The Moreau envelope based efficient first-order methods for sparse recovery

Yongchao Yu, Jigen Peng

PII: S0377-0427(17)30133-4
DOI: http://dx.doi.org/10.1016/j.cam.2017.03.014
Reference: CAM 11061

To appear in: *Journal of Computational and Applied Mathematics*

Received date: 10 July 2015
Revised date: 9 December 2016

The Moreau Envelope Based Efficient First-order Methods for Sparse Recovery

Yongchao Yu[a], Jigen Peng[a,*]

[a]*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China; Beijing Center for Mathematics and Information Interdisciplinary Sciences (BCMIIS), Beijing, 100048, China.*

Abstract

Sparse recovery from indirectly under-sampled or possibly noisy data is a burgeoning topic drawing the attention of many researchers. Since sparse recovery problems can be cast as a class of the constrained convex optimization model which minimizes a nonsmooth convex objection function in a convex closed set, fast and efficient methods for solving the constrained optimization model are highly needed. By introducing the indicator functions related to constrained sets in sparse recovery models, we reformulate these models as two general unconstrained optimization problems. To develop fast first-order methods, two smoothing approaches are proposed based on the Moreau envelope: smoothing the related indicator functions or the objection functions. By using the first smoothing approach, we obtain more proper unconstrained models for sparse recovery from noisy data. Fast iterative shrinkage-thresholding algorithm (FISTA) is applied to solve the smoothed models. When smoothing the objection functions, we propose an efficient first-order method based on FISTA to establish the rate of convergence of order $\mathcal{O}(\frac{\log k}{k})$ for the iterative sequence of values of the original objective functions. Numerical experiments have demonstrated that the two proposed smoothing methods are comparable to the state-of-the-art first-order methods with respect to accuracy and speed when applied to the sparse recovery problems such as compressed sensing, matrix completion, and robust and stable principal component analysis.

Keywords: Sparse recovery, First-order methods, Moreau envelope.

1. Introduction

1.1. Background. Owing to the well-known theory of compressed sensing [1, 2], sparse recovery has attracted much attention in the last decade in different scientific research fields such as applied mathematics [3, 4, 5], medicine [6], and in particular signal and image processing [7]. Indeed, many problems in practical applications can often be reduced to recover a unknown sparse signal $x_0 \in \mathbb{R}^n$ (a signal is sparse with respect to the canonical basis, or more general with respect to a basis or even redundant dictionary

[*] Corresponding author.

Email addresses: sparseelad@126.com (Yongchao Yu), jgpengxjtu@126.com (Jigen Peng)

[7]) from very few nonadaptive, linear and noisy data $b \in \mathbb{R}^m$, that is,

$$b = Ax_0 + e, \tag{1}$$

where A is a known $m \times n$ ($m \ll n$) sensing matrix and $e \in \mathbb{R}^n$ is a noise term. Compressed sensing [3, 4, 5] has stated that one can recover x_0 accurately and efficiently via the constrained ℓ_1 minimization problem:

$$(\mathrm{BP}_\varepsilon) \quad \begin{array}{ll} \text{minimize} & ||x||_1 \\ \text{subject to} & ||Ax - b||_2 \leq \varepsilon, \end{array} \tag{2}$$

where ε is an estimated upper bound of the noise level. When the data b are noise-free, problem (2) with $\varepsilon = 0$ is reduced to the well-known Basis Pursuit (BP) in the field of signal processing [8]:

$$(\mathrm{BP}) \quad \begin{array}{ll} \text{minimize} & ||x||_1 \\ \text{subject to} & b = Ax. \end{array} \tag{3}$$

(BP) is the convex relaxation of the original ℓ_0-minimization problem in the compressed sensing:

$$(\mathrm{P}_0) \quad \begin{array}{ll} \text{minimize} & ||x||_0 \\ \text{subject to} & b = Ax, \end{array} \tag{4}$$

where $\|x\|_0$ denotes the number of nonzero components of the signal x.

Another more theoretically efficient estimator for recovering sparse signals from noisy data is the Dantzig selector introduced in [9], which is the solution to the convex problem:

$$(\mathrm{DS}) \quad \begin{array}{ll} \text{minimize} & ||x||_1 \\ \text{subject to} & ||D^{-1}A^*(Ax - b)||_\infty \leq \gamma, \end{array} \tag{5}$$

where γ is a scalar related to the noise level and D is the diagonal matrix whose diagonal entries are the ℓ_2 norms of the columns of A. Obviously, the constraint in (5) requires that the correlation between the residual vector $r = b - Ax$ and the normalized columns of A is small. (DS) is based on the fact that, if the noise term e is stochastic, it obeys $\|D^{-1}A^*e\|_\infty \leq \gamma$ with high probability for some small scalar γ. Some researches [10, 11, 12, 13] which focus on developing numerical algorithms for solving directly three convex models, have demonstrated that these estimators are able to estimate accurately unknown sparse signals of interest.

In our tour, a closely related problem to compressed sensing is the problem of recovering an unknown low-rank matrix from incomplete, or even corrupted samples of its entries. We first focus on a special case of low-rank matrix recovery problems, namely, the low-rank matrix completion (MC) [14], which captures many applications such as the well-known Netflix problem [14], system identification [15], global positioning [16]. Let $M \in \mathbb{R}^{n \times n}$ be an unknown low-rank matrix that is partially known on a subset Ω of the complete set of entries $\{1, \cdots, n\} \times \{1, \cdots, n\}$. The exact MC problem is to seek the lowest rank matrix that agrees with M on Ω, which leads to the nonconvex optimization problem:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & \mathcal{P}_\Omega X = \mathcal{P}_\Omega M, \end{array} \tag{6}$$

where

$$[\mathcal{P}_\Omega X]_{ij} = \begin{cases} x_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{if } (i,j) \notin \Omega, \end{cases}$$

denotes the orthogonal projector onto the span of matrices vanishing outside of Ω. However, solving (6) is often numerically expensive. Notice that the nuclear norm of X defined as $\|X\|_* = \sum_{i=1}^n \sigma_i(X)$ is the convex envelop of $\text{rank}(X)$ on the set $\{X \in \mathbb{R}^{n\times n} : \ \|X\| \le 1\}$, where $\|\cdot\|$ is the operator norm of X. E. Candès et al. [14] therefore consider the following convex optimization problem:

$$\begin{array}{ll} \text{minimize} & ||X||_* \\ \text{subject to} & \mathcal{P}_\Omega X = \mathcal{P}_\Omega M. \end{array} \tag{7}$$

In practice, if we sample noisy entries $Y_{ij} = M_{ij} + E_{ij}$, $(i,j) \in \Omega$, where E_{ij} is a noise term. To recover the unknown matrix M, the modified model [15] is:

$$\begin{array}{ll} \text{minimize} & ||X||_* \\ \text{subject to} & \|\mathcal{P}_\Omega X - \mathcal{P}_\Omega Y\|_F \le \varepsilon. \end{array} \tag{8}$$

A considerable body of theoretical work [14, 15, 17] is devoted to show that, under suitable conditions, the underlying low-rank matrix M can be exactly or accurately recovered via the above two convex models with high probability when Ω is sampled uniformly at random.

The second low-rank matrix recovery problem we study is robust PCA (RPCA), which was introduced by E. Candès et al. in [18] to overcome the shortcoming of classical PCA. It is well-known that classical PCA captures the low-rank structure of the observation matrix corrupted by random Gaussian noises. However, when the observation data is corrupted by gross errors, classical PCA becomes impractical because gross errors may be arbitrary large amplitude. Under the assumption that the gross errors are sparse, RPCA wishes to decompose data matrix D ($D = L_0 + S_0$) into the low-rank matrix L_0 corresponding to the low-dimensional structure and the sparse matrix part S_0 corresponding to the sparse errors. This leads to the optimization problem:

$$\begin{array}{ll} \text{minimize} & \text{rank}(L) + \tau\|S\|_0 \\ \text{subject to} & D = L + S, \end{array} \tag{9}$$

where $\|S\|_0$ counts the number of nonzero components of the matrix S and the parameter $\tau > 0$ balances the weights of rank and sparsity. Since (9) is also a NP-hard problem. Under some conditions, it can be relaxed to the convex optimization:

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \tau\|S\|_1 \\ \text{subject to} & D = L + S. \end{array} \tag{10}$$

Problem (10) is called the robust principal component pursuit (RPCP) [18]. In many applications, however, the data matrix of interest is not only corrupted by gross sparse errors but also small dense noises caused by basic measurement inaccuracies, quantization and so on. The authors in [19] therefore studied the problem of recovering the low-rank matrix from the data matrix $D = L_0 + S_0 + E$ (E is a dense noise term), and showed that the solution to the following convex program (known as the stable principal component pursuit (SPCP))

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \tau\|S\|_1 \\ \text{subject to} & \|L + S - D\|_F \leq \varepsilon, \end{array} \tag{11}$$

gives an estimate of the low-rank matrix that is simultaneously stable to small dense noises and robust to gross sparse errors.

In this paper, our goal is to develop efficient first-order algorithms for solving various convex models mentioned above.

1.2. Related first-order methods.

Different approaches have been proposed to handle each of the above convex models such as the traditional primal-dual interior-point methods [10] for both models (BP) and (DS), interior-point algorithms for (BP$_\varepsilon$), alternating direction method of multipliers (ADMM) [11, 20, 21] for solving uniformly these models, and Bregman distance-based methods [12] for (BP). However, we here want to review a class of first-order methods which mainly involve gradient operators.

Since it is not easy to solve the constrained optimization models such as (BP) and (BP$_\varepsilon$), a popular model in the literature is the ℓ_1-regularized least squares problem (also known as the basis pursuit de-noising problem [8]):

$$(\text{QP}_\lambda) \quad \text{minimize} \quad F(x) = \tfrac{1}{2}||Ax - b||_2^2 + \lambda||x||_1, \tag{12}$$

where λ is a given positive parameter. Further, since problem (12) can be cast as a quadratic programming (QP), it is also denoted by (QP$_\lambda$). For matrix completion and robust PCA problems, we can consider their corresponding regularized unconstrained versions. One of the most widely used first-order methods for solving (12) is ISTA (iterative shrinkage-thresholding algorithm [22, 23]), which only requires one gradient evaluation and proximity operation at each iteration, and obtains $\mathcal{O}(\frac{1}{k})$ rate of convergence of the objective function $F(x)$:

$$F(x_k) - F(x^*) \sim \mathcal{O}\left(\frac{1}{k}\right),$$

where x^* is an optimal solution to (12) and k is the number of iterations. An efficient accelerated version of it is FISTA (fast iterative shrinkage-thresholding algorithm [23]) which uses linear interpolation strategy to achieve $\mathcal{O}(\frac{1}{k^2})$ iteration complexity:

$$F(x_k) - F(x^*) \sim \mathcal{O}\left(\frac{1}{k^2}\right).$$

Furthermore, FISTA can address more general problem of minimizing the objective function which is the sum of a convex smooth function with Lipschitz continuous gradient and a nonsmooth simple convex function in the sense that its proximity operator is efficiently evaluated. For example, FISTA has been extended to tackle the regularized unconstrained versions [24, 25] of models (8) and (11). However, model (QP$_\lambda$) and FISTA also have some disadvantages: first, although from convex optimization theory, solutions to both models (BP$_\varepsilon$) and (QP$_\lambda$) are same for appropriate parameters ε and λ, we do not know how to select a good parameter λ in model (QP$_\lambda$) to obtain a solution to the original model (BP$_\varepsilon$); second, FISTA can not solve the original model (BP$_\varepsilon$) since the objective function of the model does not possess Lipschitz continuous gradient.

In order not to introduce a parameter λ, NESTA developed in [13] smoothes the ℓ_1 norm and then applies the Nesterov's optimal gradient method [26] to solve the smoothed model (BP$_\varepsilon$). Obviously, single smoothing parameter in NESTA controls the closeness of solutions to model (BP$_\varepsilon$) and its smoothed version. NESTA obtains $\mathcal{O}(\frac{1}{k^2})$ rate of convergence of the objective function of the smoothed model. However, NESTA is customized to a special kind of sensing matrices A, that is, the rows of A are orthonormal, which makes projection onto the constraint set tractable. Hence, NESTA is not suitable for general sensing matrices. Most importantly, it is also not a good algorithm for solving model (DS). TFOCS [27] is a clever approach for alleviating the difficulty of projection problems via reformulating these original sparse recovery models as equivalent convex conic formulations. Unlike FISTA and NESTA, TFOCS does not provide analysis for the rate of convergence.

1.3. Our contributions. In this work, our contributions are three aspects. First, we obtain more suitable unconstrained versions for these sparse recovery models, especially for model (DS), via smoothing indicator functions of some sets related to constraint sets. Second, an efficient first-order method is developed via smoothing one of object functions for tackling the general problem of minimizing the sum of a Lipschitz continuous nonsmooth convex function and a nonsmooth simple convex function. Our method modifies FISTA by using variable smoothing parameter which is updated in each iteration, and yields the rate of convergence of order $\mathcal{O}(\frac{\log k}{k})$ for values of the objective function. At last, we apply the proposed methods to the above sparse recovery problems and test the performance of methods.

1.4. Organization. The rest of the paper is organized as follows: In Section 2, we briefly give some preliminaries from convex analysis, such as the proximity operator of a convex function and its Moreau envelope. In Section 3, two smoothing approaches are considered, and then two first-order methods are proposed. Applications of two types of methods to sparse recovery problems are given in Section 4. Some conclusions are drawn in Section 5.

2. Preliminaries

In this paper, $\mathcal{H}$ denotes a general finite dimensional Hilbert space endowed with the inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. We also denote the n-dimensional Euclidean space by $\mathbb{R}^n$ equipped with the usual scalar product $\langle x,y\rangle = \sum_{i=1}^n x_i y_i$ for $x,\, y \in \mathbb{R}^n$ and the associated norm $||x|| = \sqrt{\langle x,x\rangle}$. Furthermore, $\mathbb{R}^{m\times n}$ is the matrix space with the inner product $\langle X,Y\rangle = \mathrm{trace}(X^*Y)$ ($\|X\|_F = \sqrt{\langle X,X\rangle}$). For an extended value function $f : \mathcal{H} \to \bar{\mathbb{R}} = \mathbb{R}\cup\{+\infty\}$, we denote by $\mathrm{dom} f = \{x \in \mathcal{H} : f(x) < +\infty\}$ its effective domain. a function f is proper if $\mathrm{dom} f \neq \emptyset$. The set of all proper, lower semicontinuous convex functions from $\mathcal{H} \to \bar{\mathbb{R}}$ is denoted by $\Gamma_0(\mathcal{H})$. The indicator function of a closed nonempty convex set $C \subseteq \mathcal{H}$ at $x \in \mathcal{H}$ is defined as

$$\delta_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{if } x \notin C. \end{cases}$$

Clearly, the indicator function δ_C is in $\Gamma_0(\mathcal{H})$. A function f is strongly convex with a parameter $\rho > 0$ if for all $x, y \in \text{dom} f$ and all $\alpha \in (0, 1)$, the following inequality holds:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\rho}{2}\alpha(1-\alpha)\|x-y\|^2. \tag{13}$$

A continuously differentiable convex function f has Lipschitz continuous gradient with constant L, if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, \qquad \forall\, x, y \in \text{dom} f. \tag{14}$$

Definition 2.1 (subdifferential [28]). Let a function f be in $\Gamma_0(\mathcal{H})$. A vector g is called the subgradient of f at point $x \in \text{dom} f$ if for any $y \in \text{dom} f$, the following inequality holds:

$$f(y) - f(x) \geq \langle g, y - x \rangle. \tag{15}$$

The set of all subgradients of f at x is called the subdifferential of function f at x and is denoted as $\partial f(x)$. Furthermore, if $x \notin \text{dom} f$, clearly, $\partial f(x) = \emptyset$.

The notion of subdifferential is a generalization of the gradient of a smooth convex function and is very important to solve convex optimization problems. The first-order optimality condition is given.

Theorem 2.2. Let a function f be in $\Gamma_0(\mathcal{H})$. We have $\text{Argmin} f = \{x \in \text{dom} f \mid 0 \in \partial f(x)\}$.

We now introduce the definitions of the proximity operator and Moreau envelope of a convex function, which are at the center of the proposed first-order methods in section 3.

Definition 2.3 (proximity operator). The proximity operator of parameter β of a function f in $\Gamma_0(\mathcal{H})$ is a mapping from $\mathcal{H}$ to $\mathcal{H}$, defined as

$$\text{prox}_{\beta f}(x) = \underset{y \in \mathcal{H}}{\text{argmin}} \left\{ \frac{1}{2\beta}||y - x||^2 + f(y) \right\}. \tag{16}$$

It follows from [29] that $\text{prox}_{\beta f}$ is single-valued and firmly nonexpansive, i.e.,

$$\|\text{prox}_{\beta f}(x) - \text{prox}_{\beta f}(y)\|^2 \leq \langle \text{prox}_{\beta f}(x) - \text{prox}_{\beta f}(y), x - y \rangle, \quad \forall\, x, y \in \mathcal{H}. \tag{17}$$

Definition 2.4 (Moreau envelope [29]). The Moreau envelope of parameter β of a function f in $\Gamma_0(\mathcal{H})$ is the function $f^\beta : \mathcal{H} \to \mathbb{R}$, defined as

$$f^\beta(x) = \inf_{y \in \mathcal{H}} \left\{ \frac{1}{2\beta}||y - x||^2 + f(y) \right\}. \tag{18}$$

Note that the Moreau envelope is convex, real-valued, and continuous [29]. It is well-known that using the Moreau envelope can efficiently smooth a non-smooth convex function. Indeed, the Moreau envelope f^β of a function f in $\Gamma_0(\mathcal{H})$ converges to f as $\beta \to 0^+$. More importantly, the next proposition states that f^β is actually Fréchet differentiable on $\mathcal{H}$ [29].

Proposition 2.5. Let a function f be in $\Gamma_0(\mathcal{H})$ and $\beta > 0$. Then $f^\beta : \mathcal{H} \to \mathbb{R}$ is Fréchet differentiable on $\mathcal{H}$, and its gradient

$$\nabla f^\beta(x) = \frac{1}{\beta}\Big(x - \operatorname{prox}_{\beta f}(x)\Big), \quad \forall\, x \in \mathcal{H}, \tag{19}$$

is clearly $\frac{1}{\beta}$-Lipschitz continuous.

Proof: For $x, y \in \mathcal{H}$, we set $a = \operatorname{prox}_{\beta f}(x)$, $b = \operatorname{prox}_{\beta f}(y)$ and

$$c = (1-\alpha)a + \alpha b, \quad \alpha \in (0,1).$$

Obviously, $a, b, c \in \operatorname{dom} f$. According to Definition 2.3 and the convexity of f, the following inequalities hold:

$$\begin{aligned} f(a) &\leq f(c) + (2\beta)^{-1}\|x-c\|^2 - (2\beta)^{-1}\|x-a\|^2 \\ &\leq (1-\alpha)f(a) + \alpha f(b) + (2\beta)^{-1}[2\alpha\langle x-a, a-b\rangle + \alpha^2\|a-b\|^2]. \end{aligned}$$

Hence, we have that

$$f(b) - f(a) \geq (2\beta)^{-1}[2\langle x-a, b-a\rangle - \alpha\|a-b\|^2].$$

Letting $\alpha \to 0$, we can have the inequality:

$$f(b) - f(a) \geq \beta^{-1}\langle x-a, b-a\rangle. \tag{20}$$

Likewise, we can also obtain that:

$$f(a) - f(b) \geq \beta^{-1}\langle y-b, a-b\rangle, \tag{21}$$

that is,

$$f(b) - f(a) \leq \beta^{-1}\langle y-b, b-a\rangle. \tag{22}$$

By using Definition 2.3 and the inequality (20), we have that

$$\begin{aligned} f^\beta(y) - f^\beta(x) &= f(b) - f(a) + (2\beta)^{-1}(\|y-b\|^2 - \|x-a\|^2) \\ &\geq (2\beta)^{-1}(2\langle x-a, b-a\rangle + \|y-b\|^2 - \|x-a\|^2) \\ &= (2\beta)^{-1}(2\langle y-x, x-a\rangle + \|y-b-x+a\|^2) \\ &\geq \beta^{-1}\langle y-x, x-a\rangle. \end{aligned} \tag{23}$$

It follows from the inequality (22) that

$$\begin{aligned} f^\beta(y) - f^\beta(x) &= f(b) - f(a) + (2\beta)^{-1}(\|y-b\|^2 - \|x-a\|^2) \\ &\leq (2\beta)^{-1}(2\langle y-b, b-a\rangle + \|y-b\|^2 - \|x-a\|^2) \\ &= (2\beta)^{-1}(2\langle y-x, y-b\rangle - \|y-b-x+a\|^2) \\ &\leq \beta^{-1}\langle y-x, y-b\rangle. \end{aligned} \tag{24}$$

The two inequalities (23) and (24) lead to

$$\begin{array}{rcl} 0 & \leq & f^\beta(y) - f^\beta(x) - \beta^{-1}\langle y-x, x-a\rangle \\ & \leq & \beta^{-1}\langle y-x, (y-b)-(x-a)\rangle \\ & = & \beta^{-1}(\|y-x\|^2 + \langle y-x, a-b\rangle). \end{array}$$

In addition, the firm nonexpansiveness of $\mathrm{prox}_{\beta f}$ indicates that

$$\langle y-x, a-b\rangle \leq -\|a-b\|^2.$$

Thus,

$$0 \leq f^\beta(y) - f^\beta(x) - \beta^{-1}\langle y-x, x-a\rangle \leq \beta^{-1}\|y-x\|^2. \tag{25}$$

At last, we have that

$$\lim_{y\to x} \frac{f^\beta(y) - f^\beta(x) - \langle y-x, \beta^{-1}(x-a)\rangle}{\|y-x\|} = 0. \tag{26}$$

This indicates that

$$\nabla f^\beta(x) = \frac{1}{\beta}(x - \mathrm{prox}_{\beta f}(x)).$$

Furthermore, Lipschitz continuity follows from (17). this proof is completed. ♯

In particular, for any closed nonempty convex set C, we get that

$$\mathrm{prox}_{\beta\delta_C} = \mathcal{P}_C, \quad \delta_C^\beta(x) = \frac{1}{2\beta}\|x - \mathcal{P}_C(x)\|^2, \quad \nabla\delta_C^\beta(x) = \beta^{-1}(x - \mathcal{P}_C(x)),$$

where

$$\mathcal{P}_C(x) = \operatorname*{argmin}_{y\in C} \|y-x\|,$$

denotes the projection operator on C.

Furthermore, since some norms, such as ℓ_p $(p = 1, 2, \infty)$ for vectors, nuclear norm and operator norm for matrices, and constraint sets related to these dual norms-type balls, are often involved in our concerned sparse recovery models, it is very interesting to investigate the relationship between the proximity operator of a norm and that of the indicator function of the dual norm-type ball. The next result is essentially from the Moreau's decomposition principle [29] applied to a norm function.

Proposition 2.6. Let $\|\cdot\|_p$ denote a primal norm in $\mathcal{H}$ and $\|\cdot\|_d$ be its dual norm defined as

$$\|x\|_d = \sup_{\|y\|_p\leq 1} \langle y, x\rangle.$$

For $\beta > 0$, the following equality holds for all $x \in \mathcal{H}$:

$$\mathrm{prox}_{\beta\|\cdot\|_p}(x) + \mathcal{P}_C(x) = x, \tag{27}$$

where the set $C = \{y \in \mathcal{H} : \|y\|_d \leq \beta\}$.

Let us illustrate the proximity operators and Moreau envelopes of ℓ_1, ℓ_2, and nuclear norm of interest, respectively. For the norm $\|\cdot\|_1$ in $\mathbb{R}^n$, its proximity operator with parameter β is the well-known soft-thresholding operator

$$\operatorname{prox}_{\beta\|\cdot\|_1}(x) = \operatorname{sgn}(x) \odot \max\{|x| - \beta, 0\}; \tag{28}$$

its Moreau envelope with parameter β is sum of Huber function on each of the components, which can be expressed as

$$\|\cdot\|_1^\beta(x) = \sum_{i=1}^n H_\beta(x_i), \qquad H_\beta(y) = \begin{cases} |y| - \frac{\beta}{2}, & |y| > \beta, \\ \frac{y^2}{2\beta}, & |y| \le \beta. \end{cases} \tag{29}$$

Likewise, the proximity operator and Moreau envelope of parameter β of $\|\cdot\|_2$ in $\mathbb{R}^n$ is, respectively

$$\operatorname{prox}_{\beta\|\cdot\|_2}(x) = \begin{cases} \max\left\{1 - \frac{\beta}{\|x\|_2}, 0\right\} x, & x \neq 0, \\ 0, & x = 0. \end{cases} \tag{30}$$

$$\|\cdot\|_2^\beta(x) = \begin{cases} \|x\|_2 - \frac{\beta}{2}, & \|x\|_2 > \beta, \\ \frac{\|x\|_2^2}{2\beta}, & \|x\|_2 \le \beta. \end{cases} \tag{31}$$

It follows from [17] that the proximity operator of parameter β of nuclear norm $\|\cdot\|_*$ is defined at a matrix $X \in \mathbb{R}^{m\times n}$ of rank r as

$$\operatorname{prox}_{\beta\|\cdot\|_*}(X) = U S_\beta(\Sigma) V^T, \quad S_\beta(\Sigma) = \operatorname{diag}(\max\{\sigma_i - \beta, 0\}), \tag{32}$$

where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$ are obtained by the singular value decomposition (SVD) of X: $X = U\Sigma V^T$, and $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r)$. Furthermore, its Moreau envelope is given by the sum of Huber function on each singular value, which can be expressed as

$$\|\cdot\|_*^\beta(X) = \sum_{i=1}^r H_\beta(\sigma_i). \tag{33}$$

3. Efficient first-order methods based on the Moreau envelope

3.1. The first efficient first-order method

A unified formulation of general sparse recovery models is considered:

$$\text{minimize} \quad h(x) + \delta_C(\mathcal{A}x - d). \tag{34}$$

Here, we assume that a linear operator $\mathcal{A}$ is from a proper space $\mathcal{H}$ to itself or other proper space $\mathcal{H}'$, d is a given vector, the set C is a given closed convex set. The proximity operators of h and δ_C are also easily computed. Our first smoothing approach is to approximate the indicator function δ_C via using its Moreau envelope, which leads to a more appropriate unconstrained model for handling sparse recovery from noisy data:

$$\text{minimize} \quad h(x) + \delta_C^\beta(\mathcal{A}x - d). \tag{35}$$

It can be seen that the gradient of $\delta_C^\beta(\mathcal{A}x - d)$ is

$$\nabla(\delta_C^\beta(\mathcal{A}x - d)) = \mathcal{A}^*\nabla(\delta_C^\beta)(\mathcal{A}x - d) = \beta^{-1}\mathcal{A}^*(\mathcal{A}x - d - \mathcal{P}_C(\mathcal{A}x - d)). \tag{36}$$

Further, for $x, y \in \mathcal{H}$, we have that

$$\begin{aligned}\|\nabla(\delta_C^\beta(\mathcal{A}x - d)) - \nabla(\delta_C^\beta(\mathcal{A}y - d))\| &\leq \beta^{-1}\|\mathcal{A}\|\, \|(\mathcal{A}x - d - \mathcal{P}_C(\mathcal{A}x - d)) \\ &- (\mathcal{A}x - d - \mathcal{P}_C(\mathcal{A}x - d))\| \\ &\leq \beta^{-1}\|\mathcal{A}\|^2\, \|x - y\|, \end{aligned} \tag{37}$$

which shows that $\nabla(\delta_C^\beta(\mathcal{A}(x) - d))$ is $\beta^{-1}\|\mathcal{A}\|^2$–Lipschitz continuous.

We now apply this smoothing approach to sparse recovery models. More specifically, in model (BP$_\varepsilon$), the set C is a ℓ_2-ball in $\mathbb{R}^m$ centered at the origin with radius ε, i.e., $C = \{y \in \mathbb{R}^m : \|y\|_2 \leq \varepsilon\}$. Based on Proposition 2.6, we can easily obtain

$$\delta_C^\beta(y) = \frac{1}{2\beta}\Big\|\mathrm{prox}_{\varepsilon\|\cdot\|_2}(y)\Big\|_2^2 = \frac{1}{2\beta}\Big(\max\{\|y\|_2 - \varepsilon, 0\}\Big)^2. \tag{38}$$

Combining (2), (35) and (38), we have the following unconstrained optimization model

$$\text{minimize} \quad \tfrac{1}{2\beta}\Big(\max\{\|Ax - b\|_2 - \varepsilon, 0\}\Big)^2 + \|x\|_1. \tag{39}$$

Note that when $\varepsilon = 0$, (39) is reduced to

$$\text{minimize} \quad \tfrac{1}{2\beta}\|Ax - b\|_2^2 + \|x\|_1, \tag{40}$$

which is equivalent to the ℓ_1-regularized least squares problem (12) if the parameter λ equals to the smoothing parameter β. In other word, model (40) or original ℓ_1-regularized least squares model (12) is proper to approximate model (BP) in the free-noise case, However, model (39) is more suitable to approximate model (BP$_\varepsilon$) in the noisy case. It is expected that using slightly modified model (39) can improve the performance of sparse recovery from noisy data.

For model (DS), the set C is $\{y \in \mathbb{R}^n : \|y\|_\infty \leq \gamma\}$. Again, according to Proposition 2.6, the following model is proposed

$$\text{minimize} \quad \tfrac{1}{2\beta}\Big\|\mathrm{prox}_{\gamma\|\cdot\|_1}(D^{-1}A^*Ax - D^{-1}A^*b)\Big\|_2^2 + \|x\|_1. \tag{41}$$

To the best of our knowledge, (41) is the first reasonable unconstrained version of model (DS).

Likewise, for matrix completion problem, model (8) can be approximated by

$$\text{minimize} \quad \tfrac{1}{2\beta}\Big(\max\{\|\mathcal{P}_\Omega(X - Y)\|_F - \varepsilon, 0\}\Big)^2 + \|X\|_1. \tag{42}$$

Moreover, stable principal component pursuit (SPCA) should be modified to

$$\text{minimize} \quad \tfrac{1}{2\beta}\Big(\max\{\|L + S - D\|_F - \varepsilon, 0\}\Big)^2 + \|L\|_* + \tau\|S\|_1. \tag{43}$$

Obviously, these smoothed models are just fall into the framework of application of FISTA. We only apply FISTA to the smoothed model (DS) (41):

Algorithm 1 FISTA for the smoothed (DS) (41)

```
Input: $y_1 = x_1 \in \mathbb{R}^n$; $t_1 = 1$; $\beta > 0$; $\gamma > 0$.
Output: $x_K$.
for $k = 0$ to $K$ do
    $x_{k+1} = \mathrm{prox}_{1/L\|\cdot\|_1}\Big(y_k - L^{-1}\beta^{-1}A^*AD^{-1}\,\mathrm{prox}_{\gamma\|\cdot\|_1}\Big(D^{-1}A^*(Ay_k - b)\Big)\Big)$;
    $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
    $y_{k+1} = x_{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_{k+1} - x_k)$;
end for
```

3.2. The second efficient first-order method

Another more general framework of sparse recovery models is formulated:

$$\text{minimize} \quad H(x) = h(x) + g(x), \tag{44}$$

where we assume that the (nonsmooth) function h is L_h-Lipschitz continuous and proximity operators of h and g can be efficiently calculated. Our second approach for solving the harder model (44) is to smooth h via its Moreau envelope and then develop an efficient first-order algorithm. The proposed algorithm modifies FISTA to obtain the rate of convergence of order $\mathcal{O}(\frac{\log k}{k})$ for the iterative sequence of values of the initial objective function H. More specifically, as in FISTA, we yield two iterative sequences, x_k and y_k. We update x_k at each iteration via a proximal gradient step using variable smoothing parameter $\beta_k = 1/k$, that is,

$$x_{k+1} = \mathrm{prox}_{\beta_k g}\Big(y_k - \beta_k \nabla h^{\beta_k}(y_k)\Big),$$

and then generate y_k via a linear combination of x_{k+1} and x_k, i.e.,

$$y_{k+1} = x_{k+1} + \left(\frac{k-1}{k+1}\right)(x_{k+1} - x_k).$$

Our proposed algorithm is described in Algorithm 2 as follows:

Algorithm 2 The proposed algorithm for model (44)

```
Input: $y_1 = x_1 \in \mathrm{dom}\, g$; $\beta_1 = 1$.
Output: $x_K$.
for $k = 0$ to $K$ do
    $\beta_k = \frac{1}{k}$;
    $x_{k+1} = \mathrm{prox}_{\beta_k g}\Big(\mathrm{prox}_{\beta_k h}(y_k)\Big)$;
    $y_{k+1} = x_{k+1} + \left(\frac{k-1}{k+1}\right)(x_{k+1} - x_k)$;
end for
```

Before establishing the rate of convergence of our proposed algorithm, we prove some lemmas of importance.

Lemma 3.1. Let a function $f : \mathcal{H} \to \mathbb{R}$ be convex and L_f-Lipschitz continuous. It holds

(a) if $\beta > 0$, then $f^\beta \le f \le f^\beta + \frac{1}{2}\beta L_f^2$;
(b) if $\beta_1 > \beta_2 > 0$, then $f^\beta \le f \le f^\beta + \frac{1}{2}(\beta_1 - \beta_2)L_f^2$.

Proof: It follows from Definition 2.4 that

$$f(x) = \frac{1}{2\beta}||x-x||^2 + f(x) \ge \inf_{y\in\mathcal{H}}\left\{\frac{1}{2\beta}||y-x||^2 + f(y)\right\} = f^\beta(x), \tag{45}$$

which proves the left hand side of the property (a). Further, the Lipschitz property of f leads to

$$\frac{1}{2\beta}||y-x||^2 + f(y) + f(x) - f(x) \ge \frac{1}{2\beta}||y-x||^2 + f(x) - L_f\|y-x\|. \tag{46}$$

Hence,

$$f^\beta(x) \ge f(x) + \inf_{y\in\mathcal{H}}\left\{\frac{1}{2\beta}||y-x||^2 - L_f\|y-x\|\right\} = f(x) - \frac{1}{2}\beta L_f^2. \tag{47}$$

This completes the proof for the other side of the property (a).

According to (a) and then setting $\beta = \beta_1 - \beta_2$ and $f = f^{\beta_2}$, we obtain that

$$(f^{\beta_2})^{\beta_1-\beta_2} \le f^{\beta_2} \le (f^{\beta_2})^{\beta_1-\beta_2} + \frac{1}{2}(\beta_1-\beta_2)L_f^2. \tag{48}$$

For proving the property (b), we only require the equality $(f^{\beta_2})^{\beta_1-\beta_2} = f^{\beta_1}$. Indeed,

$$\begin{aligned}
(f^{\beta_2})^{\beta_1-\beta_2}(x) &= \inf_{y\in\mathcal{H}}\left\{\frac{1}{2(\beta_1-\beta_2)}\|y-x\|^2 + \inf_{z\in\mathcal{H}}\left\{\frac{1}{2\beta_2}\|z-y\|^2 + f(z)\right\}\right\} \\
&= \inf_{z\in\mathcal{H}}\left\{f(z) + \frac{\beta_1}{2\beta_2(\beta_1-\beta_2)}\inf_{y\in\mathcal{H}}\left\{\frac{\beta_2}{\beta_1}\|y-x\|^2 + \frac{\beta_1-\beta_2}{\beta_1}\|z-y\|^2\right\}\right\} \\
&= \inf_{z\in\mathcal{H}}\Bigg\{f(z) \\
&+ \frac{\beta_1}{2\beta_2(\beta_1-\beta_2)}\inf_{y\in\mathcal{H}}\left\{\left\|y - (\frac{\beta_2}{\beta_1}x + \frac{\beta_1-\beta_2}{\beta_1}z)\right\|^2 + \frac{\beta_2(\beta_1-\beta_2)}{\beta_1^2}\|z-x\|^2\right\}\Bigg\} \\
&= \inf_{z\in\mathcal{H}}\left\{f(z) + \frac{1}{2\beta_1}\|z-x\|^2\right\} = f^{\beta_1}(x).
\end{aligned} \tag{49}$$

Hence, the property (b) is established. ♯

Lemma 3.2. Let a function $f : \mathcal{H} \to \mathbb{R}$ be convex and have $L_{\nabla f}$-Lipschitz continuous gradient. The following inequality holds for all $x, y \in \text{dom} f$

$$f(x) \le f(y) + \langle \nabla f(y), x-y\rangle + \frac{L_{\nabla f}}{2}\|x-y\|^2. \tag{50}$$

Proof: This proof is classical. By using the mean value theorem, we have that

$$\begin{aligned} f(x) &= f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x-y \rangle \,\mathrm{d}t \\ &= f(y) + \langle \nabla f(y), x-y \rangle + \int_0^1 \langle \nabla f(y+t(x-y)) - \nabla f(y), x-y \rangle \,\mathrm{d}t. \end{aligned} \tag{51}$$

It follows from the Cauchy-Schwarz inequality and the inequality (14) that

$$\begin{aligned} f(x) &= f(y) + \int_0^1 \langle \nabla f(y + t(x-y)), x-y \rangle \,\mathrm{d}t \\ &\leq f(y) + \langle \nabla f(y), x-y \rangle + \int_0^1 \|\nabla f(y+t(x-y)) - \nabla f(y)\| \|x-y\| \,\mathrm{d}t \\ &\leq f(y) + \langle \nabla f(y), x-y \rangle + \int_0^1 L_{\nabla f} t \|x-y\| \,\mathrm{d}t \\ &= f(y) + \langle \nabla f(y), x-y \rangle + \frac{L_{\nabla f}}{2}\|x-y\|^2. \end{aligned} \tag{52}$$

We obtain the inequality (50). ♯

Lemma 3.3. Let a function f be strongly convex with paramter ρ. The following inequality holds for all $x \in \mathrm{dom} f$ and $x^* \in \mathrm{Argmin} f$

$$f(x) \geq f(x^*) + \frac{\rho}{2}\|x - x^*\|^2. \tag{53}$$

Proof: This proof is simple. For $x, y \in \mathrm{dom} f$, $z = \alpha x + (1-\alpha)y$, $\alpha \in (0,1)$, it can be induced from the inequality (13) that

$$f(x) \geq f(y) + \frac{1}{\alpha}\left[f(y + \alpha(x-y)) - f(y)\right] + \frac{\rho(1-\alpha)}{2}\|x-y\|^2. \tag{54}$$

Let g be a subgradient of f at y. We have

$$f(y + \alpha(x-y)) - f(y) \geq \alpha \langle g, x-y \rangle. \tag{55}$$

Hence,

$$f(x) \geq f(y) + \langle g, x-y \rangle + \frac{\rho(1-\alpha)}{2}\|x-y\|^2. \tag{56}$$

Tending α to 0, we get

$$f(x) \geq f(y) + \langle g, x-y \rangle + \frac{\rho}{2}\|x-y\|^2. \tag{57}$$

Setting $y = x^*$, and then letting g be 0, we get the inequality (53). ♯

The iteration complexity for values of the objective function via the proposed algorithm 2 is now presented.

Theorem 3.1. For any initial values $x_1 = y_1 \in \operatorname{dom} H$ and any optimal solution $x^* \in \operatorname{Argmin} H$, the iterative sequence of values of the objective function $H(x_k)$ generated by the proposed algorithm satisfies

$$H(x_{k+1}) - H(x^*) \leq \frac{1}{2k}\Big(\|x_1 - x^*\|^2 + L_h^2(\log k + 1)\Big), \tag{58}$$

where L_h is a Lipschitz constant of the function h in model (44).

Proof: We first set $H^{\beta_k}(x) = h^{\beta_k}(x) + g(x)$. Since h^{β_k} has Lipschitz continuous gradient with Lipschitz constant $1/\beta_k$, it follows from Lemma 3.2 that

$$H^{\beta_k}(x_{k+1}) \leq h^{\beta_k}(y_k) + \langle \nabla h^{\beta_k}(y_k), x_{k+1} - y_k\rangle + \frac{1}{2\beta_k}\|x_{k+1} - y_k\|^2 + g(x_{k+1}). \tag{59}$$

Note that the function

$$x \to \langle \nabla h^{\beta_k}(y_k), x - y_k\rangle + \frac{1}{2\beta_k}\|x - y_k\|^2 + g(x)$$

is strongly convex with parameter $1/\beta_k$ and its minimizer is nothing than x_{k+1}. Letting x^* be a minimizer of model (44) and taking $x = (1-\beta_k)x_k + \beta_k x^*$ in Lemma 3.3, we have

$$\begin{array}{rcl} H^{\beta_k}(x_{k+1}) & \leq & h^{\beta_k}(y_k) + g((1-\beta_k)x_k + \beta_k x^*) + \langle \nabla h^{\beta_k}(y_k), (1-\beta_k)x_k + \beta_k x^* - y_k\rangle \\ & + & \frac{1}{2\beta_k}\|(1-\beta_k)x_k + \beta_k x^* - y_k\|^2 - \frac{1}{2\beta_k}\|(1-\beta_k)x_k + \beta_k x^* - x_{k+1}\|^2. \end{array} \tag{60}$$

Based on the third step of the algorithm 2:

$$y_{k+1} = x_{k+1} + \frac{(1-\beta_k)\beta_{k+1}}{\beta_k}(x_{k+1} - x_k), \tag{61}$$

and the convexity of the function g, we also have the following inequality:

$$\begin{array}{rcl} H^{\beta_k}(x_{k+1}) & \leq & h^{\beta_k}(y_k) + (1-\beta_k)g(x_k) + \beta_k g(x^*) \\ & + & \langle \nabla h^{\beta_k}(y_k), (1-\beta_k)x_k + \beta_k x^* - y_k\rangle \\ & + & \frac{\beta_k}{2}\|z_k - x^*\|^2 - \frac{\beta_k}{2}\|z_{k+1} - x^*\|^2, \end{array} \tag{62}$$

where $z_k = (1 - \frac{1}{\beta_k})x_k + \frac{1}{\beta_k}y_k$. Furthermore, the convexity of the function h^{β_k} leads to

$$(1-\beta_k)h^{\beta_k}(y_k) + \langle \nabla h^{\beta_k}(y_k), (1-\beta_k)(x_k - y_k)\rangle \leq (1-\beta_k)h^{\beta_k}(x_k), \tag{63}$$

and

$$\beta_k h^{\beta_k}(y_k) + \langle \nabla h^{\beta_k}(y_k), \beta_k(x^* - y_k)\rangle \leq \beta_k h^{\beta_k}(x^*). \tag{64}$$

Hence,

$$H^{\beta_k}(x_{k+1}) \leq (1-\beta_k)H^{\beta_k}(x_k) + \beta_k H^{\beta_k}(x^*) + \frac{\beta_k}{2}\|z_k - x^*\|^2 - \frac{\beta_k}{2}\|z_{k+1} - x^*\|^2. \tag{65}$$

It can be induced from the property (a) and (b) of Lemma 3.1 that

$$\begin{aligned}\frac{H^{\beta_{k+1}}(x_{k+1})-H(x^*)}{\beta_k} &\leq \frac{H^{\beta_k}(x_{k+1})-H(x^*)}{\beta_k}+\frac{\beta_k-\beta_{k+1}}{2\beta_k}L_h^2\\ &\leq \frac{1}{2}(\|z_k-x^*\|^2-\|z_{k+1}-x^*\|^2)\\ &\quad +\frac{\beta_k-\beta_{k+1}}{2\beta_k}L_h^2+\frac{1-\beta_k}{\beta_k}(H^{\beta_k}(x_k)-H(x^*)). \qquad (66)\end{aligned}$$

Summing up the inequalities in (66) from 1 to k, we get

$$H^{\beta_{k+1}}(x_{k+1})-H(x^*) \leq \frac{\beta_k}{2}(\|z_1-x^*\|^2-\|z_{k+1}-x^*\|^2)+\frac{\beta_k}{2}\sum_{i=1}^{k}\frac{\beta_i-\beta_{i+1}}{\beta_i}L_h^2. \qquad (67)$$

Hence,

$$\begin{aligned}H(x_{k+1})-H(x^*) &\leq H^{\beta_{k+1}}(x_{k+1})-H(x^*)+\frac{\beta_{k+1}}{2}L_h^2\\ &\leq \frac{\beta_k}{2}\|x_1-x^*\|^2+\frac{\beta_{k+1}}{2}L_h^2+\frac{\beta_k}{2}\sum_{i=1}^{k}\frac{\beta_i-\beta_{i+1}}{\beta_i}L_h^2\\ &= \frac{\beta_k}{2}\Big(\|x_1-x^*\|^2+L_h^2\sum_{i=1}^{1}\frac{1}{\beta_k}\Big). \qquad (68)\end{aligned}$$

Finally, we use the elementary inequality

$$\sum_{i=1}^{k}\frac{1}{i}\leq 1+\sum_{i=2}^{k}\int_{i-1}^{i}\frac{1}{x}\,\mathrm{d}x=1+\int_{1}^{k}\frac{1}{x}\,\mathrm{d}x=1+\log k. \qquad (69)$$

to get the result (58). ♯

We also remark that we can apply FISAT to solve the minimization of sum of the smoothed function h^β and g, that is, $H^\beta = h^\beta + g$, however, this approach generates a sequence of iterates x_k such that $H(x_k)$ may not converge to the optimal objective value of (44). The proposed approach via using variable smoothing parameter β_k generates a sequence of iterates x_k which make $H(x_k)$ convergent.

In order to apply the developed method to solve sparse recovery problems mentioned above, we note that the function h in model (44) is set to be the objection function in sparse recovery models and g should be the indictor function of the constraint set. One of the most important components of our method is of computing the proximity operators of the function h and g. It is easy to obtain the proximity operator of the objective function h, however, computing the proximity operator (or the projection onto the constraint set) of the indicator function g of the constraint set is somewhat complex. Taking model (8) for matrix completion as an example. When $h = \|\cdot\|_*$, its proximity operator has been given in the above. We now derive the proximity operator of $g = \delta_C$ via a useful lemma which itself is of interest.

Lemma 3.4. Let $\mathcal{A}$ be a linear and orthogonal projection operator in $\mathcal{H}$ endowed with the inner product $\langle\cdot,\cdot\rangle$ and corresponding norm $\|\cdot\|$, and d be a vector in $\mathcal{H}$. The nonempty convex set C is $\{y \in \mathcal{H} : \|\mathcal{A}y - d\| \le \varepsilon\}$ where ε is a scalar. The projection operator $\mathcal{P}_C$ can be obtained:

$$\mathcal{P}_C(x) = \begin{cases} x, & \|\mathcal{A}x - d\| \le \varepsilon, \\ x - \frac{\lambda^*}{1+\lambda^*}\mathcal{A}(x-d), & \|\mathcal{A}x - d\| > \varepsilon. \end{cases} \tag{70}$$

where a scalar

$$\lambda^* = \frac{\|\mathcal{A}(x-d)\|}{\sqrt{\|\mathcal{A}d\|^2 + \varepsilon^2 - \|d\|^2}} - 1.$$

Proof: Given a vector x, $\mathcal{P}_C(x)$ is the solution to the following optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y - x\|^2 \\ \text{subject to} & \|\mathcal{A}y - d\| \le \varepsilon. \end{array} \tag{71}$$

Lagrangian function of problem (71) is

$$\mathcal{L}(y,\lambda) = \frac{1}{2}\|y - x\|^2 + \frac{\lambda}{2}(\|\mathcal{A}y - d\|^2 - \varepsilon^2). \tag{72}$$

A pair (y^*, λ^*) is the optimal primal-dual solution to (72) if and only if it satisfies the Karush-Kuhn-Tucker (KKT) conditions which read

$$\|\mathcal{A}y^* - d\|^2 - \varepsilon^2 \le 0, \tag{73}$$

$$\lambda^* \ge 0, \tag{74}$$

$$\lambda^*(\|\mathcal{A}y^* - d\|^2 - \varepsilon^2) = 0, \tag{75}$$

$$\mathcal{L}_y(y^*, \lambda^*) = y^* - x + \lambda^*\mathcal{A}(y^* - d) = 0. \tag{76}$$

From the equality (76) and the properties of the linear and orthogonal projection operator $\mathcal{A}$: $\mathcal{A}^2 = \mathcal{A}$, $\mathcal{A}^* = \mathcal{A}$, we derive that

$$y^* = \Big(I - \frac{\lambda^*}{\lambda^*+1}\mathcal{A}\Big)(\lambda^*\mathcal{A}d + x) = x - \frac{\lambda^*}{\lambda^*+1}\mathcal{A}(x - d). \tag{77}$$

The equality (75) indicates $\lambda^* = 0$ or $\|\mathcal{A}y^* - d\|^2 - \varepsilon^2 = 0$. Indeed, $\lambda^* = 0$ leads to $y^* = x$. Hence, a pari $(0, x)$ is the KKT point if and only if $x \in C$.

When $x \notin C$, from the equality

$$\|\mathcal{A}y^* - d\|^2 - \varepsilon^2 = 0,$$

the value of λ^* satisfies

$$\|\mathcal{A}x - d - \frac{\lambda^*}{\lambda^*+1}\mathcal{A}(x-d)\|^2 = \varepsilon^2, \tag{78}$$

which implies

$$\lambda^* = \frac{\|\mathcal{A}(x-d)\|}{\sqrt{\|\mathcal{A}d\|^2 + \varepsilon^2 - \|d\|^2}} - 1. \tag{79}$$

Now we need to demonstrate λ^* is a positive value. Since the set C is nonempty convex set, we find $y_0 \in C$, i.e., $\|\mathcal{A}y_0 - d\|^2 < \varepsilon^2$. Further,

$$\|\mathcal{A}(y_0 - d) - (d - \mathcal{A}d)\|^2 = \|\mathcal{A}(y_0 - d)\|^2 + \|d - \mathcal{A}d\|^2 < \varepsilon^2,$$

which indicates

$$\|d - \mathcal{A}d\|^2 = \|d\|^2 - \|\mathcal{A}d\|^2 < \varepsilon^2,$$

that is,

$$\|\mathcal{A}d\|^2 + \varepsilon^2 - \|d\|^2 > 0.$$

When $x \notin C$, we have that

$$\|\mathcal{A}(x - d)\|^2 - (\|\mathcal{A}d\|^2 - \|d\|^2 + \varepsilon^2) = \|\mathcal{A}x - d\|^2 - \varepsilon^2 > 0.$$

This derives

$$\lambda^* = \frac{\|\mathcal{A}(x - d)\|}{\sqrt{\|\mathcal{A}d\|^2 + \varepsilon^2 - \|d\|^2}} - 1 > 0. \tag{80}$$

This completes the proof. ♯

By setting $\mathcal{A} = \mathcal{P}_\Omega$ and letting $d = \mathcal{P}_\Omega(Y)$, we obtain from Lemma 3.4 that $\text{prox}_{\beta_k g}(X) = \mathcal{P}_C(X) = X - \text{prox}_{\varepsilon\|\cdot\|_F}(\mathcal{P}_\Omega(X - Y))$ in our method applied to matrix completion problem. Other projection operators involved in sparse recovery models are derived in the Appendix A

4. Numerical results

This section is devoted to show the numerical performance of the two proposed smoothing methods (SM1 and SM2 for short) applied to the above sparse recovery problems, that is, compressed sensing, matrix completion, and robust and stable principal component analysis. We also compare the two methods with the popular TFOCS, NESTA and FISTA in terms of accuracy and speed for sparse recovery. All the experiments are run on Thinkpad SL510 with Inter(R) Core(TM)2 Duo CPU @2.20 1.99GB RAM working on Microsoft Windows XP 2002.

4.1. Compressed Sensing

We first consider the problem of recovering sparse signals in the noiseless or noisy case. Since the projection onto the constraint set in model (DS) has no a closed-form solution, for the fair comparison, NESTA and SM2 are not applied to model (DS). Some details on this experiment are given as follows. The sensing matrix A is generated by picking randomly m rows from the $n \times n$ DCT matrix. A k-sparse signal x_0 of length n is generated by choosing randomly k nonzero coefficient positions, and sampling independently nonzero values from the standard Gaussian distribution. We denote by (n, m, k) type of signals of different length n, measurement number m, sparsity level k. Specifically, the following setting is considered: $n = 512$, 1024, $m = n/2$, $k = m/4$, i.e., $(512, 256, 64)$, $(1024, 512, 128)$. For each (n, m, k), we create 5 signals via the above way. The noiseless and noisy measurements are obtained respectively as: $b = Ax_0$ and $b = Ax_0 + e$, where

e is a Gaussian additive noise $e \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$ with noise level $\sigma = 0.05$. We also specify the parameter values involved in sparse recovery models and all the methods. According to [30], we take $\gamma = \sigma\sqrt{2\log n}$ in model (DS) and $\varepsilon = \sigma\sqrt{m + 2\sqrt{m\log m}}$ in model (BP$_\varepsilon$). The smoothing parameter β is set to 0.002 for SM1, TFOCS, NESTA. We let the parameter λ in model (QP$_\lambda$) be 0.002 in the noiseless case because of the equivalence of model (40) and (QP$_\lambda$). We do not know how to determine the parameter λ in model (QP$_\lambda$) in the noisy case, however, we try to select the most proper value to obtain the best performance for FISTA. We run 1000 iterations for all the methods and stop them when the relative ℓ_2-error between the original signal and recovered signal satisfies the inequality $||x^k - x_0||_2/||x_0||_2 <$ Tol. We take Tol as 10^{-10} in the noiseless case and as 10^{-04} in the noisy case. x^{opt} is the output of all the methods after 1000 iterations.

Figure 1 and Figure 2 exhibit that a sparse signal with $(512, 256, 64)$ is recovered via SM1 and SM2 in the noiseless and noisy cases, respectively. It can be seen that sparse signals recovery in the noisy case is a more difficult task than in the noiseless case.

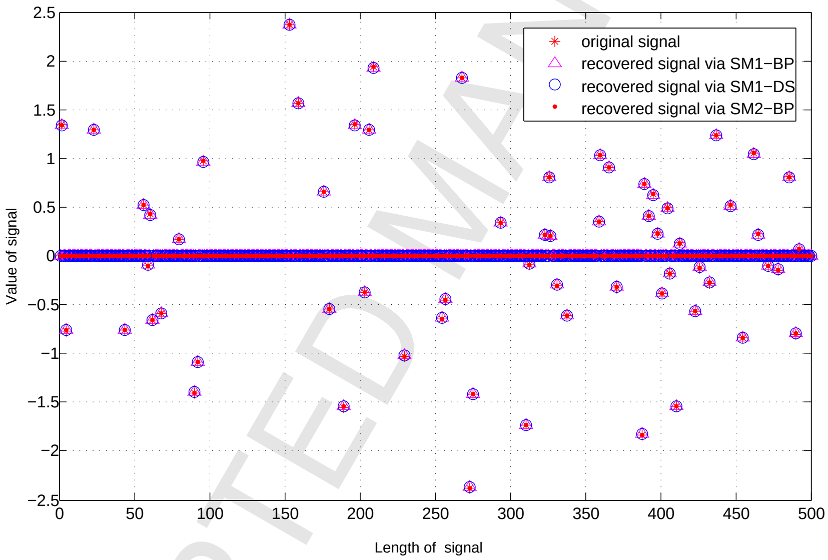

Figure 1: SM1 and SM2 applied to model (BP) or (DS) for recovering a sparse signal with $(512, 256, 64)$ in the noiseless case.

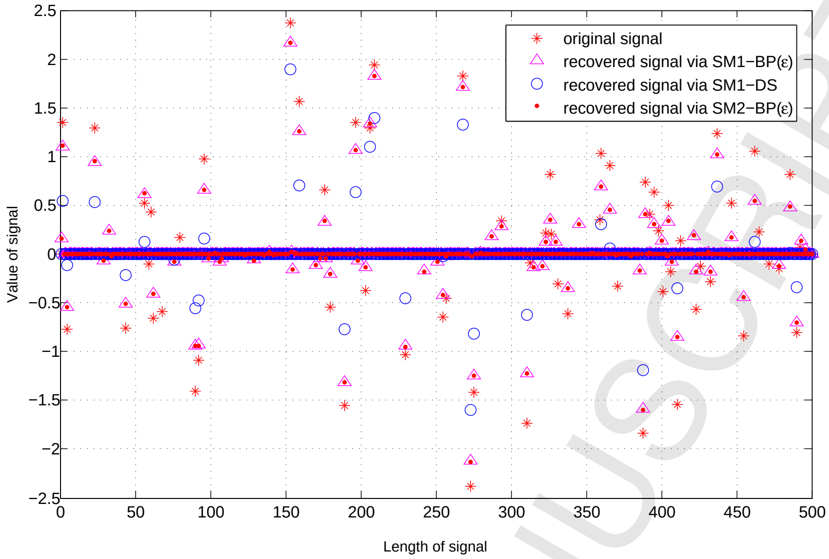

Figure 2: SM1 and SM2 applied to model BP$_\varepsilon$ or (DS) for recovering a sparse signal with $(512, 256, 64)$ in the noisy cases.

Comparison of function value errors $\|x^k\|_1 - \|x_0\|_1$ generated by different methods for recovering this sparse signal in the noisy case is shown in Figure 3. Clearly, the iterative sequences of values of the objective function generated by FISTA, SM1-BP$_\varepsilon$ and NESTA are convergent after 100 iterations. However, SM1-BP$_\varepsilon$ and NESTA are more accurate than FISTA. SM2-BP$_\varepsilon$ performs much better than all other algorithms with respect to accuracy and speed. Table 1 reports the average relative ℓ_1-error, relative ℓ_2-error, absolute ℓ_∞-error and the average elapsed time of all the methods for recovering sparse signals with $(512, 256, 64)$ or $(1024, 512, 128)$ in noiseless case. Table 1 has demonstrated that TFOCS-BP and TFOCS-DS are two best methods for recovering sparse signals with respect to accuracy, and the performance of SM1-BP (or FISTA-QP$_\lambda$), SM1-DS, especially for SM2-BP is comparable to or more better to that of NESTA. Table 2 shows the average performance of all the methods for recovering sparse with signals $(512, 256, 64)$ or $(1024, 512, 128)$ in the noisy case. It can be seen that TFOCS applied to model BP$_\varepsilon$ and (DS) shows poor performance in term of accuracy and speed. Moreover, all the methods seem to share similar performance after 1000 iterations. However, the methods applied to model BP$_\varepsilon$, except for NESTA, need less time than the methods applied to model (DS). The main reason is that solving model (DS) needs many matrix-vector products in each iteration. This observation also suggests that model (DS) has the more attractive theoretical result, however, model BP$_\varepsilon$ is capable of faster performing sparse recovery from the noisy data.

4.2. Matrix Completion

As the second example of sparse recovery, low-rank matrix completion problem is considered in this subsection. First, in the random low-rank matrix completion problem, $n \times n$ square matrices M of rank r are generated by sampling two $n \times r$ factors M_1 and

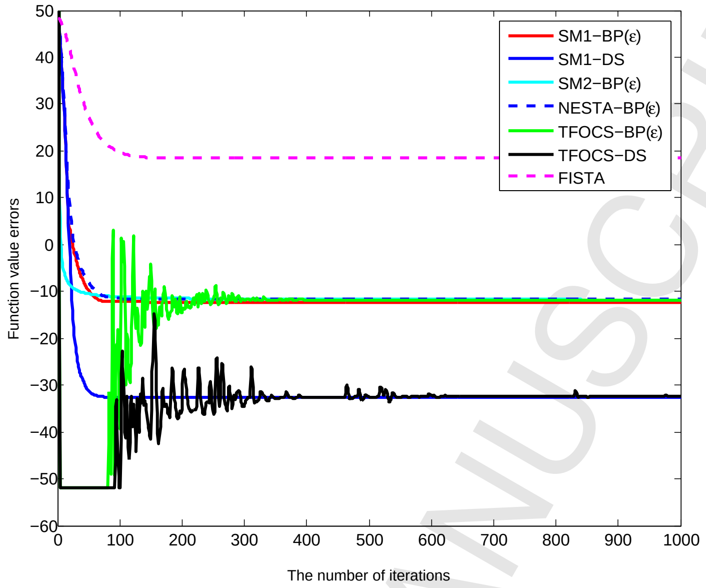

Figure 3: Comparison of the function value error $\|x^k\|_1 - \|x_0\|_1$.

$(n, m, k) = (512, 256, 64)$				
Method	$\|x^{\text{opt}} - x_0\|_1 / \|x_0\|_1$	$\|x^{\text{opt}} - x_0\|_2 / \|x_0\|_2$	$\|x^{\text{opt}} - x_0\|_\infty$	time (s)
SM1-BP	3.30e-03	2.80e-03	6.40e-03	0.4615
SM1-DS	1.33e-02	1.12e-02	2.57e-02	0.8422
SM2-BP	5.30e-03	2.30e-03	5.40e-03	0.7953
NESTA-BP	1.07e-02	4.70e-03	1.09e-02	1.4558
TFOCS-BP	7.91e-11	7.82e-11	2.38e-10	0.4280
TFOCS-DS	1.18e-08	1.19e-08	3.58e-08	0.9007
$(n, m, k) = (1024, 512, 128)$				
SM1-BP	3.20e-03	2.70e-03	4.90e-03	0.8809
SM1-DS	1.24e-02	1.05e-02	1.93e-02	1.6464
SM2-BP	5.20e-03	2.20e-03	3.90e-03	1.2851
NESTA-BP	1.02e-02	4.30e-03	7.70e-03	2.3109
TFOCS-BP	7.76e-04	1.40e-03	1.12e-02	0.9307
TFOCS-DS	2.14e-02	2.18e-02	5.82e-02	1.5450

Table 1: The average performance of all the methods for recovering signals with $(512, 256, 64)$ or $(1024, 512, 128)$ in terms of accuracy and elapsed time in the noiseless case.

M_2 independently, each having i.i.d. Gaussian entries, and then setting $M = M_1 M_2^*$. The set of observed entries Ω is sampled uniformly at random among all sets of cardinality m. We denote by m/n^2 the percentage of observed entries. Further, as suggested in [17], we denote by m/d_r the information oversampling ratio, where $d_r = r(2n - r)$ is the 'true dimensionality' of a $n \times n$ matrix of rank r. We would like to stress that because of the limitation of the computation in our experiment, we only consider two types of small-scale matrices: $n = 200$ and 300. The noiseless and noisy sampled entries are given respectively as: $Y = \mathcal{P}_\Omega M$ and $Y = \mathcal{P}_\Omega M + \mathcal{P}_\Omega E$, where E is a zero-mean Gaussian white noise with standard deviation $\sigma = 0.05$. The scale ε in model (8) is set to $\sigma\sqrt{m + 2\sqrt{m \log m}}$ and the smoothing parameter is the same as in the compressed sensing. In addition, we only use the simple Matlab command `[u,d,v]=svd(X,'econ')` to obtain the SVD of a matrix X.

Second, we also consider an image in-painting problem. We take a standard 256×256 gray-scale image Peppers and then use the SVD to obtain its low-rank approximation M with rank $r = 25$. The original Peppers image, its low-rank image and sampled low-rank image with $m/n^2 = 0.5$ are displayed in Figure 4. 1000 iterations are run for all the methods and we stop them when the inequality $||M^k - M||_F/||M||_F <$ Tol is satisfied. Tol is taken as 10^{-10} in the noiseless case and as 10^{-04} in the noisy case. All the computational results are averaged over five runs.

Figure 4 also displays the recovered low-rank Peppers image via SM2 in the noiseless case. Our computational results are reported in Table 3 and 4. These results show that it is in the noisy case that our methods proposed is comparable to TFOCS and better to NESTA. However, TFOCS has the best performance in the noiseless case.

$(n, m, k) = (512, 256, 64)$				
Method	$\|x^{\text{opt}} - x_0\|_1/\|x_0\|_1$	$\|x^{\text{opt}} - x_0\|_2/\|x_0\|_2$	$\|x^{\text{opt}} - x_0\|_\infty$	time (s)
SM1-BP$_\varepsilon$	0.2351	0.1840	0.3537	0.6352
SM1-DS	0.1263	0.4571	0.3729	0.8472
SM2-BP$_\varepsilon$	0.2361	0.2637	0.3261	0.8251
NESTA-BP$_\varepsilon$	0.2618	0.2730	0.2637	1.5273
TFOCS-BP$_\varepsilon$	0.2471	0.4829	0.3827	0.5421
TFOCS-DS	0.5172	0.3711	0.3823	0.9365
FISTA	0.3718	0.2948	0.2733	0.7235
$(n, m, k) = (1024, 512, 128)$				
SM1-BP$_\varepsilon$	0.2618	0.2371	0.4272	1.1736
SM1-DS	0.7183	0.3728	0.2638	1.6255
SM2-BP$_\varepsilon$	0.2638	0.3728	0.2638	1.2736
NESTA-BP$_\varepsilon$	0.7351	0.5163	0.3729	2.2814
TFOCS-BP$_\varepsilon$	0.3671	0.4719	0.3724	0.9361
TFOCS-DS	0.1357	0.9123	0.6281	1.6253
FISTA	0.4821	0.2637	0.1126	0.8253

Table 2: The average performance of all the methods for recovering signals with $(512, 256, 64)$ or $(1024, 512, 128)$ in terms of accuracy and elapsed time in the noisy case with $\sigma = 0.05$.

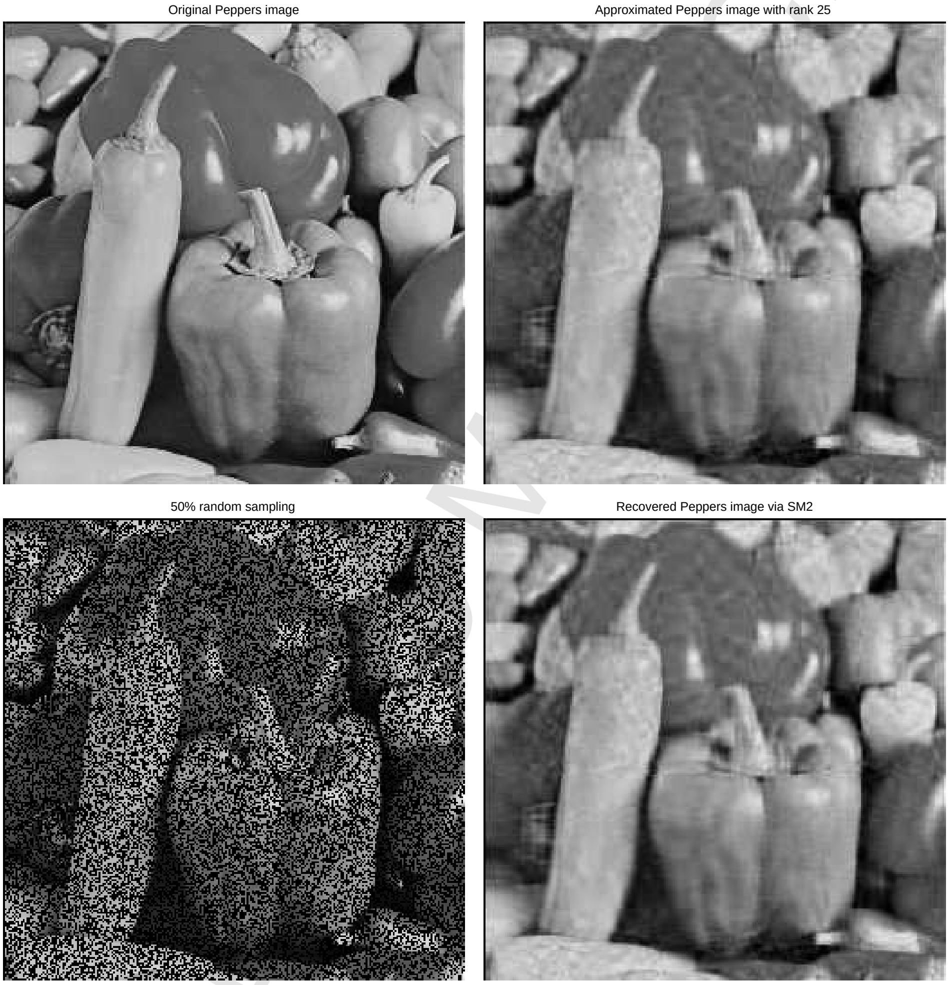
Original Peppers image
Approximated Peppers image with rank 25
50% random sampling
Recovered Peppers image via SM2

Figure 4: The original 256×256 Peppers, its low-rank approximation with rank $r = 25$, sampled low-rank image with $m/n^2 = 0.5$ and recovered image via SM2 in the noiseless case.

$(n, m/n^2, m/d_r) = (200, 0.5, 10.1)$			$(n, m/n^2, m/d_r) = (300, 0.5, 7.6)$	
Method	$\\|M^{\text{opt}} - M\\|_F / \\|M\\|_F$	time (s)	$\\|M^{\text{opt}} - M\\|_F / \\|M\\|_F$	time (s)
		$\sigma = 0$		
SM1	2.35e-02	89.3054	4.61e-02	265.1682
SM2	1.36e-03	86.3207	1.05e-02	261.4720
NESTA	3.71e-02	95.0543	2.73e-02	283.0913
TFOCS	3.58e-11	56.6236	8.16e-11	127.2538
		$\sigma = 0.05$		
SM1	2.45e-02	72.8581	2.77e-02	224.5613
SM2	1.58e-02	82.7290	3.58e-02	203.4517
NESTA	2.04e-02	87.8953	2.44e-02	262.7180
TFOCS	6.71e-02	104.2223	6.11e-02	282.5627
FISTA	3.01e-02	88.9652	5.03e-02	265.0152

Table 3: The average performance of all the methods for the random low-rank matrix completion problem in the noiseless and noisy cases.

Method	$\\|M^{\text{opt}} - M\\|_F / \\|M\\|_F$	time (s)
	$\sigma = 0$	
SM1	2.47e-02	156.2873
SM2	1.62e-03	152.8816
NESTA	2.38e-02	178.3619
TFOCS	2.85e-09	136.2722
	$\sigma = 0.05$	
SM1	2.81e-02	126.8166
SM2	3.66e-02	124.7718
NESTA	5.23e-02	152.1147
TFOCS	2.78e-02	147.2538
FISTA	3.68e-02	155.9152

Table 4: The average performance of all the methods for recovering low-rank Peppers image in the noiseless and noisy cases.

4.3. Robust and Stable Principal Component Analysis

In this subsection, we test all the methods for recovering the unknown low-rank matrix and sparse error matrix from the noiseless or noisy data matrix. We first generate $n \times n$ matrices L_0 of rank r as in the matrix completion and $n \times n$ K-sparse error matrices S_0 by the way which we generate k-sparse signals in the compressed sensing. The noiseless data matrix can be obtained via $D = L_0 + S_0$. A zero-mean Gaussian white noise E with standard deviation $\sigma = 0.05$ is added to D to get the noisy data matrix. The model parameter τ is set to 0.01. As in the matrix completion, we only consider $n = 200$ and $n = 300$. We run 1000 iterations for all the methods and terminate them when the relative error satisfies:

$$\text{RE} = \frac{\sqrt{\|L^k - L_0\|_F^2 + \|S^k - S_0\|_F^2}}{\sqrt{\|L_0\|_F^2 + \|S_0\|_F^2}} \leq \text{Tol}.$$

We also take Tol as 10^{-10} in the noiseless case and as 10^{-04} in the noisy case.

The averaged performance in terms of accuracy and elapsed time is given in table 5 and 6 which report the similar conclusions as in the matrix completion and compressed sensing.

	$(n, r, K/n^2) = (200, 5, 0.05)$		$(n, r, K/n^2) = (300, 10, 0.05)$	
Method	RE	time (s)	RE	time (s)
SM1	4.51e-02	85.7215	5.61e-02	257.2635
SM2	4.72e-03	84.1263	8.26e-03	256.3528
NESTA	2.37e-02	89.2638	5.22e-02	274.6216
TFOCS	4.71e-06	135.7238	8.16e-06	382.5627

Table 5: The average performance of all the methods for robust principal component analysis.

	$(n, r, K/n^2) = (200, 5, 0.05)$		$(n, r, K/n^2) = (300, 10, 0.05)$	
Method	RE	time (s)	RE	time (s)
SM1	2.47e-02	134.5281	4.65e-02	232.2580
SM2	2.37e-02	73.2638	2.11e-02	204.5518
NESTA	1.36e-02	75.2531	2.58e-02	246.7728
TFOCS	3.53e-02	135.2718	7.63e-02	368.2663
FISTA	2.48e-02	82.6342	1.18e-02	256.3728

Table 6: The average performance of all the methods for robust and stable principal component analysis.

5. Conclusions

In this paper, we cast sparse recovery problems of interest as two general nonsmooth convex optimization problems by introducing the indicator function related to each constrained set in sparse recovery models. Two new smoothing strategies based on the

Moreau Envelope were introduced, that is, smoothing the related indicator function or smoothing the objection function via its Moreau envelope. Two efficient first-order smoothing methods were also developed. The first method is FISTA which solves the first smoothed model, and the another method solving the second smoothed model modifies FISTA to establish the rate of convergence of order $\mathcal{O}(\frac{\log k}{k})$ for the iterative sequence of values of the original objective function. The two proposed smoothing methods are comparable to the state-of-the-art first-order methods in terms of accuracy and speed when applied to compressed sensing, matrix completion, and robust and stable principal component analysis.

Acknowledgements

The authors would like to thank the two anonymous reviewers and the associate editor and editor-in chief for their constructive comments, which significantly improved this paper. This research was supported by the National Natural Science Foundation of China under the grants 11131006, 41390450, 91330204, and 11271297, and in part supported by the National Basic Research Program of China under the grant 2013CB329404.

The Appendix A

Here, we provide other projections onto the corresponding constraint sets involved in sparse recovery models.

1. We first evaluate the projection $\mathcal{P}_C(x)$ which is the solution to the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y-x\|_2^2 \\ \text{subject to} & \|Ay-b\|_2 \le \varepsilon. \end{array} \tag{81}$$

Note that in many practical applications, it is common to take the sensing matrix A as a submatrix of a unitary transformation such as FFT, DCT, or the Hadamard transform, which admits a fast algorithm for matrix-vector products. Since the sensing matrix A satisfies $AA^* = I$, it is easy to evaluate the projection $\mathcal{P}_C(x)$. We now prove that the projection $\mathcal{P}_C$ is

$$\mathcal{P}_C(x) = x - A^*\text{prox}_{\varepsilon\|\cdot\|_2}(Ax-b).$$

Proof: The Lagrangian function of (81) is

$$\mathcal{L}(y,\lambda) = \frac{1}{2}\|y-x\|_2^2 + \frac{\lambda}{2}(\|Ay-b\|_2^2 - \varepsilon^2). \tag{82}$$

A pair (y^*,λ^*) is the optimal primal-dual solution to (82) if and only if it satisfies the Karush-Kuhn-Tucker (KKT) conditions which read

$$\|Ay^*-b\|_2^2 - \varepsilon^2 \le 0, \tag{83}$$

$$\lambda^* \ge 0, \tag{84}$$

$$\lambda^*(\|Ay^*-b\|_2^2 - \varepsilon^2) = 0, \tag{85}$$

$$\mathcal{L}_y(y^*,\lambda^*) = y^* - x + \lambda^* A^*(Ay^*-b) = 0. \tag{86}$$

It can induce from the equality (86) and $AA^* = I$ that

$$y^* = \Big(I - \frac{\lambda^*}{\lambda^*+1}A^*A\Big)(\lambda^* A^* b + x) = x - \frac{\lambda^*}{\lambda^*+1}A^*(Ax-b). \tag{87}$$

The equality (85) indicates $\lambda^* = 0$ or $\|Ay^* - b\|_2^2 - \varepsilon^2 = 0$. Indeed, $\lambda^* = 0$ leads to $y^* = x$. Hence, a pari $(0, x)$ is the KKT point if and only if $x \in C$. When $x \notin C$, from the equality $\|Ay^* - b\|^2 - \varepsilon^2 = 0$, the value of λ^* is

$$\lambda^* = \frac{\|Ax-b\|_2}{\varepsilon} - 1. \tag{88}$$

Hence, $y^* = x - A^*\text{prox}_{\varepsilon\|\cdot\|_2}(Ax - b)$. ♯

2. We second evaluate the projection $\mathcal{P}_C(x)$ which is the solution to the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y-x\|_2^2 \\ \text{subject to} & \|D^{-1}A^*(Ay-b)\|_\infty \le \gamma. \end{array} \tag{89}$$

This projection can be computed via the following equality:

$$\mathcal{P}_C(x) = x + A^*AD^{-1}z,$$

where z is a minimizer of the following minimization problem

$$\min_w \frac{1}{2}\|AD^{-1}w - (b - Ax)\|^2 + \gamma\|w\|_1.$$

Proof: The projection problem can be describe as conic form as follows:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y-x\|_2^2 \\ \text{subject to} & \mathcal{A}y - d \in \mathcal{K}, \end{array} \tag{90}$$

where the operator $\mathcal{A} : y \to (D^{-1}A^*Ay, 0)$, $d = (D^{-1}A^*b, -\gamma)$, the primal conic $\mathcal{K} = \{(u,t) : \|u\|_\infty \le t\}$. Recall that the dual of our conic form is given by

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \in \mathcal{K}^*, \end{array} \tag{91}$$

where $g(\lambda)$ is the Lagrange dual function

$$g(\lambda) = \inf_y \mathcal{L}(y,\lambda) = \inf_y \frac{1}{2}\|y-x\|_2^2 - \langle \lambda, \mathcal{A}y - d\rangle, \tag{92}$$

and $\mathcal{K}^*$ is the dual cone defined as $\mathcal{K}^* = \{(w,v) : \|w\|_1 \le v\}$. Indeed, given λ, the Lagrange dual function $g(\lambda)$ is obtained at a point $y = x + \mathcal{A}^*\lambda$ for the Lagrange function $\mathcal{L}(y,\lambda)$, that is,

$$g(\lambda) = \mathcal{L}(x + \mathcal{A}^*\lambda, \lambda) = -\frac{1}{2}\|\mathcal{A}^*\lambda\|^2 - \langle \lambda, \mathcal{A}x - d\rangle. \tag{93}$$

Hence,

$$-\max_{\lambda\in\mathcal{K}^*} g(\lambda) = \min_{\lambda\in\mathcal{K}^*} \frac{1}{2}\|\mathcal{A}^*\lambda\|^2 + \langle\lambda, \mathcal{A}x - d\rangle. \tag{94}$$

According to the definition of the operator $\mathcal{A}$, we easily get its adjoint operator $\mathcal{A}^*$: $(w, v) \to A^*AD^{-1}w$. Further, by setting $\lambda = (w, v)$, we can have that

$$\begin{aligned} -\max_{\lambda\in\mathcal{K}^*} g(\lambda) &= \min_{\|w\|_1\leq v} \frac{1}{2}\|AD^{-1}w - (b - Ax)\|_2^2 + \langle v, \gamma\rangle \\ &= \min_{w} \frac{1}{2}\|AD^{-1}w - (b - Ax)\|_2^2 + \gamma\|w\|_1. \end{aligned} \tag{95}$$

Therefore, if we can obtain the minimizer denoted by z for problem (95), the projection $\mathcal{P}(x)$ can be computed via $\mathcal{P}(x) = x + A^*AD^{-1}z$. ♯

3. We also need to evaluate the projection $\mathcal{P}_C(X_1, X_2)$ which is the solution to the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\Big(\|L - X_1\|_F^2 + \|S - X_2\|_F^2\Big) \\ \text{subject to} & \|L + S - D\|_F \leq \varepsilon. \end{array} \tag{96}$$

This projection equals to

$$\mathcal{P}_C(X_1, X_2) = \Big(X_1 - \frac{1}{2}\text{prox}_{\varepsilon\|\cdot\|_F}(X_1 + X_2 - D), X_2 - \frac{1}{2}\text{prox}_{\varepsilon\|\cdot\|_F}(X_1 + X_2 - D)\Big).$$

Proof: We generate the product space $X \times X = \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}$ with the inner product $\langle\cdot,\cdot\rangle_{X\times X} = \langle\cdot,\cdot\rangle_X + \langle\cdot,\cdot\rangle_X$ and related norm $\|\cdot\|_{X\times X}$. We introduce an linear operator $A : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$, defined at a given point $Y = (L, S)$ by $A : Y \to L + S$. Thus, this projection problem can be expressed in terms of the notation $Y = (L, S)$ and $X = (X_1, X_2)$, as

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|Y - X\|^2 \\ \text{subject to} & \|A(Y) - D\|_F \leq \varepsilon. \end{array} \tag{97}$$

Similar proof is given as in the first case. In fact, the Lagrangian function of this problem (96) and KKT conditions can be written as in the first case. It follows from the definition of the operator A, that the adjoint operator A^* is defined as $A^* : Z \to (Z, Z)$. Clearly, $AA^* = 2I$. Hence, we can obtain

$$Y^* = \Big(I - \frac{\lambda^*}{2\lambda^* + 1}A^*A\Big)(\lambda^*A^*(D) + X) = X - \frac{\lambda^*}{2\lambda^* + 1}A^*(A(X) - D). \tag{98}$$

Likewise, if $X \in C$, Y^* equals to X. When $X \notin C$, from the equality

$$\|A(Y^*) - D\|^2 - \varepsilon^2 = 0,$$

the value of λ^* is

$$\lambda^* = \frac{1}{2}\Big(\frac{\|A(X) - D\|}{\varepsilon} - 1\Big). \tag{99}$$

Hence, $Y^* = X - \frac{1}{2}A^*\text{prox}_{\varepsilon\|\cdot\|_F}(A(X) - D)$, that is,

$$\mathcal{P}_C(X_1, X_2) = \Big(X_1 - \frac{1}{2}\text{prox}_{\varepsilon\|\cdot\|_F}(X_1 + X_2 - D), X_2 - \frac{1}{2}\text{prox}_{\varepsilon\|\cdot\|_F}(X_1 + X_2 - D)\Big). \tag{100}$$

This completes the proof. ♯

References

[1] E. J. Candès, M. B. Wakin, An introduction to compressive sampling, IEEE signal processing magazine 25 (2) (2008) 21–30.

[2] D. L. Donoho, Compressed sensing, IEEE Transactions on information theory 52 (4) (2006) 1289–1306.

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

[4] E. J. Candès, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, IEEE transactions on information theory 52 (12) (2006) 5406–5425.

[5] E. J. Candès, T. Tao, Decoding by linear programming, IEEE transactions on information theory 51 (12) (2005) 4203–4215.

[6] M. Lustig, D. L. Donoho, J. M. Pauly, Sparse mri: The application of compressed sensing for rapid mr imaging, Magnetic resonance in medicine 58 (6) (2007) 1182–1195.

[7] M. Elad, Sparse and redundant representations: From theory to applications in signal and image processing, Springer Science & Business Media, 2010.

[8] S. S. Chen, D. L. Donoho, M. A. Saunders, Atomic decomposition by basis pursuit, SIAM review 43 (1) (2001) 129–159.

[9] E. J. Candès, T. Tao, The dantzig selector: Statistical estimation when p is much larger than n, The Annals of Statistics (2007) 2313–2351.

[10] E. J. Candès, J. Romberg, l1-magic: Recovery of sparse signals via convex programming, URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf 4 (2005) 14.

[11] J. Yang, Y. Zhang, Alternating direction algorithms for ℓ_1-problems in compressive sensing, SIAM journal on scientific computing 33 (1) (2011) 250–278.

[12] W. Yin, S. Osher, D. Goldfarb, J. Darbon, Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing, Siam Journal on Imaging Sciences 1 (1) (2008) 143–168.

[13] S. Becker, J. Bobin, E. J. Candès, Nesta: A fast and accurate first-order method for sparse recovery, SIAM Journal on Imaging Sciences 4 (1) (2011) 1–39.

[14] E. J. Candès, B. Recht, Exact matrix completion via convex optimization, Communications of the ACM 55 (6) (2012) 111–119.

[15] E. J. Candès, Y. Plan, Matrix completion with noise, Proceedings of the IEEE 98 (6) (2010) 925–936.

[16] A. Singer, A remark on global positioning from local distances, Proceedings of the National Academy of Sciences 105 (28) (2008) 9507–9511.

[17] J.-F. Cai, E. J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM Journal on Optimization 20 (4) (2010) 1956–1982.

[18] E. J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM (JACM) 58 (3) (2011) 11.

[19] Z. Zhou, X. Li, J. Wright, E. J. Candès, Y. Ma, Stable principal component pursuit, in: Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on, IEEE, 2010, pp. 1518–1522.

[20] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, arXiv preprint arXiv:1009.5055.

[21] J. Yang, X. Yuan, Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization, Mathematics of computation 82 (281) (2013) 301–329.

[22] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Modeling & Simulation 4 (4) (2005) 1168–1200.

[23] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM journal on imaging sciences 2 (1) (2009) 183–202.

[24] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, Y. Ma, Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix, Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) 61 (6).

[25] K.-C. Toh, S. Yun, An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems, Pacific Journal of optimization 6 (615-640) (2010) 15.

[26] Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical programming 103 (1) (2005) 127–152.

[27] S. R. Becker, E. J. Candès, M. C. Grant, Templates for convex cone problems with applications to sparse signal recovery, Mathematical programming computation 3 (3) (2011) 165.

[28] Y. Nesterov, Introductory lectures on convex optimization: A basic course, Vol. 87, Springer Science & Business Media, 2013.

[29] H. H. Bauschke, P. L. Combettes, Convex analysis and monotone operator theory in Hilbert spaces, Springer Science & Business Media, 2011.

[30] T. T. Cai, G. Xu, J. Zhang, On recovery of sparse signals via ℓ_1-minimization, IEEE Transactions on Information Theory 55 (7) (2009) 3388–3397.