Penalty methods for a class of non-Lipschitz optimization problems

Xiaojun Chen* Zhaosong Lu† Ting Kei Pong‡

June 26, 2015 (Revised: April 3, 2016)

Abstract

We consider a class of constrained optimization problems with a possibly nonconvex non-Lipschitz objective and a convex feasible set being the intersection of a polyhedron and a possibly degenerate ellipsoid. Such problems have a wide range of applications in data science, where the objective is used for inducing sparsity in the solutions while the constraint set models the noise tolerance and incorporates other prior information for data fitting. To solve this class of constrained optimization problems, a common approach is the penalty method. However, there is little theory on exact penalization for problems with nonconvex and non-Lipschitz objective functions. In this paper, we study the existence of exact penalty parameters regarding local minimizers, stationary points and ϵ-minimizers under suitable assumptions. Moreover, we discuss a penalty method whose subproblems are solved via a nonmonotone proximal gradient method with a suitable update scheme for the penalty parameters, and prove the convergence of the algorithm to a KKT point of the constrained problem. Preliminary numerical results demonstrate the efficiency of the penalty method for finding sparse solutions of underdetermined linear systems.

Keywords: Exact penalty, proximal gradient method, sparse solution, nonconvex optimization, non-Lipschitz optimization.
MSC2010 Classification: 90C30, 90C26.

1 Introduction

We consider the following constrained optimization problem:

$$\begin{aligned} \min_x \quad & \Phi(x) \\ \text{s.t.} \quad & x \in S := S_1 \cap S_2, \end{aligned} \tag{1.1}$$

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. E-mail: maxjchen@polyu.edu.hk. This author's work is supported partly by Hong Kong Research Grant Council grant PolyU5001/12p.

†Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada. Email: zhaosong@sfu.ca. This author's work is supported in part by NSERC Discovery Grant.

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. E-mail: tk.pong@polyu.edu.hk. This author's work is supported partly by Hong Kong Research Grants Council PolyU253008/15p.

where $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a nonnegative continuous function, $S_1 \subseteq \mathbb{R}^n$ is a simple polyhedron, and

$$S_2 = \{x : \ \|Ax - b\| \leq \sigma, \quad Bx \leq h\}.$$

Here $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, \sigma \geq 0, B \in \mathbb{R}^{\ell\times n}$ and $h \in \mathbb{R}^\ell$ are given matrices and vectors. We emphasize that Φ is neither necessarily convex nor locally Lipschitz continuous. To avoid triviality, we suppose that the feasible region S is nonempty.

Problem (1.1) is flexible enough to accommodate a wide range of optimization models with important applications in imaging sciences, signal processing, and statistical variable selections, etc. For example, with $S_1 = \mathbb{R}^n$ and B being vacuous, i.e., $S = S_2 = \{x : \|Ax - b\| \leq \sigma\}$, problem (1.1) reduces to the following problem

$$\begin{aligned} \min_x \quad & \Phi(x) \\ \text{s.t.} \quad & \|Ax - b\| \leq \sigma. \end{aligned} \tag{1.2}$$

This problem with $\Phi(x) = \|x\|_1$ has been studied extensively for recovering sparse signals from the possibly noisy measurements b; here, the parameter σ allows the user to *explicitly* specify the tolerance for the noise level. We refer the readers to the comprehensive review [3] for more details. In addition, we emphasize that the objective function Φ in our model (1.1) is allowed to be nonsmooth and possibly nonconvex non-Lipschitz. This enables the choice of various objective functions for inducing desirable structures in the optimal solutions. For instance, when sparsity is of concern, one popular choice of Φ is $\Phi(x) = \sum_{i=1}^n \phi(x_i)$, with ϕ being one of the widely used penalty functions, such as the bridge penalty [16, 17], the fraction penalty [13] and the logistic penalty [23]. On the other hand, we note that the simple polyhedron S_1 can be used for incorporating hard constraints/prior information that must be satisfied by the decision variables in applications. For instance, if a true solution to (1.2) is known to be in a certain interval $[l, u]$ for some $l < u$, l and $u \in \mathbb{R}^n$, then the S_1 can be chosen to be $[l, u]$ instead of just $\mathbb{R}^n$. Constraints of this kind arise naturally in applications such as image restoration, where all gray level images have intensity values ranging from 0 to 1. As shown in [1,4,24], incorporating the bound constraints can lead to substantial improvements in the quality of the restored image.

While (1.1) is a very flexible model covering a wide range of applications, this optimization problem is a constrained optimization problem, which is typically hard to solve. In the case when Φ is convex, $S_1 = \mathbb{R}^n$ and B is vacuous, i.e., (1.2), it is well known that the problem is equivalent to solving

$$\min_x \ H_\lambda(x) := \lambda\|Ax - b\|^2 + \Phi(x) \tag{1.3}$$

for some regularization parameter $\lambda > 0$, under some mild assumptions; see, for example, [11]. Unlike (1.2), for many choices of Φ, the regularized formulation (1.3) can be solved readily by various first-order methods such as the NPG method in [28]. This regularized formulation has been extensively studied in *both* cases where Φ is convex or nonconvex in the last few decades; see, for example, [3, 5–7, 9, 10, 12–14, 16, 17, 21, 23, 27, 29, 30]. Nevertheless, the equivalence between (1.2) and (1.3) does not hold in the nonconvex scenario: indeed, for nonconvex Φ and certain data (A, b, σ), there does not exist a λ so that problems (1.2) and (1.3) have a common global or local minimizer; see our Example 3.1. In particular, one cannot solve (1.2) via solving the unconstrained problem (1.3) for a suitable λ in general.

In a hope of constructing a simpler optimization problem whose local/global minimizers are closely related to (1.1), we resort to the penalty approach. While this is a standard approach, there are two important new ingredients in our work. First, although exact penalization for constrained optimization problems with a Lipschitz objective has been well studied (see, for example, [25]), to the best of our knowledge, there is little theory and development for problems with nonconvex non-Lipschitz objectives such as problem (1.1) with ϕ being the bridge penalty. Second, we consider partial penalization that keeps the constraints S_1 in (1.1). Recall that the set S_1 in (1.1) can be used to model *hard* constraints that must be satisfied or simple constraints that can be easily satisfied[1], while the set S_2 can be used to model *soft* constraints that only need to be approximately satisfied. Consequently, it can be advantageous to be able to penalize only the constraints corresponding to S_2 and *keep* the hard constraints S_1.

The penalty problem we consider is

$$\min_{x \in S_1} \; F_\lambda(x) := \lambda[(\|Ax-b\|^2 - \sigma^2)_+ + \|(Bx-h)_+\|_1] + \Phi(x) \tag{1.4}$$

for some $\lambda > 0$, where a_+ denotes the vector whose ith entry is $\max\{a_i, 0\}$ for any $a \in \Re^n$. In this paper, we derive various (partial) exact penalization results regarding (1.1) and (1.4). Specifically, under some suitable assumptions, we establish that:

(i) any local minimizer of problem (1.1) is also that of problem (1.4), provided that $\lambda \geq \lambda^*$ for some $\lambda^* > 0$;

(ii) any global minimizer of problem (1.1) is an ϵ-global minimizer of problem (1.4), provided that $\lambda \geq \lambda^*$ for some $\lambda^* > 0$;

(iii) the projection of any global minimizer of problem (1.4) onto the feasible set S of problem (1.1) produces an ϵ-global minimizer of problem (1.1), provided that $\lambda \geq \lambda^*$ for some $\lambda^* > 0$.

Consequently, problem (1.4) is an exact penalty formulation for (1.1), and an approximate solution of problem (1.1) can be obtained by solving (1.4) with $\lambda = \lambda^*$ if an exact penalty parameter λ^* is known.

In practice, the value of such λ^* is, however, generally unknown. Owing to this, we further propose a penalty method for solving (1.1) whose subproblems are (partially) smoothed and then solved approximately via a nonmonotone proximal gradient (NPG) method [28] with a suitable update scheme for the penalty and smoothing parameters. It is noteworthy that the NPG method originally studied in [28] was proposed for minimizing the sum of a possibly nonsmooth function and a smooth function whose gradient is *globally* Lipschitz continuous. Nevertheless, the gradient of the smooth component associated with our subproblems is *locally* but not globally Lipschitz continuous. We are fortunately able to show that this NPG method is indeed capable of solving a more general class of problems which includes our subproblems as a special case. In addition, we show that any accumulation point of the sequence generated by our penalty method is a KKT point of (1.1) under suitable assumptions. Finally, to benchmark our approach, we consider a sparse recovery problem and compare (1.2) with $\Phi(x) = \sum_{i=1}^n |x_i|^{\frac{1}{2}}$ solved by our penalty method against two other approaches: solving (1.2) with $\Phi(x) = \|x\|_1$ by the SPGL1 [2]

[1]This means that the projection onto S_1 is easy to compute.

for finding sparse solutions, and solving (1.3) with $\Phi(x) = \sum_{i=1}^n |x_i|^{\frac{1}{2}}$ for a suitably chosen λ. Our numerical results demonstrate that the solutions produced by our approach are sparser and have smaller recovery errors than those found by the other approaches.

The rest of the paper is organized as follows. We present notation and preliminary materials in Section 2. In Section 3, we study the existence of exact penalty parameters regarding local minimizers and ϵ-minimizers. In Section 4, we discuss the first-order optimality conditions for problems (1.1) and (1.4). We then propose a penalty method for solving problem (1.1) with an update scheme for the penalty parameters and establish its convergence to KKT points of (1.1). In Section 5, we conduct numerical experiments to test the performance of our method in sparse recovery. Concluding remarks are given in Section 6.

2 Notation and preliminaries

We use $\mathbb{R}$ and $\mathbb{R}^n$ to denote the set of real numbers and the n-dimensional Euclidean space. For any $x \in \mathbb{R}^n$, let x_i denote the ith entry of x, and $\mathrm{Diag}(x)$ denote the diagonal matrix whose ith diagonal entry is x_i, respectively. We denote the Euclidean norm of x by $\|x\|$, the ℓ_1 norm by $\|x\|_1$, the infinity norm (sup norm) by $\|x\|_\infty$, and the p quasi-norm by $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$, for any $0 < p < 1$. Moreover, we let $|x|^p$ denote the vector whose ith entry is $|x_i|^p$ and $(x)_+$ denote the vector whose ith entry is $\max\{x_i, 0\}$. Given an index set $I \subseteq \{1, \ldots, n\}$, let $\bar{I}$ denote the complement of I. For any vector x, we write $x_I \in \mathbb{R}^{|I|}$ to denote the restriction of x onto I. We also denote by A_I the matrix formed from a matrix A by picking the columns indexed by I. In addition, we use $\ker A$ to denote the null space of a matrix A.

For any closed set D, we let $\mathrm{dist}(x, D) = \inf_{y \in D} \|x - y\|$ denote the distance from x to D, and $\mathrm{conv}(D)$ denote the convex hull of D. We let $P_D(x)$ denote the set of closest points in D to $x \in \mathbb{R}^n$; this reduces to a singleton if D is additionally convex. For a closed convex set D, the normal cone at $x \in D$ is defined as

$$\mathcal{N}_D(x) := \{y : \ y^T(u - x) \leq 0 \ \ \forall u \in D\}.$$

The indicator function is denoted by δ_D, which is the function that is zero in D and is infinity elsewhere. Finally, we let $\mathbf{B}(a; r)$ denote the closed ball of radius r centered at a, i.e., $\mathbf{B}(a; r) = \{x \in \mathbb{R}^n : \ \|x - a\| \leq r\}$.

We recall from [27, Definition 8.3] that for a proper lower semicontinuous function f, the (limiting) subdifferential and horizon subdifferential are defined respectively as

$$\partial f(x) := \left\{v : \ \exists x^k \xrightarrow{f} x, \ v^k \to v \ \text{ with } \liminf_{z \to x^k} \frac{f(z) - f(x^k) - \langle v^k, z - x^k \rangle}{\|z - x^k\|} \geq 0 \ \forall k \right\},$$
$$\partial^\infty f(x) := \left\{v : \ \exists x^k \xrightarrow{f} x, \ \lambda_k v^k \to v, \lambda_k \downarrow 0 \ \text{ with } \liminf_{z \to x^k} \frac{f(z) - f(x^k) - \langle v^k, z - x^k \rangle}{\|z - x^k\|} \geq 0 \ \forall k \right\},$$

where $\lambda_k \downarrow 0$ means $\lambda_k > 0$ and $\lambda_k \to 0$, and $x^k \xrightarrow{f} x$ means both $x^k \to x$ and $f(x^k) \to f(x)$. It is well known that the following properties hold:

$$\begin{aligned} &\left\{v : \ \exists x^k \xrightarrow{f} x, \ v^k \to v \ , v^k \in \partial f(x^k)\right\} \subseteq \partial f(x), \\ &\left\{v : \ \exists x^k \xrightarrow{f} x, \ \lambda_k v^k \to v \ , \lambda_k \downarrow 0 \ , v^k \in \partial f(x^k)\right\} \subseteq \partial^\infty f(x). \end{aligned} \tag{2.1}$$

Moreover, if f is convex, the above definition of subdifferential coincides with the classical subdifferential in convex analysis [27, Proposition 8.12]. Furthermore, for a continuously differentiable f, we simply have $\partial f(x) = \{\nabla f(x)\}$, where $\nabla f(x)$ is the gradient of f at x [27, Exercise 8.8(b)]. We also use $\partial_{x_i} f(x)$ to denote the subdifferential with respect to the variable x_i. Finally, when $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ for some continuous function ϕ, we have from [27, Proposition 10.5] that

$$\partial\Phi(x) = \partial\phi(x_1) \times \partial\phi(x_2) \times \cdots \times \partial\phi(x_n). \tag{2.2}$$

For the convenience of readers, we now state our blanket assumptions on (1.1) explicitly here for easy reference.

Assumption 2.1 (**Blanket assumptions on (1.1)**). *Throughout this paper, Φ is a nonnegative continuous function. The feasible set of (1.1) is $S := S_1 \cap S_2$, where S_1 is a simple polyhedron given by $\{x : \ Dx \le d\}$, and*

$$S_2 = \{x : \ \|Ax - b\| \le \sigma, \quad Bx \le h\}.$$

Moreover, A has full row rank and there exists $x_0 \in S$ so that $\|Ax_0 - b\| < \sigma$.

We next present some auxiliary lemmas. The first lemma is a well-known result on error bound concerning S_1 and S_2, obtained as an immediate corollary of [22, Theorem 3.1].

Lemma 2.1. *There exists a $C > 0$ so that for all $x \in \mathbb{R}^n$, we have*

$$\mathrm{dist}(x, S) \le C\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1 + \|(Dx - d)_+\|_1\right].$$

Consequently, for any $x \in S_1$, we have

$$\mathrm{dist}(x, S) \le C\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1\right]. \tag{2.3}$$

The constant C in the above lemma cannot be explicitly computed in general. We next present a more explicit representation of this constant in some special cases. We start with the case where $S_1 = \mathbb{R}^n$ and B is vacuous, i.e., $S = S_2 = \{x : \ \|Ax - b\| \le \sigma\}$.

Lemma 2.2. *Suppose that $S = S_2 = \{x : \ \|Ax - b\| \le \sigma\}$. Then there exists a $C > 0$ so that for all x,*

$$\mathrm{dist}(x, S) \le \|A^\dagger\|(\|Ax - b\| - \sigma)_+ \le C(\|Ax - b\|^2 - \sigma^2)_+.$$

Indeed, C can be chosen to be $\frac{\|A^\dagger\|}{\sigma}$, where $A^\dagger = A^T(AA^T)^{-1}$ is the pseudo-inverse of A.

Proof. Notice that $S = A^\dagger \mathbf{B}(b;\sigma) + \ker A$. Moreover, for any x, $x - A^\dagger Ax \in \ker A$. Thus, we have

$$\begin{aligned}
&\mathrm{dist}(x, S) = \mathrm{dist}(A^\dagger Ax + [x - A^\dagger Ax], A^\dagger \mathbf{B}(b;\sigma) + \ker A)\\
&\ \le \mathrm{dist}(A^\dagger Ax, A^\dagger \mathbf{B}(b;\sigma)) \le \|A^\dagger\|\mathrm{dist}(Ax, \mathbf{B}(b;\sigma)) = \|A^\dagger\|(\|Ax - b\| - \sigma)_+,
\end{aligned}$$

where the last equality follows from a direct computation based on the fact that the projection from any point $u \notin \mathbf{B}(b;\sigma)$ onto $\mathbf{B}(b;\sigma)$ is $b + \sigma\frac{u-b}{\|u-b\|}$. The conclusion of the lemma now follows from the above estimate and the following simple relation:

$$(\|Ax - b\| - \sigma)_+ = \left(\frac{\|Ax - b\|^2 - \sigma^2}{\|Ax - b\| + \sigma}\right)_+ \le \frac{1}{\sigma}(\|Ax - b\|^2 - \sigma^2)_+.$$

■

We next consider the case where S is compact. We refer the readers to [8, Lemma 3.2.3] and [8, Remark 3.2.4] for an explicit finite upper bound for the constant β in (2.4) below.

Lemma 2.3. *Suppose there exist $x_s \in S$, $R > \delta > 0$ so that $\sup_{u\in \mathbf{B}(x_s;\delta)} \|Au - b\| \le \sigma$ and $S \subseteq \mathbf{B}(x_s; R)$. Then there exists $\beta > 0$ so that for all $x \in \mathbb{R}^n$, we have*

$$\operatorname{dist}(x, S) \le 2\left(1 + \frac{R}{\delta}\right)\left(\frac{\|A^\dagger\|}{\sigma}(\|Ax - b\|^2 - \sigma^2)_+ + \beta\left\|\begin{pmatrix} Bx - h \\ Dx - d\end{pmatrix}_+\right\|_1\right). \tag{2.4}$$

Proof. Let $\Omega_1 = \{x : \ Bx \le h,\ Dx \le d\}$ and $\Omega_2 = \{x : \|Ax - b\| \le \sigma\}$. Then $S = \Omega_1 \cap \Omega_2$. From the assumptions and [20, Lemma 2.1] (see also [19, Lemma 4.10]), we see that for all $x \in \mathbb{R}^n$, we have

$$\operatorname{dist}(x, S) \le 2\left(1 + \frac{R}{\delta}\right)\max\{\operatorname{dist}(x, \Omega_1), \operatorname{dist}(x, \Omega_2)\}. \tag{2.5}$$

The desired conclusion now follows from (2.5), Lemma 2.2 and [8, Lemma 3.2.3] (Hoffman error bound). ∎

We end this section with the following auxiliary lemmas concerning the function $t \mapsto t^p$, $0 < p < 1$.

Lemma 2.4. *Let $0 < p < 1$. For any nonnegative numbers s and t, it holds that*

$$|s^p - t^p| \le |s - t|^p.$$

Proof. Without loss of generality, we may assume that $s \ge t$. Consider $h(r) := 1 - r^p - (1-r)^p$ for $r \in [0, 1]$. Simple differential calculus shows that $h(r) \le h(0) = 0 = h(1)$ whenever $r \in [0, 1]$. The desired conclusion then follows by setting $r = \frac{t}{s}$. ∎

Lemma 2.5. *Let $0 < p < 1$. Then the following statements hold.*

(i) Let $h(t) = |t|^p$. Then $\partial h(0) = \partial^\infty h(0) = \mathbb{R}$.

(ii) Let $H(x) = \sum_{i=1}^n |x_i|^p$ and fix any $x^* \in \mathbb{R}^n$. Let $I := \{i : \ x^* \neq 0\}$. Then

$$\partial^\infty H(x^*) = \{v : \ v_i = 0 \text{ for } i \in I\}.$$

Proof. We first prove (i). Consider the set

$$\hat\partial h(0) := \left\{s \in \mathbb{R} : \ \liminf_{t\to 0} \frac{|t|^p - st}{|t|} \ge 0\right\}.$$

Since $\liminf_{t\to 0} |t|^{p-1} = \infty$ due to $0 < p < 1$, we see immediately that $\hat\partial h(0) = \mathbb{R}$. Since we have from [27, Theorem 8.6] that $\hat\partial h(0) \subseteq \partial h(0)$ and that $\partial^\infty h(0)$ contains the recession cone of $\hat\partial h(0)$, we conclude further that $\partial h(0) = \partial^\infty h(0) = \mathbb{R}$.

We next prove (ii). Part (i) together with the fact that $\hat\partial h(0) = \mathbb{R}$ and [27, Corollary 8.11] shows that $h(t) = |t|^p$ is regular at 0. In addition, h is clearly regular at any $t \neq 0$. Then, according to [27, Proposition 10.5], we have

$$\partial^\infty H(x^*) = \partial^\infty h(x_1^*) \times \cdots \partial^\infty h(x_n^*),$$

from which the conclusion follows immediately. ∎

3 Exact Penalization

Problem (1.1) is a constrained optimization problem, which can be difficult to solve when the constraint set S is complicated. In the case when Φ is convex, $\sigma > 0$, $S_1 = \mathbb{R}^n$ and B is vacuous, i.e., (1.2), it is well known that the problem is equivalent to solving the unconstrained optimization problem (1.3) for some suitable $\lambda > 0$; see, for example, [11]. However, as we will illustrate in the next example, this is no longer true for a general nonconvex Φ.

Example 3.1. Consider the following one-dimensional optimization problem:

$$\begin{array}{ll} \min\limits_{t} & \phi(t) \\ \text{s.t.} & |t-a| \le \gamma a \end{array} \tag{3.1}$$

for some $a > 0$ and $\gamma \in (0,1)$. Assume that ϕ is strictly increasing on $[0,\infty)$.

It is clear that $t^* = (1-\gamma)a$ is the global minimizer of (3.1). Suppose that ϕ is twice continuously differentiable at t^*. Then it is easy to check from the first-order optimality condition that t^* is a stationary point of

$$\min_{t} \ \lambda(t-a)^2 + \phi(t) \tag{3.2}$$

only when $\lambda = \phi'(t^*)/(2\gamma a)$, which is nonnegative since ϕ is monotone. Next, the second derivative of the objective of (3.2) with $\lambda = \phi'(t^*)/(2\gamma a)$ at t^* is given by

$$2\lambda + \phi''(t^*) = \frac{\phi'(t^*)}{\gamma a} + \phi''(t^*). \tag{3.3}$$

If this quantity is negative, then t^* cannot be a local minimizer of (3.2) even for $\lambda = \phi'(t^*)/(2\gamma a)$, and consequently, t^* cannot be a local minimizer of (3.2) for any $\lambda > 0$.

Some concrete examples of ϕ and a such that (3.3) is negative are given below, where the ϕ's are building blocks for widely used nonconvex regularization functions.

1. bridge penalty $\phi(t) = |t|^p$ for $0 < p < 1$ [16, 17].
   For any $a > 0$, (3.3) equals $p(t^*)^{p-2}(p-2+1/\gamma)$. Hence, (3.3) is negative if $p < 2 - 1/\gamma$. Since p is positive, this can happen when $\gamma > 1/(2-p)$;

2. fraction penalty $\phi(t) = \alpha|t|/(1+\alpha|t|)$ for $\alpha > 0$ [13].
   For any $a > 0$, a direct computation shows that (3.3) equals $(\alpha/\gamma a)\,(1+\alpha t^*)^{-3}\,[1 + (1-3\gamma)\alpha a]$, which is negative when $1 + (1-3\gamma)\alpha a < 0$. Since a and α are both positive, this can happen when $\gamma > (1+\alpha a)/(3\alpha a)$;

3. logistic penalty $\phi(t) = \log(1+\alpha|t|)$ for $\alpha > 0$ [23].
   For any $a > 0$, (3.3) equals $(\alpha/\gamma a)\,(1+\alpha t^*)^{-2}\,[1+(1-2\gamma)\alpha a]$, which is negative if $1 + (1-2\gamma)\alpha a < 0$. Since a and α are both positive, this can happen when $\gamma > (1+\alpha a)/(2\alpha a)$.

Example 3.1 shows that the negativity of ϕ'' prevents us from building a relationship between (1.2) and (1.3) regarding global or local minimizers. In general, we cannot always find a λ such that the intersection of the sets of global (local) minimizers of (1.2) and (1.3) is nonempty, when ϕ is monotone and concave in $[0,\infty)$.

In order to build a simpler optimization problem whose local/global minimizers are related to the constrained problem (1.1) (which contains (1.2) as a special case) when Φ is possibly nonconvex, we adopt the penalty approach. We hereby emphasize again that there is little theory concerning the penalty approach when Φ is non-Lipschitz. Moreover, it is not common in the literature to consider *partial* penalization that keeps part of the constraints, S_1, in the penalized problem (1.4). In this section, we shall study various (partial) exact penalization results concerning the problems (1.1) and (1.4), for both locally Lipschitz and non-Lipschitz objectives Φ.

3.1 A general penalization result

We first present some results regarding exact penalty reformulation for a general optimization problem. These results will be applied in subsequent subsections to derive various exact penalization results. The following lemma is similar to [18, Proposition 4]. For self-contained purpose, we provide a simple proof.

Lemma 3.1. *Consider the problem*

$$\min_{x\in\Omega_1\cap\Omega_2} f(x), \tag{3.4}$$

where Ω_1 and Ω_2 are two nonempty closed sets in $\mathbb{R}^n$. Assume that f is Lipschitz continuous in Ω_1 with a Lipschitz constant $L_f > 0$, and moreover, problem (3.4) has at least one optimal solution. Suppose in addition that there is a function $Q : \Omega_1 \to \mathbb{R}_+$ satisifying

$$Q(x) \geq \mathrm{dist}(x, \Omega_1 \cap \Omega_2) \quad \forall x \in \Omega_1; \quad Q(x) = 0 \quad \forall x \in \Omega_1 \cap \Omega_2. \tag{3.5}$$

Then it holds that:

(i) if x^* is a global minimizer of (3.4), then x^* is a global minimizer of

$$\min_{x\in\Omega_1} \; f(x) + \lambda Q(x) \tag{3.6}$$

whenever $\lambda \geq L_f$;

(ii) if x^* is a global minimizer of (3.6) for some $\lambda > L_f$, then x^* is a global minimizer of (3.4).

Proof. Since f is Lipschitz continuous in Ω_1 with a Lipschitz constant $L_f > 0$, it follows that for all $\lambda \geq L_f$,

$$f(x) + \lambda \,\mathrm{dist}(x, \Omega_1 \cap \Omega_2) \;\geq\; f(y) \qquad \forall x \in \Omega_1,\ \forall y \in P_{\Omega_1\cap\Omega_2}(x),$$

which together with (3.5) implies that for any $\lambda \geq L_f$,

$$f(x) + \lambda Q(x) \;\geq\; f(y) \qquad \forall x \in \Omega_1,\ \forall y \in P_{\Omega_1\cap\Omega_2}(x).$$

Using this relation, one can observe that for all $\lambda \geq L_f$,

$$\begin{aligned}\inf_{x\in\Omega_1} \{f(x) + \lambda Q(x)\} &\geq \inf_{x\in\Omega_1, y\in P_{\Omega_1\cap\Omega_2}(x)} f(y) \;=\; \inf_{x\in\Omega_1\cap\Omega_2} f(x) \\ &= \inf_{x\in\Omega_1\cap\Omega_2} f(x) + \lambda Q(x) \geq \inf_{x\in\Omega_1} \{f(x) + \lambda Q(x)\},\end{aligned}$$

where the second equality follows from the fact that $Q(x) = 0$ for all $x \in \Omega_1 \cap \Omega_2$. Statement (i) follows immediately from this relation.

We next prove statement (ii). Suppose that $x^* \in \Omega_1$ is a global minimizer of (3.6) for some $\lambda > L_f$. Using this and $Q(x) = 0$ on $\Omega_1 \cap \Omega_2$, we have

$$f(x^*) + \lambda Q(x^*) \le f(x), \tag{3.7}$$

for any $x \in \Omega_1 \cap \Omega_2$. This together with (3.5) implies that for any $x \in P_{\Omega_1 \cap \Omega_2}(x^*)$,

$$f(x^*) + \lambda \operatorname{dist}(x^*, \Omega_1 \cap \Omega_2) \le f(x).$$

Using this relation and Lipschitz continuity of f, one can obtain that for any $x \in P_{\Omega_1 \cap \Omega_2}(x^*)$,

$$\operatorname{dist}(x^*, \Omega_1 \cap \Omega_2) \le \frac{1}{\lambda}(f(x) - f(x^*)) \le \frac{L_f}{\lambda}\|x - x^*\| = \frac{L_f}{\lambda}\operatorname{dist}(x^*, \Omega_1 \cap \Omega_2),$$

which along with $\lambda > L_f$ yields $\operatorname{dist}(x^*, \Omega_1 \cap \Omega_2) = 0$, that is, $x^* \in \Omega_1 \cap \Omega_2$. In addition, by (3.7) and $Q(x^*) \ge 0$, one can see that $f(x) \ge f(x^*)$ for any $x \in \Omega_1 \cap \Omega_2$. Hence, x^* is a global minimizer of (3.4). ∎

We next state a result regarding the local minimizers of problems (3.4) and (3.6), whose proof is similar to that of Lemma 3.1 and thus omitted.

Corollary 3.1. *Assume that f is locally Lipschitz continuous in Ω_1 and Q satisfies (3.5). Suppose that x^* is a local minimizer of (3.4). Then there exists a $\lambda^* > 0$ such that x^* is a local minimizer of (3.6) whenever $\lambda \ge \lambda^*$.*

3.2 When Φ is locally Lipschitz continuous

In this subsection, we consider the case where Φ is locally Lipschitz continuous and derive the corresponding exact regularization results concerning models (1.1) and (1.4). This covers a lot of regularization functions used in practice, including many difference-of-convex functions; see, for example, [14, 29].

Our main result concerns local and global minimizers of models (1.1) and (1.4).

Theorem 3.1 (Local & global minimizers). *Suppose that Φ is locally Lipschitz continuous in S_1 and x^* is a local minimizer of (1.1). Then there exists a $\lambda^* > 0$ such that x^* is a local minimizer of (1.4) whenever $\lambda \ge \lambda^*$. If Φ is indeed globally Lipschitz continuous in S_1, then there exists a $\lambda^* > 0$ such that any global minimizer of (1.1) is a global minimizer of (1.4) whenever $\lambda \ge \lambda^*$; moreover, if x^* is a global minimizer of (1.4) for some $\lambda > \lambda^*$, then x^* is a global minimizer of (1.1).*

Proof. From Lemma 2.1, we see that there exists a $C > 0$ so that for all $x \in S_1$, we have

$$\operatorname{dist}(x, S) \le C\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1\right].$$

The first conclusion now follows immediately from Corollary 3.1 by setting $f(x) = \Phi(x)$, $Q(x) = C\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1\right]$, $\Omega_1 = S_1$ and $\Omega_2 = S_2$, while the second conclusion follows from Lemma 3.1. ∎

Remark 3.1. *It is not hard to see from the proof of Theorem 3.1 that with an explicit error bound modulus $C > 0$ in (2.3), the λ^* in the theorem can be chosen to be CL, where L is a local (resp., global) Lipschitz constant Φ.* ■

In the next example, we present explicit exact penalty functions for problem (3.1) with some specific choices of ϕ.

Example 3.2. Notice that the fraction penalty function and the logistic penalty function considered in Example 3.1 are (globally) Lipschitz continuous, and have a Lipschitz constant α. From Theorem 3.1 and Remark 3.1, we conclude that any global minimizer of (3.1) is a global minimizer of the problem

$$\min_t \ \lambda(|t-a|^2 - \gamma^2 a^2)_+ + \phi(t),$$

whenever $\lambda \geq \frac{\alpha}{\gamma a}$, since C can be chosen to be $\frac{\|A^\dagger\|}{\sigma} = \frac{1}{\gamma a}$ by Lemma 2.2. The bridge penalty function, on the other hand, is locally Lipschitz continuous everywhere except at 0. Since $\gamma \in (0,1)$, t^p is Lipschitz continuous on $[(1-\gamma)a/2, \infty)$ with Lipschitz constant $p[(1-\gamma)a/2]^{p-1}$. From Theorem 3.1 and Remark 3.1, we conclude that any local minimizer of (3.1) is a local minimizer of the problem

$$\min_t \ \lambda(|t-a|^2 - \gamma^2 a^2)_+ + \phi(t),$$

whenever $\lambda \geq \frac{p[(1-\gamma)a/2]^{p-1}}{\gamma a}$.

3.3 When Φ is not locally Lipschitz continuous at some points

In this subsection, we suppose that $\Phi(x)$ is not locally Lipschitz continuous at some points. To proceed, we make an assumption for Φ that will be used subsequently.

Assumption 3.1. *The function $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ is continuous and nonnegative with $\phi(0) = 0$, and is locally Lipschitz continuous everywhere except at 0. Moreover, for any $L > 0$, there exists an $\epsilon > 0$ such that whenever $|t| < \epsilon$, we have*

$$\phi(t) \geq L|t|. \tag{3.8}$$

It is not hard to show that the widely used bridge penalty function $|x|^p$, for $0 < p < 1$, satisfies this assumption.

Theorem 3.2 (**Local minimizers**). *Suppose that x^* is a local minimizer of (1.1) with a Φ satisfying Assumption 3.1. Then there exists a $\lambda^* > 0$ such that x^* is a local minimizer of (1.4) whenever $\lambda \geq \lambda^*$.*

Proof. Suppose first that $x^* = 0$. Fix any bounded neighborhood U of 0 and any $\lambda > 0$. Let L denote a Lipschitz constant the function $x \mapsto \lambda[(\|Ax-b\|^2 - \sigma^2)_+ + \|(Bx-h)_+\|_1]$ on U. For this L, by Assumption 3.1, there exists a neighborhood $V \subseteq U$ of zero such that $\Phi(x) \geq L\|x\|_1$ whenever $x \in V$. Then for any $x \in V$, we have

$$\begin{aligned}
&\lambda[(\|Ax-b\|^2 - \sigma^2)_+ + \|(Bx-h)_+\|_1] + \Phi(x)\\
&\geq \lambda[(\|Ax-b\|^2 - \sigma^2)_+ + \|(Bx-h)_+\|_1] + L\|x\|_1\\
&\geq \lambda[(\|Ax-b\|^2 - \sigma^2)_+ + \|(Bx-h)_+\|_1] + L\|x\|\\
&\geq \lambda[(\|b\|^2 - \sigma^2)_+ + \|(-h)_+\|_1],
\end{aligned}$$

where the last inequality follows from the definition of L being a Lipschitz constant. This shows that $x^* = 0$ is a local minimizer of (1.4) for any $\lambda > 0$. Thus, to complete the proof, from now on, we assume that $x^* \neq 0$. Let I denote the support of x^*, i.e., $I := \{i : x_i^* \neq 0\}$. Then $I \neq \emptyset$.

Since x^* is a local minimizer of (1.1), it follows that x_I^* is a local minimizer of the following optimization problem:

$$\begin{array}{ll} \min\limits_{x_I} & \sum_{i\in I} \phi(x_i) \\ \text{s.t.} & \|A_I x_I - b\| \le \sigma, \quad B_I x_I \le h, \quad D_I x_I \le d. \end{array} \tag{3.9}$$

Let $\tilde{\epsilon} = \frac{1}{2}\min\{|x_i^*| : i \in I\} > 0$. Then we can choose a small $\delta > 0$ so that x_I^* is a local minimizer of (3.9) and that $\min_{i\in I}|x_i| > \tilde{\epsilon}$ for all $x_I \in \mathbf{B}(x_I^*;\delta)$. Next, consider $\Omega_1 = \mathbf{B}(x_I^*;\delta)$ and $\Omega_2 = \{x_I : \|A_I x_I - b\| \le \sigma, \ B_I x_I \le h, \ D_I x_I \le d\}$. Then we have from [19, Lemma 4.9] that

$$\mathrm{dist}(x_I, \Omega_1 \cap \Omega_2) \le 4\ \mathrm{dist}(x_I, \Omega_2)$$

for all $x_I \in \Omega_1$. Using this and applying Corollary 3.1 with $f(x_I) = \sum_{i\in I}\phi(x_i)$, the Ω_1 and Ω_2 as defined above, and

$$Q(x_I) = 4C\left[(\|A_I x_I - b\|^2 - \sigma^2)_+ + \|(B_I x_I - h)_+\|_1 + \|(D_I x_I - d)_+\|_1\right]$$

for the C given in Lemma 2.1, we conclude that there exists a $\lambda^* > 0$ so that for any $\lambda \ge \lambda^*$, there is a neighborhood U_I of 0 with $U_I \subseteq \mathbf{B}(0;\frac{\delta}{2})$ such that $G_\lambda^I(x_I) \ge G_\lambda^I(x_I^*)$ whenever $x_I \in x_I^* + U_I$, where

$$G_\lambda^I(x_I) = \lambda\left[(\|A_I x_I - b\|^2 - \sigma^2)_+ + \|(B_I x_I - h)_+\|_1 + \|(D_I x_I - d)_+\|_1\right] + \sum_{i\in I}\phi(x_i).$$

We now show that x^* is a local minimizer of (1.4) with $\lambda \ge \lambda^*$. Fix any $\epsilon > 0$ and any $\lambda \ge \lambda^*$. Consider the (bounded) neighborhood $U := U_I \times (-\epsilon,\epsilon)^{n-|I|}$ of 0 and let M be a Lipschitz constant the function

$$g_\lambda(x) = \lambda\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1 + \|(Dx - d)_+\|_1\right]$$

over $x^* + U$. Taking $L = M$ in Assumption 3.1, we see that there exists an $\epsilon_0 \in (0,\epsilon)$ such that (3.8) holds with M in place of L whenever $|t| < \epsilon_0$. Then, for any $v \in U_I \times (-\epsilon_0,\epsilon_0)^{n-|I|}$ with $x^* + v \in S_1$, we have

$$\begin{aligned} F_\lambda(x^* + v) &= F_\lambda\left(x^* + \begin{pmatrix} v_I \\ v_{\bar{I}} \end{pmatrix}\right) = g_\lambda\left(x^* + \begin{pmatrix} v_I \\ v_{\bar{I}} \end{pmatrix}\right) + \sum_{i\in I}\phi(x_i^* + v_i) + \sum_{i\notin I}\phi(v_i) \\ &\ge g_\lambda\begin{pmatrix} x_I^* + v_I \\ 0 \end{pmatrix} - M\|v_{\bar{I}}\| + \sum_{i\in I}\phi(x_i^* + v_i) + M\|v_{\bar{I}}\|_1 \\ &\ge G_\lambda^I(x_I^*) = F_\lambda(x^*), \end{aligned}$$

where the first inequality follows from the Lipschitz continuity of g_λ with Lipschitz constant M and (3.8) with $L = M$, and the last inequality follows from the local optimality of x_I^*, while the second and the last equalities follow from $\|(D(x^* + v) - d)_+\|_1 = 0$ since $x^* + v \in S_1$. This shows that x^* is locally optimal for (1.4) with $\lambda \ge \lambda^*$, and completes the proof. ■

We next study ϵ-minimizers of (1.1) and (1.4), which are defined as follows.

Definition 3.1. *Let $\epsilon > 0$.*

1. *We say that x_ϵ is an ϵ-minimizer of (1.1), if $x_\epsilon \in S$ and $\Phi(x_\epsilon) \leq \inf_{x\in S} \Phi(x) + \epsilon$.*
2. *We say that x_ϵ is an ϵ-minimizer of (1.4), if $x_\epsilon \in S_1$ and $F_\lambda(x_\epsilon) \leq \inf_{x\in S_1} F_\lambda(x) + \epsilon$.*

In order to establish results concerning ϵ-minimizers, we also need the following definition.

Definition 3.2. *We say that a globally Lipschitz continuous function Ψ with a Lipschitz constant L is an (L, ϵ)-approximation to Φ if $0 \leq \Psi(x) - \Phi(x) \leq \epsilon$ for all x.*

As a concrete example of such an approximation, consider the case where $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ with $\phi(t) = |t|^p$ for some $0 < p < 1$. We can consider the following smoothing function of $|t|$:

$$\psi_\mu(t) = \begin{cases} |t| & \text{if } |t| \geq \mu, \\ \frac{t^2}{2\mu} + \frac{\mu}{2} & \text{otherwise.} \end{cases}$$

Notice that for a fixed $\mu > 0$, the minimum and maximum values of $\psi_\mu(t) - |t|$ are attained at $|t| \geq \mu$ and $t = 0$, respectively. Let

$$\Psi_\mu(x) = \sum_{i=1}^n \psi_\mu(x_i)^p.$$

Then we have from the above discussion and Lemma 2.4 that

$$0 \leq \Psi_\mu(x) - \|x\|_p^p \leq n\left(\frac{\mu}{2}\right)^p. \tag{3.10}$$

Moreover, for a fixed $\mu > 0$, the function Ψ_μ is continuously differentiable. The maximum value of $|(\psi_\mu(t)^p)'|$ is attained at $t = \mu$, and hence we have

$$|\Psi_\mu(x) - \Psi_\mu(y)| \leq \sqrt{n}p\mu^{p-1}\|x - y\|. \tag{3.11}$$

The inequalities (3.10) and (3.11) show that Ψ_μ is a $(\sqrt{n}p\mu^{p-1}, n(\mu/2)^p)$-approximation to Φ when $\phi(t) = |t|^p$.

From the definition of an (L, ϵ)-approximation Ψ, it is easy to show that any global minimizer of

$$\begin{aligned} \min_{x\in S_1} \quad & \Psi(x) \\ \text{s.t.} \quad & \|Ax - b\| \leq \sigma, \quad Bx \leq h, \end{aligned} \tag{3.12}$$

is an ϵ-minimizer of (1.1). Conversely, any global minimizer x^* of (1.1) is an ϵ-minimizer of (3.12). Our next result concerns the global minimizers of (1.1) and the ϵ-minimizers of (1.4).

Theorem 3.3 (**ϵ-minimizers**). *Suppose that Φ admits an $(L, \epsilon/2)$-approximation Ψ. Then for any global minimizer x^* of (1.1), there exists a $\lambda^* > 0$ so that x^* is an ϵ-minimizer of (1.4) whenever $\lambda \geq \lambda^*$, i.e.,*

$$F_\lambda(x^*) \leq \inf_{x\in S_1} F_\lambda(x) + \epsilon; \tag{3.13}$$

in particular, one can take $\lambda^* = CL$, where C is the constant in Lemma 2.1.

Proof. From the definition of an $(L, \epsilon/2)$-approximation, we see that any global minimizer x^* of (1.1) is an $\epsilon/2$-minimizer of (3.12). Moreover, since Ψ is globally Lipschitz continuous with Lipschitz constant L, we have for any $x \in S_1$ that

$$\tilde{L}\ \mathrm{dist}(x, S) + \Psi(x) = \tilde{L}\ \|x - P_S(x)\| + \Psi(x) \geq \Psi(P_S(x)) \geq \Psi(x^*) - \frac{\epsilon}{2},$$

where $\tilde{L}$ is any number greater than or equal to L and the second inequality follows from the $\epsilon/2$-optimality of x^* for (3.12). This shows that x^* is an $\epsilon/2$-minimizer of the optimization problem

$$\min_{x \in S_1}\ \tilde{L}\ \mathrm{dist}(x, S) + \Psi(x).$$

Combining this fact with Lemma 2.1, it is not hard to show that x^* is an $\epsilon/2$-minimizer of

$$\min_{x \in S_1}\ C\tilde{L}\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1\right] + \Psi(x).$$

Using this and the fact that $0 \leq \Psi(x) - \Phi(x) \leq \epsilon/2$ for all x, we have further that for all $x \in S_1$,

$$\begin{aligned}
F_{C\tilde{L}}(x) &= C\tilde{L}\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1\right] + \Phi(x) \\
&\geq C\tilde{L}\left[(\|Ax - b\|^2 - \sigma^2)_+ + \|(Bx - h)_+\|_1\right] + \Psi(x) - \frac{\epsilon}{2} \\
&\geq C\tilde{L}\left[(\|Ax^* - b\|^2 - \sigma^2)_+ + \|(Bx^* - h)_+\|_1\right] + \Psi(x^*) - \frac{\epsilon}{2} - \frac{\epsilon}{2} \\
&= F_{C\tilde{L}}(x^*) - \epsilon,
\end{aligned}$$

i.e., (3.13) holds with $\lambda^* = CL$. ∎

So far we have shown that if x^* is locally or globally optimal for (1.1), then it is also optimal in some sense for (1.4), when λ is sufficiently large. Conversely, it is clear that if x^* is optimal (locally or being an ϵ-minimizer) for (1.4) for some $\lambda > 0$, and x^* is also feasible for (1.1), then it is also optimal for (1.1). Our next result studies the case when x^* is not necessarily feasible for (1.1).

Theorem 3.4 (**ϵ-minimizers feasible for (1.1)**). *Suppose that $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ with ϕ being Hölder continuous for some $0 < p < 1$, i.e., there exists a $K > 0$ such that*

$$|\phi(s) - \phi(t)| \leq K|s - t|^p$$

for any s, $t \in \mathbb{R}$. Take any $\epsilon > 0$ and fix any $\tilde{x} \in S$. Consider any

$$\lambda > \frac{K^{\frac{1}{p}} C\Phi(\tilde{x})}{(n^{\frac{p}{2}-1}\epsilon)^{\frac{1}{p}}},$$

with C chosen as in Lemma 2.1. Then for any global minimizer x_λ of (1.4), the projection $P_S(x_\lambda)$ is an ϵ-minimizer of (1.1).

Proof. We first note from the global optimality of x_λ that $F_\lambda(x_\lambda) \leq F_\lambda(\tilde{x})$, from which we immediately obtain that

$$(\|Ax_\lambda - b\|^2 - \sigma^2)_+ + \|(Bx_\lambda - h)_+\|_1 \leq \frac{1}{\lambda}F_\lambda(x_\lambda) \leq \frac{1}{\lambda}F_\lambda(\tilde{x}) = \frac{1}{\lambda}\Phi(\tilde{x}). \tag{3.14}$$

Next, for the projection $P_S(x_\lambda)$, we have

$$\begin{aligned}
&\Phi(P_S(x_\lambda)) - \Phi(x_\lambda) \le K \sum_{i=1}^n |[P_S(x_\lambda)]_i - [x_\lambda]_i|^p = nK\frac{1}{n}\sum_{i=1}^n \left(|[P_S(x_\lambda)]_i - [x_\lambda]_i|^2\right)^{\frac{p}{2}} \\
&\le nK\left(\frac{1}{n}\sum_{i=1}^n |[P_S(x_\lambda)]_i - [x_\lambda]_i|^2\right)^{\frac{p}{2}} = Kn^{1-\frac{p}{2}}\|P_S(x_\lambda) - x_\lambda\|^p \\
&= Kn^{1-\frac{p}{2}}\mathrm{dist}^p(x_\lambda, S) \le KC^p n^{1-\frac{p}{2}}\left[(\|Ax_\lambda - b\|^2 - \sigma^2)_+ + \|(Bx_\lambda - h)_+\|_1\right]^p \\
&\le Kn^{1-\frac{p}{2}}\left(\frac{C\Phi(\tilde{x})}{\lambda}\right)^p,
\end{aligned} \tag{3.15}$$

where the first inequality follows from the assumption on Hölder continuity, the second one holds due to the concavity of the function $t \mapsto t^{\frac{p}{2}}$ for nonnegative t, the third inequality follows from Lemma 2.1 and the fact that $x_\lambda \in S_1$, while the last one follows from (3.14). On the other hand, for any $x \in S$, we have from the optimality of x_λ for (1.4) and the definition of F_λ that $F_\lambda(x_\lambda) \le F_\lambda(x) = \Phi(x)$. From this we see immediately that

$$\Phi(x_\lambda) \le F_\lambda(x_\lambda) \le \inf_{x\in S} \Phi(x).$$

Combining this with (3.15), we obtain further that

$$0 \le \Phi(P_S(x_\lambda)) - \inf_{x\in S}\Phi(x) \le Kn^{1-\frac{p}{2}}\left(\frac{C\Phi(\tilde{x})}{\lambda}\right)^p < \epsilon,$$

from our choice of λ. This shows that $P_S(x_\lambda)$ is an ϵ-minimizer of (1.1). ∎

From Lemma 2.4, it is easy to see that $t \mapsto |t|^p$, $0 < p < 1$, is Hölder continuous with $K = 1$. Thus, we have the following immediate corollary when $\Phi(x) = \|x\|_p^p$, $0 < p < 1$.

Corollary 3.2. *Suppose that $\Phi(x) = \|x\|_p^p$ for some $0 < p < 1$. Take any $\epsilon > 0$ and fix any $\tilde{x} \in S$. Consider any*

$$\lambda > \frac{C\|\tilde{x}\|_p^p}{(n^{\frac{p}{2}-1}\epsilon)^{\frac{1}{p}}},$$

with C chosen as in Lemma 2.1. Then for any global minimizer x_λ of (1.4), the projection $P_S(x_\lambda)$ is an ϵ-minimizer of (1.1).

4 Algorithm

In this section we propose a penalty method for solving problem (1.1). Based on our discussion in the previous section, a natural penalty method for solving (1.1) would be to solve the problem (1.4) once with an exact penalty parameter λ. This approach is, however, not appealing in practice because such λ may be hard to estimate, or it may be over-estimated and the resulting penalty problem becomes very ill-conditioned. To circumvent these potential difficulties, we propose a practical penalty method that solves a sequence of penalty subproblems in the form of (1.4) with a gradually increased penalty parameter. In addition, the approximate solution of the current subproblem will be used as the starting point for solving the next subproblem.

Our algorithm is presented in Section 4.2, where we show that any cluster point of the sequence generated from our algorithm is a KKT point of problem (1.1), under a suitable constraint qualification. To prepare for our convergence analysis, we start by discussing the first-order optimality conditions for problems (1.1) and (1.4) and describing the constraint qualification in Section 4.1.

4.1 First-order optimality conditions

In this subsection, we discuss the first-order optimality conditions for problems (1.1) and (1.4).

We first look at model (1.4). Since the objective is a sum of a locally Lipschitz continuous function and the lower semicontinuous function $\Phi + \delta_{S_1}$, it follows from [27, Theorem 8.15], [27, Theorem 10.1] and [27, Exercise 10.10] that at any locally optimal solution $\bar{x}$ of (1.4), we have

$$0 \in \partial(\lambda(\|A\cdot - b\|^2 - \sigma^2)_+)(\bar{x}) + \partial(\lambda\|(B\cdot - h)_+\|_1)(\bar{x}) + \partial(\Phi + \delta_{S_1})(\bar{x}). \tag{4.1}$$

This motivates the following definition.

Definition 4.1 (First-order stationary point of (1.4)). *We say that x^* is a first-order stationary point of (1.4) if $x^* \in S_1$ and (4.1) is satisfied with x^* in place of $\bar{x}$.*

In the special case where $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ with $\phi(t) = |t|^p$, it is easy to check that $\partial\phi(t) = \{p\,\mathrm{sign}(t)\,|t|^{p-1}\}$ whenever $t \neq 0$ and, from Lemma 2.5 (i), we have $\partial\phi(0) = \mathbb{R}$. Moreover, for the first subdifferential in (4.1), we have the following explicit expression

$$\partial(\lambda(\|A\cdot - b\|^2 - \sigma^2)_+)(\bar{x}) = \begin{cases} 0 & \text{if } \|A\bar{x} - b\| < \sigma, \\ \mathrm{conv}\{0, 2\lambda A^T(A\bar{x} - b)\} & \text{if } \|A\bar{x} - b\| = \sigma, \\ 2\lambda A^T(A\bar{x} - b) & \text{otherwise.} \end{cases} \tag{4.2}$$

Thus, in the case when B is vacuous and $S_1 = \mathbb{R}^n$, we have that x^* is a first-order stationary point of (1.4) if and only if

$$0 = 2\nu\lambda[A^T(Ax^* - b)]_i + p\,\mathrm{sign}(x_i^*)\,|x_i^*|^{p-1}, \quad \forall i \in I \tag{4.3}$$

with $I = \{i : \ x_i^* \neq 0\}$ for some ν satisfying

$$\nu \begin{cases} = 0 & \text{if } \|Ax^* - b\| < \sigma, \\ \in [0, 1] & \text{if } \|Ax^* - b\| = \sigma, \\ = 1 & \text{otherwise.} \end{cases}$$

This is because the inclusion (4.1) is trivial for $i \notin I$. Using the definition of I, it is not hard to see that (4.3) is further equivalent to

$$0 = 2\nu\lambda\mathrm{Diag}(x^*)A^T(Ax^* - b) + p|x^*|^p, \tag{4.4}$$

with the same ν defined above.

We next turn to the KKT points of (1.1). We recall from [27, Theorem 8.15] that at any locally optimal solution $\bar{x}$ of (1.1), we have

$$0 \in \mathcal{N}_{S_2}(\bar{x}) + \partial(\Phi + \delta_{S_1})(\bar{x}), \tag{4.5}$$

assuming the following constraint qualification holds:

$$-\partial^\infty(\Phi + \delta_{S_1})(\bar{x}) \cap \mathcal{N}_{S_2}(\bar{x}) = \{0\}. \tag{4.6}$$

This motivates the following definition.

Definition 4.2 (KKT point of (1.1)). *We say that x^* is a KKT point of (1.1) if $x^* \in S$ and (4.5) is satisfied with x^* in place of $\bar{x}$.*

Since there exists x_0 with $\|Ax_0 - b\| < \sigma$, in the case when B is vacuous and $S_1 = \mathbb{R}^n$, we have

$$\mathcal{N}_S(\bar{x}) = \begin{cases} \{\mu A^T(A\bar{x} - b) : \ \mu \geq 0\} \neq \{0\} & \text{if } \|A\bar{x} - b\| = \sigma, \\ \{0\} & \text{if } \|A\bar{x} - b\| < \sigma; \end{cases} \tag{4.7}$$

see, for example, Theorem 1.3.5 in [15, Section D]. In the special case where $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ with $\phi(t) = |t|^p$ and that B is vacuous and $S_1 = \mathbb{R}^n$, similarly as above, one can see that an x^* satisfying $\|Ax^* - b\| = \sigma$ is a KKT point of (1.1) if and only if there exists a $\mu \geq 0$ so that

$$0 = \mu[A^T(Ax^* - b)]_i + p\,\mathrm{sign}(x_i^*)\,|x_i^*|^{p-1}, \quad \forall i \in I,$$

with $I = \{i : \ x_i^* \neq 0\}$. This condition is further equivalent to

$$0 = \mu \mathrm{Diag}(x^*)A^T(Ax^* - b) + p|x^*|^p. \tag{4.8}$$

On the other hand, we recall from Lemma 2.5 (ii) that

$$\partial^\infty \Phi(x^*) = \{v : \ v_i = 0 \text{ for } i \in I\}.$$

Since $\mathcal{N}_S(x^*) = \{\mu A^T(Ax^* - b) : \ \mu \geq 0\}$, the constraint qualification (4.6) is equivalent to $[A^T(Ax^* - b)]_i$ being nonzero for some $i \in I$. From the definition of I, this constraint qualification can be equivalently formulated as

$$\mathrm{Diag}(x^*)A^T(Ax^* - b) \neq 0. \tag{4.9}$$

On passing, recall from Proposition 5.3.1 (i) and Remark 5.3.2 in [15, Section A] that we have

$$\mathcal{N}_{S_2}(x) = \mathcal{N}_{\|A\cdot - b\| \leq \sigma}(x) + \mathcal{N}_{B\cdot \leq h}(x)$$

at any $x \in S_2$, thanks to the existence of $x_0 \in S$ with $\|Ax_0 - b\| < \sigma$ by our blanket assumption. It is then not hard to see from the definitions that any first-order stationary point of (1.4) that lies in S is a KKT point of (1.1). Conversely, any KKT point of (1.1) is a first-order stationary point of (1.4) for some $\lambda > 0$.

Before ending this subsection, we comment on the magnitude of the nonzero entries of a first-order stationary point x^* of (1.4), assuming $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ for some continuous function ϕ. To facilitate comparison with existing work, we focus on the case where B is vacuous and $S_1 = \mathbb{R}^n$. Note that in this case, the definition of $F_\lambda(x)$ reduces to

$\lambda(\|Ax - b\|^2 - \sigma^2)_+ + \Phi(x)$. Then it follows from the stationarity of x^* and (4.1) that there exists $0 \le \nu \le 1$ so that at any i with $x_i^* \neq 0$, we have for some $\xi_i \in \partial\phi(x_i^*)$,

$$-\xi_i = 2\nu\lambda[A^T(Ax^* - b)]_i.$$

Let $x^\diamond$ be chosen so that $F_\lambda(x^*) \le F_\lambda(x^\diamond)$. Then for each i with $x_i \neq 0$,

$$\begin{aligned} |\xi_i| &\le 2\lambda\|A^T(Ax^* - b)\| \le 2\lambda\|A\|\|Ax^* - b\| \\ &\le 2\sqrt{\lambda}\|A\|\sqrt{(\lambda\|Ax^* - b\|^2 - \lambda\sigma^2)_+ + \lambda\sigma^2} \\ &\le 2\sqrt{\lambda}\|A\|\sqrt{F_\lambda(x^*) + \lambda\sigma^2} \le 2\sqrt{\lambda}\|A\|\sqrt{F_\lambda(x^\diamond) + \lambda\sigma^2}, \end{aligned} \tag{4.10}$$

where the fourth inequality follows from the nonnegativity of Φ, and the last inequality follows from the choice of $x^\diamond$. A concrete lower bound can be derived for some specific ϕ. For example, consider $\phi(t) = |t|^p$ for $p \in (0,1)$. Then we have from (4.10) that for $x_i^* \neq 0$,

$$p|x_i^*|^{p-1} \le 2\sqrt{\lambda}\|A\|\sqrt{F_\lambda(x^\diamond) + \lambda\sigma^2} \Longrightarrow |x_i^*| \ge \left(\frac{p}{2\sqrt{\lambda}\|A\|\sqrt{F_\lambda(x^\diamond) + \lambda\sigma^2}}\right)^{\frac{1}{1-p}} > 0. \tag{4.11}$$

Since local minimizers of (1.1) are local minimizers of (1.4) for some $\lambda^* > 0$ according to Theorem 3.2, and local minimizers of (1.4) are first-order stationary points of (1.4), the above discussion also gives a lower bound on the magnitude of the nonzero entries of the local minimizers of (1.1) when B is vacuous and $S_1 = \mathbb{R}^n$.

Remark 4.1. *In the recent paper [7], the authors derived a lower bound on the magnitudes of the nonzero entries of any first-order stationary point $\hat{x}$ of (1.3) with $H_\lambda(x) = \lambda\|Ax - b\|^2 + \|x\|_p^p$ for some $0 < p < 1$. Their lower bound is given by*

$$|\hat{x}_i| \ge \left(\frac{p}{2\sqrt{\lambda}\|A\|\sqrt{H_\lambda(\tilde{x})}}\right)^{\frac{1}{1-p}} > 0, \ \textit{for } \hat{x}_i \neq 0,$$

with $\tilde{x}$ chosen so that $H_\lambda(\hat{x}) \le H_\lambda(\tilde{x})$; see [7, Theorem 2.3]. This lower bound is similar to (4.11) except that $F_\lambda(x^\diamond) + \lambda\sigma^2$ is replaced by $H_\lambda(\tilde{x})$. Notice that when $x^\diamond = \tilde{x}$, we always have $F_\lambda(x^\diamond) + \lambda\sigma^2 \ge H_\lambda(x^\diamond)$, and these two values are the same if $\|Ax^\diamond - b\| \ge \sigma$. In particular, when $x^\diamond = \tilde{x} = 0$ and $\|b\| \ge \sigma$, the guaranteed lower bounds for both models are the same and is given by $\left(\frac{p}{2\lambda\|A\|\|b\|}\right)^{\frac{1}{1-p}}$. ■

4.2 Penalty method for solving (1.1)

In this subsection, we present details of our penalty method for solving (1.1). Before proceeding, we make the following assumption on Φ and S_1, which is standard in guaranteeing the sequence generated by an algorithm is bounded.

Assumption 4.1. *The function $\Phi + \delta_{S_1}$ has bounded level sets.*

Based on our previous discussions, an ϵ-minimizer of (1.1) can be obtained by finding a globally optimal solution of (1.4) with a sufficiently large λ. This approach is, however, not appealing because such λ may be hard to estimate, or it may be over-estimated and the resulting penalty problem becomes very ill-conditioned. Instead, it is natural to solve a sequence of problems in the form of (1.4) in which λ gradually increases. This scheme is commonly used in the classical penalty method. Also, notice that the first part of the objective of (1.4) is convex but nonsmooth. For an efficient implementation, we solve a sequence of partially smooth counterparts of (1.4) in the form of

$$\min_{x \in S_1} \; F_{\lambda,\mu}(x) := f_{\lambda,\mu}(x) + \Phi(x) \tag{4.12}$$

for some $\lambda, \mu > 0$, where

$$f_{\lambda,\mu}(x) := h_{\lambda,\mu}(\|Ax-b\|^2 - \sigma^2) + \sum_{i=1}^{\ell} h_{\lambda,\mu}([Bx-h]_i) \text{ with } \; h_{\lambda,\mu}(s) := \lambda \max_{0 \le t \le 1} \left\{ st - \frac{\mu}{2} t^2 \right\},$$

where the function $h_{\lambda,\mu}(\cdot)$ is a μ-smoothing for the function $s \to \lambda \cdot s_+$; see [26, Eq. 4] and the discussions therein.

It is not hard to show that for all $x \in \mathbb{R}^n$,

$$0 \;\le\; f_{\lambda,\mu}(x) \;\le\; \lambda[(\|Ax-b\|^2 - \sigma^2)_+ + \|(Bx-h)_+\|_1] \;\le\; f_{\lambda,\mu}(x) + \frac{\ell+1}{2}\lambda\mu, \tag{4.13}$$

and

$$\nabla f_{\lambda,\mu}(x) = 2h'_{\lambda,\mu}(\|Ax-b\|^2 - \sigma^2)A^T(Ax-b) + \sum_{i=1}^{\ell} h'_{\lambda,\mu}([Bx-h]_i) b_i, \tag{4.14}$$

where b_i is the column vector formed from the ith row of B, and the function $h'_{\lambda,\mu}$ satisfies

$$h'_{\lambda,\mu}(s) = \lambda \min\left\{ \max\left\{ \frac{s}{\mu}, 0 \right\}, 1 \right\}, \tag{4.15}$$

$$|h'_{\lambda,\mu}(s_1) - h'_{\lambda,\mu}(s_2)| \le \frac{\lambda}{\mu}|s_1 - s_2| \qquad \forall s_1,\; s_2 \in \mathbb{R}. \tag{4.16}$$

To solve (4.12), we consider an adaptation of the nonmonotone proximal gradient (NPG) method proposed in [28]. In [28], the NPG method was proposed to solve a class of unconstrained problems in the form of

$$\min_x f(x) + P(x), \tag{4.17}$$

where f and P are finite-valued functions in $\mathbb{R}^n$, and moreover, f is differentiable in $\mathbb{R}^n$ and its gradient is globally Lipschitz continuous in $\mathbb{R}^n$. The convergence analysis for the NPG method conducted in [28] relies on the global Lipschitz continuity of ∇f. Though the objective of (4.12) is in the same form as that of (4.17), we observe from (4.14) that $\nabla f_{\lambda,\mu}$ is locally but not globally Lipschitz continuous in $\mathbb{R}^n$. It thus appears that the NPG method [28] may not be applicable to our problem (4.12). We are, however, fortunately able to show in Appendix A that this NPG method is indeed capable of solving a more general class of problems that satisfies Assumption A.1. We next verify that Assumption A.1 holds for problem (4.12) with $f = f_{\lambda,\mu}$ and $P = \Phi + \delta_{S_1}$. As a consequence, the NPG method is applicable to our problem (4.12).

First, it is easy to see that Assumption A.1 (ii) holds. Let $x^0 \in S_1$ be arbitrarily chosen. It follows from (4.13) that $f_{\lambda,\mu}(x) \ge 0$, which implies that

$$\Omega(x^0) := \left\{ x \in S_1 : F_{\lambda,\mu}(x) \le F_{\lambda,\mu}(x^0) \right\} \subseteq \left\{ x \in S_1 : \Phi(x) \le F_{\lambda,\mu}(x^0) \right\}. \tag{4.18}$$

The set on the right hand side is nonempty and bounded by Assumption 4.1, and hence $\Omega(x^0)$ is nonempty and compact. Since $f_{\lambda,\mu} + \Phi$ is a continuous function, it is uniformly continuous and bounded below in $\Omega(x^0)$. Consequently, Assumption A.1 (iii) holds. One can also easily verify that Assumption A.1 (iv) holds using the compactness of $\Omega(x^0)$ and the nonnegativity of Φ. Finally, it is routine to show that $\nabla f_{\lambda,\mu}$ is locally Lipschitz continuous. This together with the compactness of $\Omega(x^0)$ shows that Assumption A.1 (i) also holds. Therefore, the NPG method can be suitably applied to solving problem (4.12).

We now establish a convergence result for the NPG method applied to problem (4.12).

Theorem 4.1. *Suppose that Assumption 4.1 holds. Given any $x^0 \in S_1$, let $\{x^k\}$ be the sequence generated by the NPG method applied to problem (4.12). There hold:*

(i) $\{x^k\}$ is bounded;

(ii) Any accumulation point x^* of $\{x^k\}$ is a first-order stationary point of problem (4.12), that is, it satisfies

$$0 \in \nabla f_{\lambda,\mu}(x^*) + \partial(\Phi + \delta_{S_1})(x^*). \tag{4.19}$$

Proof. (i) It follows from (4.18) and Proposition A.1 (i) with $f = f_{\lambda,\mu}$ and $P = \Phi + \delta_{S_1}$ that

$$\{x^k\} \subseteq \{x \in S_1 : \ F_{\lambda,\mu}(x) \le F_{\lambda,\mu}(x^0)\} \subseteq \{x \in S_1 : \ \Phi(x) \le F_{\lambda,\mu}(x^0)\}$$

and hence $\{x^k\}$ is bounded.

(ii) In view of Proposition A.1 (ii), $\bar{L}_k \le \tilde{L}$ for some $\tilde{L} > 0$ and all $k \ge 0$. It follows from (A.4) with $f = f_{\lambda,\mu}$ and $P = \Phi + \delta_{S_1}$, together with [27, Theorem 10.1] and [27, Exercise 10.10] that

$$0 \in \nabla f_{\lambda,\mu}(x^k) + \bar{L}_k(x^{k+1} - x^k) + \partial(\Phi + \delta_{S_1})(x^{k+1}).$$

Suppose that x^* is an accumulation point of $\{x^k\}$. Then there exists a subsequence $\mathcal{K}$ such that $\{x^k\}_{\mathcal{K}} \to x^*$. Upon taking limits as $k \in \mathcal{K} \to \infty$ on both sides of the above inclusion and using Theorem A.1 and (2.1), we see that (4.19) holds. ∎

We are now ready to present a penalty method for solving problem (1.1).

Penalty method for problem (1.1):

Let x^{feas} be an arbitrary feasible point of problem (1.1). Choose $x^0 \in S_1$, $\lambda_0 > 0$, $\mu_0 > 0$, $\epsilon_0 > 0$, $\rho > 1$ and $\theta \in (0, 1)$ arbitrarily. Set $k = 0$ and $x^{0,0} = x^0 \in S_1$.

1) If $F_{\lambda_k,\mu_k}(x^{k,0}) > F_{\lambda_k,\mu_k}(x^{\text{feas}})$, set $x^{k,0} = x^{\text{feas}}$. Apply the NPG method with $x^{k,0}$ as the initial point to find an approximate stationary point x^k to problem (4.12) with $\lambda = \lambda_k$ and $\mu = \mu_k$ satisfying

$$\text{dist}(0, \nabla f_{\lambda_k,\mu_k}(x^k) + \partial(\Phi + \delta_{S_1})(x^k)) \le \ \epsilon_k. \tag{4.20}$$

2) Set $\lambda_{k+1} = \rho\lambda_k$, $\mu_{k+1} = \theta\mu_k$, $\epsilon_{k+1} = \theta\epsilon_k$ and $x^{k+1,0} = x^k$.

3) Set $k \leftarrow k + 1$ and go to step 1).

end

Remark 4.2. *By virtue of Theorem 4.1, an x^k satisfying (4.20) can be found by the NPG method within a finite number of iterations. Therefore, the sequence $\{x^k\}$ is well defined.* ■

Convergence results for the above penalty method for solving problem (1.1) are presented in the next theorem. The arguments in the proof are standard and similar to the standard convergence analysis of the classical penalty methods, except that we make use of (i) the feasible point $x^{\rm feas}$ to guarantee that any limit point is feasible for (1.1); (ii) the constraint qualification (4.6) to guarantee the boundedness of "Lagrange multipliers". For completeness, we include the proof in Appendix B.

Theorem 4.2. *Suppose that Assumption 4.1 holds. Let $\{x^k\}$ be generated by the above penalty method for solving problem (1.1). There hold:*

(i) $\{x^k\}$ is bounded;

(ii) Any accumulation point x^* of $\{x^k\}$ is a feasible point of problem (1.1).

(iii) Suppose that $\{x^k\}_{\mathcal{K}} \to x^*$ for some subsequence $\mathcal{K}$ and that the constraint qualification (4.6) holds at x^*. Then x^* is a KKT point of problem (1.1).

5 Numerical simulations

In this section, we consider the problem of recovering a sparse solution of an underdetermined linear system from noisy measurements. In the literature, this is typically done via solving (1.2) or (1.3) with a specific sparsity inducing function Φ, e.g., the ℓ_1 norm or the $\ell_{1/2}$ quasi-norm; see, for example, [2, 3, 6, 7] and references therein. Here, we propose using the model (1.2) (a special case of (1.1) with $S_1 = \mathbb{R}^n$ and B being vacuous) with $\Phi(x) = \sum_{i=1}^n |x_i|^p$, $p = 1/2$. We solve this problem using our penalty method proposed in Subsection 4.2, which involves solving a sequence of subproblems in the form of (1.4). We benchmark our method against two other approaches:

1. the solver SPGL1 [2] (Version 1.8) that solves (1.2) with $\Phi(x) = \|x\|_1$;
2. the quadratic penalty method that solves (1.3) with $\Phi(x) = \sum_{i=1}^n |x_i|^{1/2}$ and some suitable $\lambda > 0$.

All codes are written in MATLAB, and the experiments were performed in MATLAB version R2014a on a cluster with 32 processors (2.9 GHz each) and 252G RAM.

For our penalty method, we set $x^0 = e$, the vector of all ones, $\lambda_0 = \mu_0 = \epsilon_0 = 1$, $\rho = 2$ and $\theta = 1/\rho$. We also set $x^{\rm feas} = A^\dagger b$, which we take as an input to the algorithm and does not count this computation in our CPU time below. For the NPG method for solving the unconstrained subproblem (4.12) at $\lambda = \lambda_k$ and $\mu = \mu_k$, we set $L_{\min} = 1$, $L_{\max} = 10^8$, $\tau = 2$, $c = 10^{-4}$, $M = 4$, $L_0^0 = 1$ and, for any $l \ge 1$,

$$L_l^0 := \min\left\{\max\left\{\frac{[x^{k,l} - x^{k,l-1}]^T[\nabla f_{\lambda_k,\mu_k}(x^{k,l}) - \nabla f_{\lambda_k,\mu_k}(x^{k,l-1})]}{\|x^{k,l} - x^{k,l-1}\|^2}, L_{\min}\right\}, L_{\max}\right\}.$$

The NPG method is terminated (at the lth inner iteration) when

$$\|{\rm Diag}(x^{k,l})\nabla f_{\lambda_k,\mu_k}(x^{k,l}) + p|x^{k,l}|^p\|_\infty \le \sqrt{\epsilon_k} \ \text{ and } \ \frac{|F_{\lambda_k,\mu_k}(x^{k,l}) - F_{\lambda_k,\mu_k}(x^{k,l-1})|}{\max\{1, |F_{\lambda_k,\mu_k}(x^{k,l})|\}} \le \min\{\epsilon_k^2, 10^{-4}\}.$$

Note that the first condition above means the first-order optimality condition (4.8) is approximately satisfied. The penalty method itself is terminated when

$$\max\left\{(\|Ax^k - b\|^2 - \sigma^2)_+, 0.01\epsilon_k\right\} \le 10^{-6},$$

with the ϵ_{k+1} in step 2) of the penalty method updated as $\max\{\theta\epsilon_k, 10^{-6}\}$ (instead of $\theta\epsilon_k$) in our implementation.

For the aforementioned SPGL1 [2], we use the default settings. For the quadratic penalty model (1.3), as discussed in our Example 3.1, there may be no $\lambda > 0$ so that the local minimizers of (1.3) are closely related to those of (1.2). However, one can observe as λ increases from 0 to ∞, the residual $\|A\tilde{x}(\lambda) - b\|$ changes from $\|b\|$ to 0, where $\tilde{x}(\lambda)$ is an optimal solution of (1.3). Thus, a possibly best approximate solution to (1.1) offered by model (1.3) appears to be the one corresponding to the least λ such that $\|A\tilde{x}(\lambda) - b\| \le \sigma$. However, such a λ is typically unknown. Instead, we solve a sequence of problem (1.3) along an increasing sequence of λ, and terminate when the approximate solution is approximately feasible for (1.2). Specifically, we apply the same scheme described in our penalty method but with H_λ in place of $F_{\lambda,\mu}$ and $\lambda\|Ax - b\|^2$ in place of $f_{\lambda,\mu}$, and we use exactly the same parameter settings as above. For ease of reference, we call this approach and our proposed penalty method as "Inexact Penalty" and "Exact Penalty" methods, respectively.

We consider randomly generated instances. First, we generate a matrix $\tilde{A} \in \mathrm{IR}^{K\times N}$ with i.i.d. standard Gaussian entries. The matrix A is then constructed so that its rows form an orthonormal basis for the row space of $\tilde{A}$. Next, we generate a vector $v \in \mathrm{IR}^T$ with i.i.d. standard Gaussian entries. We choose an index set I of size T at random and define a vector $\hat{x} \in \mathrm{IR}^N$ by setting $\hat{x}_I = v$ and $\hat{x}_{\bar{I}} = 0$. The measurement b is then set to be $A\hat{x} + \delta\xi$ for some $\delta > 0$, with each entry of ξ following again the standard Gaussian distribution. Finally, we set $\sigma = \delta\|\xi\|$ so that the resulting feasible set will contain the sparse vector $\hat{x}$.[2]

In our tests below, we set $(K, N, T) = (120i, 512i, 20i)$ for each $i = 12, 14, ..., 30$ and generate 10 random instances for each such (K, N, T) as described above. The computational results reported are averaged over the 10 instances, and they are reported in Tables 1, 2 and 3, which present results for $\delta = 10^{-2}$, 5×10^{-3} and 10^{-3}, respectively. For all three methods, we report the number of nonzero entries (**nnz**) in the approximate solution x obtained, computed using the MATLAB function `nnz`, the recovery error (**err**) $\|x - \hat{x}\|$, and the CPU time in seconds. We also report the function value $\Phi(x)$ at termination (**fval**) for the penalty methods. One can observe from these tables that our penalty method usually produces sparser solutions with smaller recovery errors than the other two approaches though it is in general slower than SPGL1. Moreover, in contrast with the method "Inexact Penalty", our penalty method achieves smaller objective values. These phenomena indeed reflect the intrinsic advantage of our approach.

6 Concluding remarks

Optimization models in finding sparse solutions to underdetermined systems of linear equations have stimulated development in signal processing and image sciences. The

[2]In our simulations, all random instances satisfy $\|b\| > \sigma$, which implies that the origin is excluded from the feasible region of the problem.

Table 1: Comparing the penalty method and SPGL1, $\delta = 10^{-2}$

Data			SPGL1			Inexact Penalty				Exact Penalty			
K	N	T	**nnz**	**err**	CPU	**fval**	**nnz**	**err**	CPU	**fval**	**nnz**	**err**	CPU
1440	6144	240	719	1.2e+00	0.69	2.89e+02	859	9.2e-01	15.27	1.90e+02	219	5.1e-01	5.08
1680	7168	280	837	1.3e+00	0.80	3.38e+02	998	1.0e+00	17.44	2.23e+02	257	5.5e-01	5.79
1920	8192	320	943	1.4e+00	1.06	3.87e+02	1139	1.1e+00	23.85	2.57e+02	294	5.7e-01	7.37
2160	9216	360	1050	1.5e+00	1.27	4.35e+02	1290	1.1e+00	28.91	2.87e+02	330	6.1e-01	10.37
2400	10240	400	1188	1.6e+00	1.53	4.82e+02	1430	1.2e+00	34.38	3.17e+02	366	6.6e-01	11.80
2640	11264	440	1266	1.6e+00	1.87	5.31e+02	1568	1.3e+00	43.91	3.49e+02	402	6.7e-01	13.98
2880	12288	480	1404	1.7e+00	2.20	5.78e+02	1712	1.3e+00	51.89	3.81e+02	439	7.0e-01	20.21
3120	13312	520	1500	1.7e+00	2.79	6.28e+02	1849	1.4e+00	64.28	4.15e+02	474	7.4e-01	21.67
3360	14336	560	1656	1.8e+00	2.92	6.75e+02	2000	1.4e+00	64.65	4.46e+02	514	7.7e-01	24.77
3600	15360	600	1755	1.9e+00	3.28	7.24e+02	2137	1.5e+00	75.72	4.78e+02	546	7.9e-01	25.12

Table 2: Comparing the penalty method and SPGL1, $\delta = 5 \times 10^{-3}$

Data			SPGL1			Inexact Penalty				Exact Penalty			
K	N	T	**nnz**	**err**	CPU	**fval**	**nnz**	**err**	CPU	**fval**	**nnz**	**err**	CPU
1440	6144	240	727	6.1e-01	0.78	2.54e+02	738	4.4e-01	10.40	1.94e+02	228	2.5e-01	4.68
1680	7168	280	827	6.7e-01	0.97	2.94e+02	865	4.9e-01	13.20	2.23e+02	266	2.7e-01	5.67
1920	8192	320	960	7.2e-01	1.31	3.39e+02	988	5.3e-01	18.56	2.57e+02	304	2.9e-01	7.93
2160	9216	360	1068	7.5e-01	1.58	3.83e+02	1104	5.5e-01	23.95	2.92e+02	342	3.0e-01	11.55
2400	10240	400	1195	7.9e-01	1.89	4.28e+02	1230	5.8e-01	29.73	3.26e+02	378	3.2e-01	11.47
2640	11264	440	1320	8.4e-01	2.35	4.66e+02	1352	6.1e-01	35.31	3.54e+02	416	3.5e-01	15.63
2880	12288	480	1422	8.7e-01	2.78	5.10e+02	1472	6.4e-01	40.89	3.88e+02	455	3.6e-01	16.76
3120	13312	520	1580	9.3e-01	3.23	5.54e+02	1600	6.7e-01	46.70	4.22e+02	496	3.7e-01	20.15
3360	14336	560	1668	9.5e-01	3.43	5.94e+02	1715	6.9e-01	52.10	4.53e+02	530	3.8e-01	24.81
3600	15360	600	1794	9.8e-01	3.89	6.40e+02	1841	7.2e-01	54.26	4.87e+02	570	3.9e-01	26.36

Table 3: Comparing the penalty method and SPGL1, $\delta = 10^{-3}$

Data			SPGL1			Inexact Penalty				Exact Penalty			
K	N	T	**nnz**	**err**	CPU	**fval**	**nnz**	**err**	CPU	**fval**	**nnz**	**err**	CPU
1440	6144	240	743	1.3e-01	1.24	2.02e+02	345	6.1e-02	5.63	1.95e+02	236	4.9e-02	6.49
1680	7168	280	880	1.4e-01	1.47	2.38e+02	396	6.5e-02	6.35	2.30e+02	275	5.5e-02	6.75
1920	8192	320	995	1.4e-01	1.93	2.74e+02	460	7.0e-02	8.21	2.64e+02	315	5.8e-02	8.84
2160	9216	360	1120	1.5e-01	2.08	3.08e+02	511	7.3e-02	9.36	2.97e+02	354	6.1e-02	11.23
2400	10240	400	1232	1.6e-01	2.59	3.41e+02	573	7.9e-02	11.51	3.28e+02	393	6.4e-02	13.60
2640	11264	440	1410	1.7e-01	2.96	3.73e+02	631	8.3e-02	13.78	3.59e+02	431	6.8e-02	17.26
2880	12288	480	1476	1.7e-01	3.71	4.08e+02	687	8.6e-02	15.82	3.93e+02	472	7.0e-02	18.00
3120	13312	520	1613	1.9e-01	4.13	4.42e+02	742	9.0e-02	18.29	4.26e+02	511	7.5e-02	23.66
3360	14336	560	1720	1.9e-01	4.81	4.78e+02	803	9.4e-02	21.97	4.61e+02	551	7.7e-02	28.99
3600	15360	600	1857	2.0e-01	5.17	5.07e+02	863	9.8e-02	24.26	4.87e+02	591	7.9e-02	27.44

constrained optimization model (1.2) and regularization model (1.3) have been widely used in this context when the data has noise. The existence of a regularization parameter λ such that problems (1.2) and (1.3) have a common global minimizer is known if the function Φ is convex. However, when Φ is nonconvex, such a λ does not always exist, as shown in Example 3.1. In this paper, we proposed a new penalty model (1.4) for the more general problem (1.1) where Φ can be nonconvex nonsmooth, perhaps even non-Lipschitz. We studied the existence of exact penalty parameters for (1.1) regarding local minimizers, stationary points and ϵ-minimizers. Moreover, we proposed a new penalty method which solves the constrained problem (1.1) by solving a sequence of (1.4) via the proximal gradient algorithm, with an update scheme for the penalty parameters. We also proved the convergence of the penalty method to a KKT point of (1.1). Preliminary numerical results showed that our penalty method is efficient for finding sparse solutions to underdetermined systems.

A Convergence of a nonmonotone proximal gradient method

In this appendix, we consider an algorithm for solving the following optimization problem

$$\min_{x} F(x) := f(x) + P(x), \tag{A.1}$$

where f and P satisfy the following assumptions:

Assumption A.1. *(i) f is continuously differentiable in $\mathcal{U}(x^0;\Delta)$ for some $x^0 \in \operatorname{dom} P := \{x : P(x) < \infty\}$ and $\Delta > 0$, and moreover, there exists some $L_f > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \le L_f \|x - y\|, \qquad \forall x, y \in \mathcal{U}(x^0;\Delta), \tag{A.2}$$

where

$$\begin{aligned} \mathcal{U}(x^0,\Delta) &:= \left\{x : \|x - z\| \le \Delta \text{ for some } z \in \Omega(x^0)\right\}, \\ \Omega(x^0) &:= \left\{x \in \mathbb{R}^n : \ F(x) \le F(x^0)\right\}. \end{aligned}$$

(ii) P is a proper lower semicontinuous function in $\mathbb{R}^n$.

(iii) F is bounded below and uniformly continuous in $\Omega(x^0)$.

(iv) The quantities A, B and C defined below are finite:

$$A := \sup_{x \in \Omega(x^0)} \|\nabla f(x)\|, \quad B := \sup_{x \in \Omega(x^0)} P(x), \quad C := \inf_{x \in \mathbb{R}^n} P(x). \tag{A.3}$$

The algorithm we consider is a nonmonotone proximal gradient method, presented as follows.

Algorithm 1: Nonmonotone proximal gradient (NPG) method for (A.1)

Let x^0 be given in Assumption A.1. Choose $L_{\max} \ge L_{\min} > 0$, $\tau > 1$, $c > 0$ and an integer $M \ge 0$ arbitrarily. Set $k = 0$.

1) Choose $L_k^0 \in [L_{\min}, L_{\max}]$ arbitrarily. Set $L_k = L_k^0$.

   1a) Solve the subproblem

$$u \in \operatorname{Arg}\min_{x} \left\{ \langle \nabla f(x^k), x - x^k\rangle + \frac{L_k}{2}\|x - x^k\|^2 + P(x) \right\}.^{3} \tag{A.4}$$

   1b) If

$$F(u) \le \max_{[k-M]_+ \le i \le k} F(x^i) - \frac{c}{2}\|u - x^k\|^2 \tag{A.5}$$

   is satisfied, then go to step 2).

[3] This problem has at least one optimal solution due to Assumption A.1 (ii) and (iv).

1c) Set $L_k \leftarrow \tau L_k$ and go to step 1a).

2) Set $x^{k+1} \leftarrow u$, $\bar{L}_k \leftarrow L_k$, $k \leftarrow k+1$ and go to step 1).

end

Although the NPG method has been analyzed in [28], the analysis there relies on the assumption that ∇f is globally Lipschitz continuous in $\mathbb{R}^n$. In our Assumption A.1, ∇f is, however, not necessarily globally Lipschitz continuous and thus the analysis in [28] does not apply directly to problem (A.1). We next show that the NPG method is still convergent for problem (A.1) under Assumption A.1.

Proposition A.1. *Let x^k be the approximate solution generated at the end of the kth iteration, and let*

$$\bar{L} := \max\{L_{\max}, \tau \underline{L}, \tau(L_f + c)\}, \qquad \underline{L} := \frac{2A\Delta + 2(B-C)}{\Delta^2}, \tag{A.6}$$

where A, B, C and Δ are given in Assumption A.1. Under Assumption A.1, there hold:

(i) x^{k+1} is well defined and $F(x^{k+1}) \le F(x^0)$ for all $k \ge 0$;

(ii) $\bar{L}_k$ is well defined and satisfies $\bar{L}_k \le \bar{L}$ for all $k \ge 0$.

(iii) For each $k \ge 0$, the inner termination criterion (A.5) is satisfied after at most

$$\left\lfloor \frac{\log(\bar{L}) - \log(L_{\min})}{\log \tau} + 1 \right\rfloor$$

inner iterations.

Proof. For convenience, whenever x^k is well defined with $F(x^k) \le F(x^0)$, set

$$x^{k+1}(L) \in \operatorname*{Arg\,min}_{x \in \mathbb{R}^n} \left\{ \langle \nabla f(x^k), x - x^k \rangle + \frac{L}{2}\|x - x^k\|^2 + P(x) \right\} \qquad \forall L > 0. \tag{A.7}$$

By (A.7), one can then observe that

$$\langle \nabla f(x^k), x^{k+1}(L) - x^k \rangle + P(x^{k+1}(L)) + \frac{L}{2}\|x^{k+1}(L) - x^k\|^2 \le P(x^k),$$

which along with (A.3) yields

$$\frac{L}{2}\|x^{k+1}(L) - x^k\|^2 - \|\nabla f(x^k)\|\|x^{k+1}(L) - x^k\| + C - P(x^k) \le 0.$$

Hence, we obtain that

$$\|x^{k+1}(L) - x^k\| \le \frac{\|\nabla f(x^k)\| + \sqrt{\|\nabla f(x^k)\|^2 + 2L(P(x^k) - C)}}{L}. \tag{A.8}$$

We now prove statements (i) and (ii) by induction. Indeed, for $k = 0$, we know that $x^0 \in \Omega(x^0)$. Using this relation, (A.3) and (A.8) with $k = 0$, one can have

$$\|x^1(L) - x^0\| \le \frac{A + \sqrt{A^2 + 2L(B-C)}}{L}.$$

In view of this inequality and (A.6), it is not hard to verify that

$$\|x^1(L) - x^0\| \le \Delta, \qquad \forall L \ge \underline{L}.$$

Using this relation and (A.2), we have

$$f(x^1(L)) \le f(x^0) + \langle \nabla f(x^0), x^1(L) - x^0 \rangle + \frac{L_f}{2}\|x^1(L) - x^0\|^2, \qquad \forall L \ge \underline{L}.$$

It follows from this relation and (A.7) that for all $L \ge \underline{L}$,

$$\begin{aligned}
&F(x^1(L)) = f(x^1(L)) + P(x^1(L)) \\
&\le f(x^0) + \langle \nabla f(x^0), x^1(L) - x^0 \rangle + \tfrac{L_f}{2}\|x^1(L) - x^0\|^2 + P(x^1(L)) \\
&= f(x^0) + \langle \nabla f(x^0), x^1(L) - x^0 \rangle + \tfrac{L}{2}\|x^1(L) - x^0\|^2 + P(x^1(L)) + \tfrac{L_f - L}{2}\|x^1(L) - x^0\|^2 \\
&\le f(x^0) + P(x^0) + \tfrac{L_f - L}{2}\|x^1(L) - x^0\|^2 \ = \ F(x^0) + \tfrac{L_f - L}{2}\|x^1(L) - x^0\|^2,
\end{aligned}$$

where the second inequality follows from (A.7). Using this relation, one can immediately observe that

$$F(x^1(L)) \le F(x^0) - \frac{c}{2}\|x^1(L) - x^0\|^2, \qquad \forall L \ge \hat{L}, \tag{A.9}$$

where

$$\hat{L} := \max\{\underline{L}, L_f + c\}.$$

This shows that (A.5) must be satisfied after finitely many inner iterations. Moreover, from the definition of $\bar{L}_0$, we must have either $\bar{L}_0 = L_0^0$ or $\bar{L}_0/\tau < \hat{L}$. This together with $L_0^0 \le L_{\max}$ implies $\bar{L}_0 \le \max\{L_{\max}, \tau\hat{L}\}$, and hence statement (ii) holds for $k = 0$. We also see from (A.9) that $F(x^1) = F(x^1(\bar{L}_0)) \le F(x^0)$. Hence, statement (i) also holds for $k = 0$.

We now suppose that statements (i) and (ii) hold for all $k \le K$ for some $K \ge 0$. It remains to show that they also hold for $k = K+1$. Indeed, using the induction hypothesis, we have $x^{K+1} \in \Omega(x^0)$. In view of this relation and a similar argument as for $k = 0$, one can show that statement (ii) holds for $k = K+1$. By the induction hypothesis, we know that $F(x^{k+1}) \le F(x^0)$ for all $k \le K$. Using this relation and (A.5) with $k = K+1$, one can conclude that $F(x^{K+2}) \le F(x^0)$ and hence statement (i) holds for $k = K+1$. This completes the induction.

Finally we prove statement (iii). Let n_k denote the total number of inner iterations executed at the kth outer iteration. One can observe that

$$L_{\min}\tau^{n_k - 1} \le L_k^0 \tau^{n_k - 1} = \bar{L}_k.$$

The conclusion then immediately follows from this relation and statement (ii). ■

We end our discussion with a convergence result for the NPG method, which can be proved similarly as in [28, Lemma 4].

Theorem A.1. *Let x^k be the approximate solution generated at the end of the kth iteration. Under Assumption A.1, there holds $\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$.*

B Proof of Theorem 4.2

In this section, we present the proof of Theorem 4.2.

Proof. (i) By Proposition A.1, we know that $F_{\lambda_k,\mu_k}(x^k) \le F_{\lambda_k,\mu_k}(x^{k,0})$. In addition, from step 1) of the above penalty method, one has $F_{\lambda_k,\mu_k}(x^{k,0}) \le F_{\lambda_k,\mu_k}(x^{\rm feas})$. It then follows that $F_{\lambda_k,\mu_k}(x^k) \le F_{\lambda_k,\mu_k}(x^{\rm feas})$. Using this relation along with (4.13) and the facts that $\|Ax^{\rm feas} - b\| \le \sigma$ and $Bx^{\rm feas} \le h$, one can have

$$\Phi(x^k) \ \le\ F_{\lambda_k,\mu_k}(x^k) \ \le\ F_{\lambda_k,\mu_k}(x^{\rm feas}) = \Phi(x^{\rm feas}).$$

Moreover, we also have $x^k \in S_1$ from the definition. Hence, $\{x^k\}$ is bounded since $\Phi + \delta_{S_1}$ has bounded level sets.

(ii) Let x^* be an accumulation point of $\{x^k\}$. Then there exists a subsequence $\{x^k\}_{\mathcal{K}} \to x^*$. Using $F_{\lambda_k,\mu_k}(x^k) \le F_{\lambda_k,\mu_k}(x^{\rm feas})$, (4.13) and the definition of $F_{\lambda,\mu}$, we have

$$\begin{aligned}
\lambda_k(\|Ax^k - b\|^2 - \sigma^2)_+ \ &+\ \lambda_k\|(Bx^k - h)_+\|_1 \le f_{\lambda_k,\mu_k}(x^k) + \tfrac{\ell+1}{2}\lambda_k\mu_k \\
&\le\ F_{\lambda_k,\mu_k}(x^k) + \tfrac{\ell+1}{2}\lambda_k\mu_k \le F_{\lambda_k,\mu_k}(x^{\rm feas}) + \tfrac{\ell+1}{2}\lambda_k\mu_k \\
&=\ \Phi(x^{\rm feas}) + \tfrac{\ell+1}{2}\lambda_k\mu_k.
\end{aligned}$$

It then follows that

$$(\|Ax^k - b\|^2 - \sigma^2)_+ + \|(Bx^k - h)_+\|_1 \le \frac{\Phi(x^{\rm feas})}{\lambda_k} + \frac{\ell+1}{2}\mu_k.$$

Taking limits on both sides of this inequality as $k \in \mathcal{K} \to \infty$, one has $(\|Ax^* - b\|^2 - \sigma^2)_+ \le 0$ and $\|(Bx^* - h)_+\|_1 \le 0$. Hence x^* is a feasible point of problem (1.1).

(iii) Let $I_* := \{i : (Bx^* - h)_i = 0\}$. Then $(Bx^*)_i < h_i$ for all $i \notin I_*$ and we have

$$\mathcal{N}_{B\cdot\le h}(x^*) = \left\{\sum_{i\in I_*} y_i b_i : \ y \ge 0\right\},$$

where b_i denotes the column vector formed from the ith row of B. Moreover, for all sufficiently large $k \in \mathcal{K}$, we have $(Bx^k)_i < h_i$ for all $i \notin I_*$. Using this and (4.15), we have $w_i^k := h'_{\lambda_k,\mu_k}([Bx^k - h]_i) = 0$ for $i \notin I_*$ and all sufficiently large k. This together with (4.20) and (4.14) implies that for all $k \in \mathcal{K}$ sufficiently large, there exists $\xi^k \in \partial(\Phi + \delta_{S_1})(x^k)$ so that

$$\left\| 2h'_{\lambda_k,\mu_k}(\|Ax^k - b\|^2 - \sigma^2)A^T(Ax^k - b) + \xi^k + \sum_{i\in I_*} w_i^k b_i \right\| \le\ \epsilon_k. \tag{B.1}$$

We consider two different cases: $\|Ax^* - b\| < \sigma$ or $\|Ax^* - b\| = \sigma$.

Case 1. Suppose first that x^* satisfies $\|Ax^* - b\| < \sigma$. Then $\|Ax^k - b\| < \sigma$ for all sufficiently large $k \in \mathcal{K}$. Using this relation and (4.15), we have $h'_{\lambda_k,\mu_k}(\|Ax^k - b\|^2 - \sigma^2) = 0$ for all sufficiently large $k \in \mathcal{K}$. Hence, the relation (B.1) reduces to

$$\left\| \xi^k + \sum_{i\in I_*} w_i^k b_i \right\| \le\ \epsilon_k. \tag{B.2}$$

We suppose to the contrary that $\|\xi^k\|$ is unbounded. Without loss of generality, assume that $\{\|\xi^k\|\}_{\mathcal{K}} \to \infty$ and that $\lim\limits_{k\in\mathcal{K}} \frac{\xi^k}{\|\xi^k\|} = \xi^*$ for some ξ^*. Divide both sides of (B.2) by $\|\xi^k\|$

and pass to the limit, making use of $\epsilon_k \to 0$, (2.1) and the closeness of the conical hull of the finite set $\{b_i : i \in I_*\}$, we see further that $\xi^* \in \partial^\infty(\Phi + \delta_{S_1})(x^*)$ and

$$-\xi^* \in \left\{ \sum_{i \in I_*} y_i b_i : y \geq 0 \right\} = \mathcal{N}_{B \cdot \leq h}(x^*) = \mathcal{N}_{S_2}(x^*),$$

where the second equality follows from the fact that $\|Ax^* - b\| < \sigma$. Since $\|\xi^*\| = 1$, this is a contradiction to (4.6). This shows that $\|\xi^k\|$ is bounded. By passing to the limit along a convergent subsequence in (B.2), using (2.1) and the closedness of finitely generated cones, we obtain

$$0 \in \partial(\Phi + \delta_{S_1})(x^*) + \left\{ \sum_{i \in I_*} y_i b_i : y \geq 0 \right\} = \partial(\Phi + \delta_{S_1})(x^*) + \mathcal{N}_{S_2}(x^*),$$

i.e., x^* is a KKT point of (1.1).

Case 2. Suppose now that x^* satisfies $\|Ax^* - b\| = \sigma$. Observe from (4.15) that $h'_{\lambda_k, \mu_k}(\|Ax^k - b\|^2 - \sigma^2) \geq 0$ for all k. Let $t_k := 2h'_{\lambda_k, \mu_k}(\|Ax^k - b\|^2 - \sigma^2)$ for notational simplicity, and suppose for contradiction that the sequence $\{\|\xi^k\|\}_\mathcal{K}$ is unbounded. Without loss of generality, assume that $\{\|\xi^k\|\}_\mathcal{K} \to \infty$. It follows from (B.1) that

$$\left\| \frac{t_k}{\|\xi^k\|} A^T(Ax^k - b) + \frac{1}{\|\xi^k\|}\xi^k + \sum_{i \in I_*} \frac{w_i^k}{\|\xi^k\|} b_i \right\| \leq \frac{\epsilon_k}{\|\xi^k\|}. \tag{B.3}$$

We claim that $\{\frac{t_k}{\|\xi^k\|}\}_\mathcal{K}$ is bounded. Suppose to the contrary and without loss of generality that $\{\frac{t_k}{\|\xi^k\|}\}_\mathcal{K} \to \infty$. Dividing both sides of (B.3) by $\frac{t_k}{\|\xi^k\|}$, passing to the limit and using the closedness of finitely generated cones, we see that

$$0 \in A^T(Ax^* - b) + \mathcal{N}_{B \cdot \leq h}(x^*). \tag{B.4}$$

This means that x^* is an optimal solution of the problem

$$\begin{array}{ll} \min\limits_x & \frac{1}{2}\|Ax - b\|^2 \\ \text{s.t.} & Bx \leq h. \end{array}$$

Since $\|Ax^* - b\| = \sigma$, this contradicts our assumption that there is $x_0 \in S$ with $\|Ax_0 - b\| < \sigma$. This contradiction shows that $\{\frac{t_k}{\|\xi^k\|}\}_\mathcal{K}$ is bounded. By passing to a further subsequence if necessary, we may now assume without loss of generality that

$$\lim_{k \in \mathcal{K}} \frac{t_k}{\|\xi^k\|} = t_*, \quad \text{and} \quad \lim_{k \in \mathcal{K}} \frac{\xi^k}{\|\xi^k\|} = \xi^*.$$

Note that $\xi^* \in \partial^\infty(\Phi + \delta_{S_1})(x^*)$ due to (2.1). Taking limit on both sides of (B.3) along this subsequence and making use again of the closedness of finitely generated cones, we see further that

$$-\xi^* \in t_* A^T(Ax^* - b) + \left\{ \sum_{i \in I_*} y_i b_i : y \geq 0 \right\} \subseteq \mathcal{N}_{\|A \cdot - b\| \leq \sigma}(x^*) + \mathcal{N}_{B \cdot \leq h}(x^*) = \mathcal{N}_{S_2}(x^*), \tag{B.5}$$

where the set inclusion follows from the fact that $\|Ax^* - b\| = \sigma$ and the existence of $x_0 \in S$ with $\|Ax_0 - b\| < \sigma$; this latter condition also gives the last equality in (B.5). Since $\|\xi^*\| = 1$, the relation (B.5) together with $\xi^* \in \partial^\infty(\Phi + \delta_{S_1})(x^*)$ contradicts (4.6). Thus, the sequence $\{\|\xi^k\|\}_\mathcal{K}$ is bounded.

Next, we claim that $\{t_k\}_\mathcal{K}$ is bounded. Assume again to the contrary that $\{t_k\}_\mathcal{K}$ is unbounded and assume without loss of generality that $\{t_k\}_\mathcal{K} \to \infty$. From (B.1), we have

$$\left\| A^T(Ax^k - b) + \frac{1}{t_k}\xi^k + \sum_{i \in I_*} \frac{w_i^k}{t_k} b_i \right\| \le \frac{\epsilon_k}{t_k}. \tag{B.6}$$

Passing to the limit in (B.6) and using the boundedness of ξ^k as well as the closedness of finitely generated cones, we arrive at (B.4). A contradiction can then be derived similarly as before. Thus, we conclude that $\{t_k\}_\mathcal{K}$ is bounded.

Let π^* be an accumulation point of $\{t_k\}_\mathcal{K}$. Without loss of generality, assume that $\{t_k\}_\mathcal{K} \to \pi^*$. Since $t_k \ge 0$ for all k, one has $\pi^* \ge 0$. Taking limits on both sides of (B.1) as $k \in \mathcal{K} \to \infty$, invoking (2.1), the boundedness of $\{\xi^k\}_{k \in \mathcal{K}}$ and the closedness of finitely generated cones, one can see that

$$0 \in \pi^* A^T(Ax^* - b) + \partial(\Phi + \delta_{S_1})(x^*) + \mathcal{N}_{B \cdot \le h}(x^*) \subseteq \partial(\Phi + \delta_{S_1})(x^*) + \mathcal{N}_{S_2}(x^*).$$

This shows that x^* is a KKT point of (1.1). ■

References

[1] A. Beck and M. Teboulle, A fast gradient-based algorithms for constrained total variation image denosing and deblurring problems, *IEEE Trans. Image Process.* 18, pp. 2419–2434 (2009).

[2] E. van den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.* 31, pp. 890–912 (2008).

[3] A. M. Bruckstein, D. L. Donoho and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.* 51, pp. 34–81 (2009).

[4] R. H. Chan, M. Tao and X. M. Yuan. Constrained total variation deblurring models and fast algorithms based on alternating direction method of multipliers. *SIAM J. Imaging Sci.* 6, pp. 680–697 (2013).

[5] X. Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.* Ser. B 134, pp. 71–99 (2012).

[6] X. Chen, D. Ge, Z. Wang and Y. Ye. Complexity of unconstrained L_2-L_p minimization. *Math. Program.* 143, pp. 371–383 (2014).

[7] X. Chen, F. Xu and Y. Ye. Lower bound theory of nonzero entries in solutions of l_2-l_p minimization. *SIAM J. Sci. Comput.* 32, pp. 2832–2852 (2010).

[8] F. Fachinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems. I and II.* Springer (2003).

[9] J. Fan. Comments on "Wavelets in Statistic: A review" by A. Antoniadis. *Stat. Method. Appl.* 6, pp. 131–138 (1997).

[10] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 96, pp. 1348–1360 (2001).

[11] M. P. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM J. Optim.* 18, pp. 1326–1350 (2007).

[12] D. Ge, X. Jiang and Y. Ye A note on the complexity of L_p minimization. *Math. Program.* 21, pp. 1721–1739 (2011).

[13] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, pp. 357–383 (1992).

[14] P. Gong, C. Zhang, Z. Lu, J. Huang and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. The 30th International Conference on Machine Learning (ICML 2013).

[15] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis.* Springer, (2001).

[16] J. Huang, J. L. Horowitz and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Stat.* 36, pp. 587–613 (2008).

[17] K. Knight and W. J. Fu. Asymptotics for lasso-type estimators. *Ann. Stat.* 28, pp. 1356–1378 (2000).

[18] H. A. Le Thi, T. P. Dinh and H. V. Ngai. Exact penalty and error bounds in DC programming. *J. Global Optim.* 52, pp. 509–535 (2012).

[19] C. Li, K. F. Ng and T. K. Pong. The SECQ, linear regularity, and the strong CHIP for an infinite system of closed convex sets in normed linear spaces. *SIAM J. Optim.* 18, pp. 643–665 (2007).

[20] G. Li, A. K. C. Ma and T. K. Pong. Robust least square semidefinite programming with applications. *Comput. Optim. & Appl.* 58, pp. 347–379 (2014).

[21] Z. Lu. Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. *Math. Program.* 147, pp. 277–307 (2014).

[22] X.-D. Luo and Z.-Q. Luo. Extension of Hoffman's error bound to polynomial systems. *SIAM J. Optim.* 4, pp. 383–392 (1994).

[23] M. Nikolova, M. K. Ng, S. Zhang and W. Ching. Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.* 1, pp. 2–25 (2008).

[24] M. Ng, P. Weiss and X. Yuan. Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods. *SIAM J. Sci. Comput.* 32, pp. 2710–2736 (2010).

[25] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, 2nd, New York (2006).

[26] M. Ç. Pinar and S. A. Zenios. On smoothing exact penalty functions for convex constrained optimization. *SIAM J. Optim.* 4, pp. 486–511 (1994).

[27] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer (1998).

[28] S. J. Wright, R. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE T. Signal Proces.* 57, pp. 2479–2493 (2009).

[29] P. Yin, Y. Lou, Q. He and J. Xin. Minimization of ℓ_{1-2} for compressed sensing. *SIAM J. Sci. Comput.* 37, pp. A536–A563 (2015).

[30] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* 38, pp. 894–942 (2010).