb{R}^n)$. The following inequalities hold for all $x \in \mathbb{R}^n$.

(i) $F_\beta(x) \le F(x) - \frac{\beta}{2}\|R_\beta(x)\|^2,\ \forall\beta > 0$;

(ii) $F(T_\beta(x)) \le F_\beta(x) - \frac{\beta}{2}(1 - \beta L_f)\|R_\beta(x)\|^2,\ \forall\beta > 0$;

(iii) $F(T_\beta(x)) \le F_\beta(x),\ \forall\beta \in (0, 1/L_f)$.

Furthermore,

(iv) $F_\beta(x) = F(x),\ \forall\beta > 0,\ \forall x \in S_F$;

(v) $\inf F_\beta = \inf F$, and $\operatorname{argmin} F \subseteq \operatorname{argmin} F_\beta,\ \forall\beta \in (0, 1/L_f]$;

(vi) $\operatorname{argmin} F = \operatorname{argmin} F_\beta,\ \forall\beta \in (0, 1/L_f)$.

For the sake of completeness, the simple and direct proofs of these properties in Proposition 4.1 are also provided in the Appendix. Proposition 4.1 shows that, if $\beta \in (0, 1/L_f)$, the problems of minimizing F and F_β are equivalent. We now discuss the differentiability of F_β. To this end, we assume that the function f has a L_f-Lipschitz continuous gradient and is twice-continuously differentiable in $\mathbb{R}^n$. It follows from [32] that, for any given point x and $\beta \in (0, 1/L_f)$, the matrix

$$Q_\beta(x) = I - \beta\nabla^2 f(x) \tag{23}$$

is symmetric and positive definite. Thus, we have following important result from [32].

Theorem 4.1 Let f have a L_f-Lipschitz continuous gradient and be twice-continuously differentiable in $\mathbb{R}^n$. Then, F_β is continuously differentiable, and its gradient is

$$\nabla F_\beta(x) = Q_\beta(x)R_\beta(x). \tag{24}$$

If $\beta \in (0, 1/L_f)$, the sets of stationary points of F and F_β are equal.

Remark 4.1 To compute the gradient ∇F_β, we need to use the Hessian matrix of f, however, in nonconvex penalized least squares problems and some other sparse optimization, we only perform matrix-vector products with $\nabla^2 f$, and do not use the computation of the full Hassian. More remarks on the Hessian matrix of f, we refer the reader to [32].

Proposition 4.1 and Theorem 4.1 indicate that, if $\beta \in (0, 1/L_f)$, problem (15) is equivalent to the minimization of the unconstrained continuously differentiable function F_β. In the rest of this section, we focus on each of problems (4) with the first eight penalties in Table 1 and the TOP penalty, and construct corresponding equivalent unconstrained smooth optimization problems. We consider the first decomposition form (6) of penalties and then write problem (4) as

$$\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{2}\|Ax-b\|^2 + H(x) + p'_{\lambda,\tau}(0^+)\|x\|_1. \tag{25}$$

We denote by

$$f(x) = \frac{1}{2}\|Ax-b\|^2 + H(x),\ g(x) = p'_{\lambda,\tau}(0^+)\|x\|_1. \tag{26}$$

Obviously, problem (26) is a special case of the general optimization problem (15). Since the function H has a L_H-Lipschitz continuous gradient and is twice-continuously differentiable, we can have

$$L_f = \|A^TA\| + L_H,\ \nabla f(x) = A^T(Ax-b) + (h'(x_1), h'(x_2), \cdots, h'(x_n))^T, \tag{27}$$

and

$$\nabla^2 f(x) = A^TA + \mathrm{diag}(h''(x_1), h''(x_2), \cdots, h''(x_n)), \tag{28}$$

where $\mathrm{diag}(\cdot)$ denotes a diagonal matrix. For the convex function $g = p'_{\lambda,\tau}(0^+)\|x\|_1$, its proximity operator with parameter β is the well-known soft-thresholding operator with parameter $\beta p'_{\lambda,\tau}(0^+)$, that is,

$$\mathrm{prox}_{\beta g}(x) = \mathrm{sign}(x) \odot \max\{|x| - \beta p'_{\lambda,\tau}(0^+), 0\}, \tag{29}$$

where $\odot$ denotes elementwise multiplication. Moreover, its Moreau envelope with parameter β is sum of the well-known Huber function on each of the components, that is,

$$g^\beta(x) = \sum_{i=1}^n H_\beta(x_i),\ H_\beta(z) = \begin{cases} |z| - \frac{\beta p'_{\lambda,\tau}(0^+)}{2}, & |z| > \beta p'_{\lambda,\tau}(0^+), \\ \frac{z^2}{2\beta p'_{\lambda,\tau}(0^+)}, & |z| \le \beta p'_{\lambda,\tau}(0^+). \end{cases} \tag{30}$$

Based on the equality (23), Proposition 4.1 and Theorem 4.1, we can obtain the equivalent smooth optimization problem of (4).

Theorem 4.2 Let the scalar penalty function $p_{\lambda,\tau}$ in problem (4) be one of the first six functions in Table 1 or the TOP function. Then, the smooth optimization problem equivalent to problem (4) is

$$\min_{x\in\mathbb{R}^n} F_\beta(x) = \frac{1}{2}\|Ax-b\|^2 + H(x) - \frac{\beta}{2}\|A^T(Ax-b) + \nabla H(x)\|^2 + g^\beta(x - \beta(A^T(Ax-b) + \nabla H(x))), \tag{31}$$

where $\beta \in \left(0, \frac{1}{\|A^TA\|+L_H}\right)$, the smooth function g^β is defined in (30) and the gradient operator ∇H is given as

$$\nabla H(x) = (h'(x_1), h'(x_2), \cdots, h'(x_n)).$$

In particular, for the exponential penalty and the smoothed-ℓ_p penalty, we can calculate the proximity operator of ℓ_1-norm, instead of the hard proximity operators of these two penalties, to construct the equivalent smooth optimization problems of (4) with these two penalties.

When the penalty in problem (4) is SCAD or MCP, by utilizing the decomposition form (8), we can express problem (4) as

$$\min_{x\in\mathbb{R}^n} F(x) = f(x) + \Phi(x), \tag{32}$$

where the convex function Φ is defined in (8), and we denote by

$$f(x) = \frac{1}{2}\|Ax-b\|^2 - \frac{\omega}{2}\|x\|^2. \tag{33}$$

It is easy to see that

$$L_f = \|A^TA - \omega I\|,\ \nabla f(x) = A^T(Ax-b) - \omega x,\ \nabla^2 f(x) = A^TA - \omega I. \tag{34}$$

To construct the smooth function of the objective function of problem (32), we also need to calculate the proximity operator of Φ. To this end, we present a simple statement about the proximity operator of a weakly convex function.

Lemma 4.1 Let ϕ defined in $\mathbb{R}^n$ be a weakly convex function with constant $\omega \geq 0$. The convex function $\bar{\phi}$ is defined as

$$\bar{\phi} = \phi(x) + \frac{\omega}{2}\|x\|^2.$$

Then, the proximity operators of the two functions ϕ and $\bar{\phi}$ satisfy(provided $\beta\omega < 1$)

$$\mathrm{prox}_{\beta\phi}(x) = \mathrm{prox}_{\frac{\beta}{1-\beta\omega}\bar{\phi}}\Big(\frac{1}{1-\beta\omega}x\Big),\ \mathrm{prox}_{\beta\bar{\phi}}(x) = \mathrm{prox}_{\frac{\beta}{1+\beta\omega}\phi}\Big(\frac{1}{1+\beta\omega}x\Big). \tag{35}$$

Lemma 4.1 can be shown by making use of the definition of the proximity operator. It can be seen that, if the nonconvex function is weakly convex, its proximity operator with parameter $\beta < \omega^{-1}$ is also single-valued. Since the function Φ is convex and sparable, we only calculate the proximity operator of the scalar convex function φ defined in (7). It follows from Theorem 2.2 and Lemma 4.1 that the proximity operator of the scalar convex function φ can be derived from the proximity operator of the scalar SCAD or MCP function.

For the scalar SCAD function, its proximity operator with parameter $\beta < \tau - 1$ can be given as

$$\mathrm{prox}_{\beta p_{\lambda,\tau}(|\cdot|)}(z) = \begin{cases} 0, & |z| \leq \lambda\beta, \\ \mathrm{sign}(z)(|z| - \lambda\beta), & \lambda\beta < |z| \leq \lambda(\beta+1), \\ \mathrm{sign}(z)\frac{(\tau-1)|z| - \lambda\beta\tau)}{\tau-\beta-1}, & \lambda(\beta+1) < |z| < \lambda\tau, \\ z, & |z| \geq \lambda\tau. \end{cases}$$

The proximity operator of the scalar MCP is

$$\mathrm{prox}_{\beta p_{\lambda,\tau}(|\cdot|)}(z) = \begin{cases} 0, & |z| \leq \lambda\beta, \\ \mathrm{sign}(z)\frac{\tau(|z|-\lambda\beta)}{\tau-\beta}, & \lambda\beta < |z| < \lambda\tau, \\ z, & |z| \geq \lambda\tau. \end{cases}$$

Note that the conditions $\beta < \tau - 1$ and $\beta < \tau$ ensure the single-valuedness of proximity operators of the scalar SCAD and MCP function, respectively. We now present the following important result.

Theorem 4.3 Let the penalty function $p_{\lambda,\tau}$ in problem (4) be the weakly convex scalar SCAD or MCP function with constant $\omega \geq 0$. The function Φ is defined in (8). Then, the smooth optimization problem equivalent to problem (4) is

$$\min_{x\in\mathbb{R}^n} F_\beta(x) = \frac{1}{2}\|Ax-b\|^2 - \frac{\omega}{2}\|x\|^2 - \frac{\beta}{2}\|A^T(Ax-b)-\omega x\|^2 + \Phi^\beta(x-\beta(A^T(Ax-b)-\omega x)), \quad (36)$$

where $\beta \in \big(0, \min\big\{\frac{1}{\|A^TA-\omega I\|}, \frac{1}{\omega}\big\}\big)$, the convex function Φ^β is the Moreau envelope of Φ.

Theorem 4.2 and Theorem 4.3 indicate that most of nonconvex penalized least squares problems can be equivalent to smooth optimization problems, which opens up the possibility of extending many existing smooth methods to solve these nonconvex penalized least squares problems.

5 Extensions and conclusion

Our approach above for nonconvex penalized least squares problems can be extended to treat other types of sparse optimization problems. A direct example of interest is the nonconvex sparse logistic regression model [15] expressed as

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^m \log\Big(1+\exp(-b_i\langle a_i, x\rangle)\Big) + \sum_{i=1}^n p_{\lambda,\tau}(|x_i|), \quad (37)$$

where $x \in \mathbb{R}^n$, $a_i \in \mathbb{R}^n$, and $b_i \in \{-1, 1\}$. In machine learning applications, one can utilize the model to find a linear classifier for points a_i. Let us define a new matrix $A \in \mathbb{R}^{m\times n}$ taking $-b_ia_i$ as i-th row, and define a new function $\bar{f}(y) = \sum_{i=1}^m \log(1+\exp(y_i))$. We set $f(x) = \bar{f}(Ax)$ which is the first term of the objective function of model (37). Since $\bar{f}$ is separable, it is easy to see that the function f has a $\frac{1}{4}\|A^TA\|$-Lipschitz continuous gradient and is twice-continuously differentiable in $\mathbb{R}^n$. When the nonconvex function $p_{\lambda,\tau}$ is one of the first eight penalties in Table 1, model (37) can be transformed into an equivalent smooth optimization.

Another important problem is the nonconvex penalized matrix least squares model [37] which aims to recover a low-rank matrix from a small number of noisy linear measurements. This model has the form

$$\min_{X\in\mathbb{R}^{m\times n}} \frac{1}{2}\|\mathcal{A}(X)-b\|^2 + \sum_{i=1}^n p_{\lambda,\tau}(\sigma_i(X)), \quad (38)$$

where $\mathcal{A}: \mathbb{R}^{m\times n} \to \mathbb{R}^s$ with $s \ll mn$ and the assumption $n \leq m$ is a linear operator, $\sigma_i(X)$ is the i-th component of the vector $\sigma(X) = (\sigma_1(X), \cdots, \sigma_n(X))$ of all singular values arranged in descending order. A well-known application of model (38) is the low-rank matrix completion [38]. For the nonconvex function $p_{\lambda,\tau}$ is one of the first six penalties in Table 1, we can consider its first decomposition form (5) and then obtain

$$P(\sigma(X)) = H(\sigma(X)) + p_{\lambda,\tau}(0^+)\|X\|_*. \quad (39)$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix. Obviously, the function H is absolutely symmetric [39]. It follows from [35] and [39] that the matrix function $H\circ\sigma$ is concave, and its gradient can expressed as

$$\nabla(H\circ\sigma)(X) = U_X \mathrm{diag}(\nabla H(\sigma(X)))V_X^T,$$

with U_X, V_X satisfying the singular value decomposition (SVD) $X = U_X\mathrm{diag}(\sigma(X))V_X^T$. However, the decomposition (39) may be not practical in theory and algorithm. Indeed, we do not know whether the matrix function $H\circ\sigma$ is twice-continuously differentiable in $\mathbb{R}^{m\times n}$ and what form $\nabla^2(H\circ\sigma)$ is, although the function H itself is twice-continuously differentiable in $\mathbb{R}^n$. This prevents us from constructing

the corresponding equivalent smooth optimization model. In addition, when we apply the well-known proximal gradient algorithm to solve the new model

$$\min_{X\in\mathbb{R}^{m\times n}} \underbrace{\frac{1}{2}\|\mathcal{A}(X)-b\|^2 + H(\sigma(X))}_{f(X)} + \underbrace{p_{\lambda,\tau}(0^+)\|X\|_*}_{g(X)}, \tag{40}$$

although we can use the SVD to obtain the closed-form solution of the proximity operator of g, the algorithm may be not very efficient because of the most computationally expensive task of computing the SVD twice at each iteration. However, when the nonconvex function $p_{\lambda,\tau}$ is chose to be the scalar SCAD or MCP penalty, model (38) can be expressed as

$$\min_{X\in\mathbb{R}^{m\times n}} \underbrace{\frac{1}{2}\|\mathcal{A}(X)-b\|^2 - \frac{\omega}{2}\|X\|_F^2}_{f(X)} + \underbrace{\Phi(\sigma(X))}_{g(X)}. \tag{41}$$

Since the function Φ is convex and absolutely symmetric, the composite function $\Phi\circ\sigma$ is also convex. It follows from our approach that model (41) can be equivalent to corresponding smooth optimization problem.

In this paper, we further study properties of most of popular nonconvex penalties, give a new penalty function, and make use of the well-known the Moreau envelope technique to prove that many nonconvex penalized least squares problems and other sparse optimization models such as nonconvex sparse logistic regression models and nonconvex penalized matrix least squares models can be transform equivalently into corresponding smooth unconstrained optimization problems in the sense that sets of globally optimal solutions, and optimal values of the original problem and the corresponding smooth optimization problem are equal, respectively. There exist some future research tasks. For example,

(1) the new TOP penalty function is a modified version of the SCAD, MCP and Capped-ℓ_1 penalties. It is necessary to study further its statistical properties and give more detailed comparison with other penalties;

(2) to prove that the matrix model (40) is equivalent to a smooth optimization problem, we need to investigate whether $\nabla^2(H\circ\sigma)$ exists and what form it may be;

(3) it is also very practical to study the numerical performance of using the classical smooth methods for solving these equivalent nonconvex smooth problems to address various sparse tasks.

Acknowledgements

This research was supported by the National Natural Science Foundation of China under the Grant nos. 61901404 and 11771347, and supported by the Nanhu Scholars Program for Young Scholars of XYNU.

Appendix A

Proof of Proposition 4.1. We first prove the property (i). The forward-backward envelope F_β defined in (24) can be expressed as

$$F_\beta(x) = f(x) + g(T_\beta(x)) - \beta\langle\nabla f(x), R_\beta(x)\rangle + \frac{\beta}{2}\|R_\beta(x)\|^2.$$

It follows from the convex optimization problem (19) that

$$R_\beta(x) - \nabla f(x) \in \partial g(T_\beta(x)),$$

which indicates that

$$\begin{aligned} f(x)+g(x) &\geq f(x)+g(T_\beta(x))+\langle R_\beta(x)-\nabla f(x), x-T_\beta(x)\rangle \\ &= F_\beta(x)+\frac{\beta}{2}\|R_\beta(x)\|^2. \end{aligned}$$

Thus the property (i) holds. It can be seem from the inequality (22) that

$$F_\beta(x) \geq f(T_\beta(x))+g(T_\beta(x))-\frac{L_f}{2}\|T_\beta(x)-x\|^2+\frac{\beta}{2}\|R_\beta(x)\|^2,$$

which proves the the properties (ii) and (iii). Based on the equality (16) and properties (i) and (ii), we have the property (iv).

To prove (v), we consider $\beta \in (0, 1/L_f]$ and $\bar{x} \in \operatorname{argmin} F$. Based on properties (i) and (iii), we have

$$F_\beta(\bar{x}) = F(\bar{x}) \leq F(T_\beta(x)) \leq F_\beta(x),\ \forall x \in \mathbb{R}^n.$$

Thus, $\bar{x}$ is also a minimizer of F_β, and $\min F_\beta = \min F$ if $\operatorname{argmin} F \neq \emptyset$. Suppose that $\operatorname{argmin} F = \emptyset$. It follows from (i) that $\inf F_\beta \leq \inf F$. If there exists $x \in \mathbb{R}^n$ such as $F_\beta(x) \leq \inf F$, the property (ii) indicates that

$$F(T_\beta(x)) \leq \inf F,$$

which is a contradiction with $\operatorname{argmin} F = \emptyset$. Thus, $\inf F_\beta = \inf F$. The proof of (v) is completed.

If $\beta \in (0, 1/L_f)$ and $\bar{x} \in \operatorname{argmin} F_\beta$, we can have from (i) and (ii) that

$$F_\beta(T_\beta(\bar{x})) \leq F(T_\beta(\bar{x})) \leq F_\beta(\bar{x})-\frac{1-\beta L_f}{2\beta}\|\bar{x}-T_\beta(\bar{x})\|^2,$$

This indicates that $\bar{x} = T_\beta(\bar{x})$. Thus, we further have that

$$F_\beta(\bar{x}) = F_\beta(T_\beta(\bar{x})) \leq F(\bar{x}) \leq F_\beta(\bar{x}),$$

that is, $F_\beta(\bar{x}) = F(\bar{x})$. Since $F_\beta \leq F$ and $\bar{x} \in \operatorname{argmin} F_\beta$, we have $\bar{x} \in \operatorname{argmin} F$. Together with (v), we can prove (iv). ♯

Proof of Theorem 4.1. The gradient of F_β can be obtained from the expression (23) for F_β and the gradient of the Moreau envelope g^β. Indeed, we have

$$\begin{aligned} \nabla F_\beta(x) &= \nabla f(x)-\beta\nabla^2 f(x)\nabla f(x)+\beta^{-1}(I-\beta\nabla^2 f(x))(x-\beta\nabla f(x)-T_\beta(X)) \\ &= Q_\beta(x)R_\beta(x). \end{aligned}$$

If $\beta \in (0, 1/L_f)$, then $Q_\beta(x)$ is nonsingular for all $x \in \mathbb{R}^n$. Thus, $\nabla F_\beta(x) = 0$ if and only if $R_\beta(x) = 0$, that is, the sets of stationary points of F and F_β are equal. ♯

References

[1] P. Bühlmann, S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

[2] T. Hastie, R. Tibshirani, M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* CRC press, 2015.

[3] E. Candès, J. Romberg, T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 2006, 59(8): 1207-1223.

[4] S. Foucart, H. Rauhut. *A mathematical introduction to compressive sensing*. Basel: Birkhäuser, 2013.

[5] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer New York, 2010.

[6] D. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 2006, 52(4): 1289-1306.

[7] R. Mazumder, J. Friedman, T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 2011, 106(495): 1125-1138.

[8] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, series B*, 1994, 58(1): 267-288.

[9] S. Chen, D. Donoho, M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 1998, 20(1): 33-61.

[10] B. Efron, T. Hastie, I. Johnstone, et al. Least angle regression. *Annals of Statistics*, 2004, 32(2): 407-499.

[11] M. Figueiredo, R. Nowak, S. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, 2007, 1(4): 586-598.

[12] A. Beck, M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2009, 2(1): 183-202.

[13] P. Combettes. V. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Journal on Multiscale Modeling & Simulation*, 2005, 4(4): 1168-1200.

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

[15] C. Gao, N. Wang, Q. Yu, et al. A Feasible Nonconvex Relaxation Approach to Feature Selection. *AAAI*, 2011: 356-361.

[16] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 2010, 38(2): 894-942.

[17] Zhang T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 2010, 11(Mar): 1081-1107.

[18] R. Chartrand, V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 2008, 24(3): 035020.

[19] S. Foucart, M. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 2009, 26(3): 395-407.

[20] Q. Sun. Recovery of sparsest signals via ℓ_q-minimization. *Applied and Computational Harmonic Analysis*, 2012, 32(3): 329-341.

[21] J. Peng, S. Yue, H. Li. NP/CMP Equivalence: A Phenomenon Hidden Among Sparsity Models ℓ_0 Minimization and ℓ_p Minimization for Information Processing. *IEEE Transactions on Information Theory*, 2015, 61(7): 4028-4033.

[22] Y. Zhao, D. Li. Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems. *SIAM Journal on Optimization*, 2012, 22(3): 1065-1088.

[23] I. Selesnick, I. Bayram. Sparse signal estimation by maximally sparse convex optimization. *IEEE Transactions on Signal Processing*, 2014, 62(5): 1078-1092.

[24] S. Zhang, J. Xin. Minimization of Transformed L_1 Penalty: Closed Form Representation and Iterative Thresholding Algorithms. *arXiv preprint arXiv:1412.5240*, 2014.

[25] H. Li, Q. Zhang, A. Cui, et al. Minimization of fraction function penalty in compressed sensing. *arXiv preprint arXiv:1705.06048*, 2017.

[26] E. Candès, M. Wakin, S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 2008, 14(5-6): 877-905.

[27] H. Zou, R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 2008, 36(4): 1509-1533.

[28] Z. Xu, X. Chang, F. Xu, et al. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on neural networks and learning systems*, 2012, 23(7): 1013-1027.

[29] C. Lu, Y. Wei, Z. Lin, et al. Proximal Iteratively Reweighted Algorithm with Multiple Splitting for Nonconvex Sparsity Optimization. *AAAI*, 2014: 1251-1257.

[30] H. Thi, T. Dinh, H. Le, X. Vo. DC approximation approaches for sparse optimization. *European Journal of Operational Research*, 2015, 244(1): 26-46.

[31] G. Gasso, A. Rakotomamonjy, S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing*, 2009, 57(12): 4686-4698.

[32] L. Stella, A. Themelis, P. Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 2017, 67(3): 443-487.

[33] P. Bradley, O. Mangasarian. Feature selection via concave minimization and support vector machines. *ICML*, 1998, 98: 82-90.

[34] M. Malek-Mohammadi, A. Koochakzadeh, M. Babaie-Zadeh, et al. Successive concave sparsity approximation for compressed sensing. *IEEE Transactions on Signal Processing*, 2016, 64(21): 5657-5671.

[35] H. Bauschke, P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, second edition, 2017.

[36] R. Rockafellar, R. Wets. *Variational analysis*. Springer Science & Business Media, 2009.

[37] C. Lu, J. Tang, S. Yan, et al. Generalized nonconvex nonsmooth low-rank minimization. *CVPR*, 2014: 4130-4137.

[38] E. Candès, B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 2009, 9(6):717-772.

[39] A. Lewis, H. Sendov. Nonsmooth analysis of singular values. Part I: Theory. *Set-Valued Analysis*, 2005, 13(3): 213-241.