Mathematical Programming (2019) 176:339–367
https://doi.org/10.1007/s10107-018-1327-8

FULL LENGTH PAPER

Series B

A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems

Tianxiang Liu[1] · Ting Kei Pong[1] · Akiko Takeda[2,3]

Received: 16 October 2017 / Accepted: 31 August 2018 / Published online: 8 September 2018

Abstract

We consider a class of nonconvex nonsmooth optimization problems whose objective is the sum of a smooth function and a finite number of nonnegative proper closed possibly nonsmooth functions (whose proximal mappings are easy to compute), some of which are further composed with linear maps. This kind of problems arises naturally in various applications when different regularizers are introduced for inducing simultaneous structures in the solutions. Solving these problems, however, can be challenging because of the coupled nonsmooth functions: the corresponding proximal mapping can be hard to compute so that standard first-order methods such as the proximal gradient algorithm cannot be applied efficiently. In this paper, we propose a successive difference-of-convex approximation method for solving this kind of problems. In this algorithm, we approximate the nonsmooth functions by their Moreau envelopes in each iteration. Making use of the simple observation that Moreau envelopes of nonnegative proper closed functions are continuous *difference-of-convex* functions, we can then approximately minimize the approximation function by first-order methods with suitable majorization techniques. These first-order methods can be implemented efficiently thanks to the fact that the proximal mapping of *each* nonsmooth function is easy to compute. Under suitable assumptions, we prove that the sequence generated by our method is bounded and any accumulation point is a stationary point of the objective. We also discuss how our method can be applied to concrete applications such as nonconvex fused regularized optimization problems and simultaneously structured matrix optimization problems, and illustrate the performance numerically for these two specific applications.

Keywords Moreau envelope · Difference-of-convex approximation · Proximal mapping · Simultaneous structures

Ting Kei Pong is supported in part by Hong Kong Research Grants Council PolyU153085/16p. Akiko Takeda is supported by Grant-in-Aid for Scientific Research (C), 15K00031.

✉ Akiko Takeda
takeda@mist.i.u-tokyo.ac.jp; akiko.takeda@riken.jp

Extended author information available on the last page of the article

Mathematics Subject Classification 90C30 · 65K05 · 90C26

1 Introduction

In this paper, we consider the following possibly nonconvex nonsmooth optimization problem:

$$\underset{\boldsymbol{x}}{\text{minimize}}\ F(\boldsymbol{x}) := f(\boldsymbol{x}) + P_0(\boldsymbol{x}) + \sum_{i=1}^{m} P_i(\mathcal{A}_i\boldsymbol{x}), \tag{1}$$

with the objective satisfying the following assumptions (see the next section for notation and definitions):

A1. $f : \mathbb{R}^n \to \mathbb{R}$ is an L-smooth function i.e., there exists a constant $L > 0$ so that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{v})\| \le L\|\boldsymbol{x} - \boldsymbol{v}\|$$

for any $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^n$.

A2. $\mathcal{A}_i : \mathbb{R}^n \to \mathbb{R}^{n_i}$, $i = 1, \ldots, m$, are linear mappings and $P_i : \mathbb{R}^{n_i} \to \mathbb{R}_+ \cup \{\infty\}$, $i = 0, \ldots, m$, are proper closed functions. The functions P_i, $i = 0, \ldots, m$, are continuous in their respective domains, and

$$\text{dom } P_0 \cap \bigcap_{i=1}^{m} \mathcal{A}_i^{-1}(\text{dom } P_i) \neq \emptyset.$$

Moreover, the proximal mapping of γP_i is easy to compute for every $\gamma > 0$ and for each $i = 0, \ldots, m$. The sets dom P_i, $i = 1, \ldots, m$, are closed.

A3. The function $f + P_0$ is level-bounded, i.e., for each $r \in \mathbb{R}$, the set $\{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) + P_0(\boldsymbol{x}) \le r\}$ is bounded.

Problem (1) arises in many contemporary applications such as structured low rank matrix recovery problems (see, for example, [18]), nonconvex fused regularized optimization problems (see, for example, [21] and Example 2 in Sect. 4) and simultaneously structured matrix optimization problems (see, for example, [23] and Example 5 in Sect. 4). In these applications, the P_i's are used for inducing desirable structures in the solutions and they are typically functions whose proximal mappings are easy to compute. If only one such function appears in (1), i.e., $m = 0$, then some standard first-order methods such as the proximal gradient algorithm or its variants can be applied to solving (1) efficiently, because these algorithms only require the computation of ∇f and the proximal mapping of γP_0 ($\gamma > 0$) in each iteration. However, in all the aforementioned applications, there are always more than one such structure-inducing functions in (1) (i.e., $m \ge 1$) and the $\mathcal{A}_i$'s might not always be identity mappings. Then the proximal gradient algorithm and its variants cannot be applied efficiently, because the proximal mapping of $\boldsymbol{x} \mapsto P_0(\boldsymbol{x}) + \sum_{i=1}^{m} P_i(\mathcal{A}_i\boldsymbol{x})$ can be hard to compute in general.

When the function f and the P_i's are all convex functions, one alternative approach for solving (1) is the alternating direction method of multipliers (ADMM); see, for

example, [9,10]. This method can be applied to (1) by suitably introducing slack variables that transform the problem into a linearly constrained problem, and each iteration only requires computing the proximal mappings of f and γP_i's, as well as an update of an auxiliary (dual) variable. However, it is known that the ADMM does not necessarily converge if the P_i's are nonconvex and $m \geq 1$; see, for example, [13, Example 7]. In the case when P_i's are nonconvex but globally Lipschitz for $i = 0, \ldots, m$, and $\mathcal{A}_i$ is the identity mapping for all i, a new method for solving (1) was introduced in a series of work [32,33]. Their method is based on the so-called proximal average of P_i's, and each iteration involves only the computations of ∇f and the proximal mappings of γP_i's. However, it was only shown that any accumulation point of the sequence generated by their method is a stationary point of a certain smooth approximation of (1). Moreover, their method was designed for the case when P_i's are globally Lipschitz, and the convergence behavior of their method is unknown when some non-Lipschitz functions such as the ℓ_p quasi-norm or the indicator function of some closed sets (such as the set of all k-sparse vectors) are present in (1).

In this paper, we propose a new method for solving (1) that is ready to take advantage of the ease of proximal mapping computations and has convergence guarantee under suitable assumptions, without imposing convexity nor globally Lipschitz continuity on P_i's. We call our method the successive difference-of-convex approximation method (SDCAM). In this method, we construct an approximation to the objective of (1) in each iteration using the Moreau envelopes of the $\lambda_{i,t} P_i$, $i = 1, \ldots, m$, where t is the number of iteration and $\{\lambda_{i,t}\}$ are nonincreasing positive sequences satisfying $\lim_{t\to\infty} \lambda_{i,t} = 0$; a suitable approximate stationary point of this approximation function is then taken to be the next iterate $\boldsymbol{x}^{t+1}$ of our algorithm. The point $\boldsymbol{x}^{t+1}$ can be found efficiently by recalling that the Moreau envelopes involved, despite being nonsmooth in general due to the possible nonconvexity of the P_i's, are continuous difference-of-convex functions. Thus, one can incorporate majorization techniques in some standard first-order methods such as the proximal gradient algorithm for finding $\boldsymbol{x}^{t+1}$ in each iteration. Moreover, when such first-order methods are applied, the main computational cost per inner iteration typically only depends on the computations of ∇f and the proximal mappings of γP_i, $i = 0, \ldots, m$, $\gamma > 0$, which are inexpensive in many applications. This suggests that the SDCAM can be applied efficiently for solving (1). More details of this algorithm will be discussed in Sect. 3, where we also prove that the sequence $\{\boldsymbol{x}^t\}$ generated is bounded and any accumulation point is a stationary point of (1) under suitable assumptions.

The rest of the paper is organized as follows. In Sect. 2, we introduce notation and some preliminary results. Our SDCAM is presented and its convergence is analyzed under suitable assumptions in Sect. 3. We then discuss how our method can be applied to various kinds of structured optimization problems including some nonconvex fused regularized optimization problems, some simultaneously sparse and low rank matrix optimization problems, and the low rank nearest correlation matrix problem, in Sect. 4. We also perform numerical experiments on some of these applications to demonstrate the efficiency of our algorithm in Sect. 5. Finally, we present some concluding remarks in Sect. 6.

2 Notation and preliminaries

In this paper, vectors and matrices are represented in bold lower case letters and upper case letters, respectively. The inner product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathbb{R}^n$ are denoted by $\boldsymbol{a}^\top \boldsymbol{b}$ or $\boldsymbol{b}^\top \boldsymbol{a}$, and we use $\|\boldsymbol{a}\|_0$, $\|\boldsymbol{a}\|_1$ and $\|\boldsymbol{a}\|$ to denote the number of nonzero entries, the ℓ_1 norm and the ℓ_2 norm of $\boldsymbol{a}$, respectively. Moreover, we use $\mathrm{Diag}(\boldsymbol{a})$ to denote the diagonal matrix whose diagonal is $\boldsymbol{a}$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times n}$, their Hadamard (entrywise) product is denoted by $\boldsymbol{A} \circ \boldsymbol{B}$. We also use $\|\boldsymbol{A}\|_*$ and $\|\boldsymbol{A}\|_F$ to denote the nuclear norm and the Fröbenius norm of $\boldsymbol{A}$, respectively, and let $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{R}^{mn\times 1}$ denote the vectorization of $\boldsymbol{A}$, which is obtained by stacking the columns of $\boldsymbol{A}$ on top of one another. Furthermore, we use $\sigma_{\max}(\boldsymbol{A})$ to denote the largest singular value of $\boldsymbol{A}$. The space of symmetric $n \times n$ matrices is denoted by $\mathcal{S}^n$. For a matrix $\boldsymbol{X} \in \mathcal{S}^n$, we use $\mathrm{diag}(\boldsymbol{X}) \in \mathbb{R}^n$ to denote its diagonal and $\lambda_{\max}(\boldsymbol{X})$ to denote its largest eigenvalue. We write $\boldsymbol{X} \succeq 0$ if $\boldsymbol{X}$ is positive semidefinite. For a linear operator $\mathcal{A}$, we let $\mathcal{A}^*$ denote its adjoint.

A function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is said to be proper if $\mathrm{dom}\, h := \{\boldsymbol{x} : h(\boldsymbol{x}) < \infty\} \neq \emptyset$. Such a function is said to be closed if it is lower semicontinuous. Following [25, Definition 8.3], for a proper function h, the limiting and horizon subdifferentials at $\boldsymbol{x} \in \mathrm{dom}\, h$ are defined respectively as

$$\partial h(\boldsymbol{x}) = \left\{\boldsymbol{u} : \exists \boldsymbol{u}^t \to \boldsymbol{u}, \boldsymbol{x}^t \xrightarrow{h} \boldsymbol{x} \text{ with } \boldsymbol{u}^t \in \hat{\partial} h(\boldsymbol{x}^t) \text{ for each } t\right\},$$
$$\partial^\infty h(\boldsymbol{x}) = \left\{\boldsymbol{u} : \exists \alpha_t \downarrow 0, \alpha_t \boldsymbol{u}^t \to \boldsymbol{u}, \boldsymbol{x}^t \xrightarrow{h} \boldsymbol{x} \text{ with } \boldsymbol{u}^t \in \hat{\partial} h(\boldsymbol{x}^t) \text{ for each } t\right\},$$

where $\hat{\partial} h(\boldsymbol{w}) := \left\{\boldsymbol{u} : \liminf_{\boldsymbol{y}\to\boldsymbol{w}, \boldsymbol{y}\neq\boldsymbol{w}} \frac{h(\boldsymbol{y})-h(\boldsymbol{w})-\boldsymbol{u}^\top(\boldsymbol{y}-\boldsymbol{w})}{\|\boldsymbol{y}-\boldsymbol{w}\|} \geq 0\right\}$, and the notation $\boldsymbol{x}^t \xrightarrow{h} \boldsymbol{x}$ means $\boldsymbol{x}^t \to \boldsymbol{x}$ and $h(\boldsymbol{x}^t) \to h(\boldsymbol{x})$. We also define $\partial h(\boldsymbol{x}) = \partial^\infty h(\boldsymbol{x}) := \emptyset$ when $\boldsymbol{x} \notin \mathrm{dom}\, h$. It is easy to show that at any $\boldsymbol{x} \in \mathrm{dom}\, h$, the limiting and horizon subdifferentials have the following robustness property:

$$\begin{aligned} \left\{\boldsymbol{u} : \exists \boldsymbol{u}^t \to \boldsymbol{u}, \boldsymbol{x}^t \xrightarrow{h} \boldsymbol{x} \text{ with } \boldsymbol{u}^t \in \partial h(\boldsymbol{x}^t) \text{ for each } t\right\} &\subseteq \partial h(\boldsymbol{x}), \\ \left\{\boldsymbol{u} : \exists \alpha_t \downarrow 0, \alpha_t \boldsymbol{u}^t \to \boldsymbol{u}, \boldsymbol{x}^t \xrightarrow{h} \boldsymbol{x} \text{ with } \boldsymbol{u}^t \in \partial h(\boldsymbol{x}^t) \text{ for each } t\right\} &\subseteq \partial^\infty h(\boldsymbol{x}). \end{aligned} \tag{2}$$

The limiting subdifferential at $\boldsymbol{x}$ reduces to $\{\nabla h(\boldsymbol{x})\}$ if h is continuously differentiable at $\boldsymbol{x}$ [25, Exercise 8.8(b)], and reduces to the convex subdifferential if h is proper convex [25, Proposition 8.12].

For a proper closed function h with $\inf h > -\infty$, we will also need its Moreau envelope for any given $\lambda > 0$, which is defined as

$$e_\lambda h(\boldsymbol{x}) := \inf_{\boldsymbol{y}} \left\{\frac{1}{2\lambda}\|\boldsymbol{x}-\boldsymbol{y}\|^2 + h(\boldsymbol{y})\right\}.$$

This function is finite everywhere [25, Theorem 1.25]. It is not hard to see that

$$e_\lambda h(\boldsymbol{x}) \leq h(\boldsymbol{x}) \tag{3}$$

for all $\boldsymbol{x}$. The infimum in the definition of Moreau envelope is attained at the so-called proximal mapping of λh at $\boldsymbol{x}$, which is defined as

$$\mathbf{prox}_{\lambda h}(\boldsymbol{x}) := \operatorname*{Argmin}_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \frac{1}{2\lambda}\|\boldsymbol{x}-\boldsymbol{u}\|^2 + h(\boldsymbol{u}) \right\}.$$

This set is always nonempty because h is proper closed and bounded below [25, Theorem 1.25]. Let $\boldsymbol{\zeta}_\lambda \in \mathbf{prox}_{\lambda h}(\boldsymbol{x})$. Then we have from [25, Theorem 10.1] and [25, Exercise 8.8(c)] that

$$\frac{1}{\lambda}(\boldsymbol{x}-\boldsymbol{\zeta}_\lambda) \in \partial h(\boldsymbol{\zeta}_\lambda). \tag{4}$$

Furthermore, we have the following simple lemma, which should be well known. We provide a short proof for self-containedness.

Lemma 1 *Let h be a proper closed function with $\inf h > -\infty$ and let $\boldsymbol{x}^* \in \operatorname{dom} h$. Suppose that $\boldsymbol{x}^t \to \boldsymbol{x}^*$, $\lambda_t \downarrow 0$ and pick any $\boldsymbol{\zeta}^t \in \mathbf{prox}_{\lambda_t h}(\boldsymbol{x}^t)$ for each t. Then it holds that $\boldsymbol{\zeta}^t \in \operatorname{dom} h$ for all t and $\boldsymbol{\zeta}^t \to \boldsymbol{x}^*$.*

Proof Under the assumptions, we have the following inequality:

$$\frac{1}{2\lambda_t}\|\boldsymbol{x}^t-\boldsymbol{\zeta}^t\|^2 + \inf h \le \frac{1}{2\lambda_t}\|\boldsymbol{x}^t-\boldsymbol{\zeta}^t\|^2 + h(\boldsymbol{\zeta}^t) = e_{\lambda_t} h(\boldsymbol{x}^t) \le \frac{1}{2\lambda_t}\|\boldsymbol{x}^t-\boldsymbol{x}^*\|^2 + h(\boldsymbol{x}^*).$$

Hence, we have $\boldsymbol{\zeta}^t \in \operatorname{dom} h$ for all t and

$$\begin{aligned}\|\boldsymbol{\zeta}^t-\boldsymbol{x}^*\| &\le \|\boldsymbol{\zeta}^t-\boldsymbol{x}^t\| + \|\boldsymbol{x}^t-\boldsymbol{x}^*\| \\ &\le \sqrt{2\lambda_t(h(\boldsymbol{x}^*)-\inf h) + \|\boldsymbol{x}^t-\boldsymbol{x}^*\|^2} + \|\boldsymbol{x}^t-\boldsymbol{x}^*\| \to 0.\end{aligned}$$

□

Finally, recall that for a nonempty closed set C, the indicator function is defined as

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

We define the (limiting) normal cone at any $\boldsymbol{x} \in C$ as $N_C(\boldsymbol{x}) := \partial\delta_C(\boldsymbol{x})$. We let $\operatorname{dist}(\boldsymbol{x}, C) := \inf_{\boldsymbol{y}\in C}\|\boldsymbol{y}-\boldsymbol{x}\|$. The set of points in the nonempty closed set C that are closest to a given $\boldsymbol{x}$ is denoted by $\mathbf{proj}_C(\boldsymbol{x})$. One can observe that $\mathbf{proj}_C = \mathbf{prox}_{\delta_C}$. The set $\mathbf{proj}_C(\boldsymbol{x})$ at a given $\boldsymbol{x}$ is always nonempty for a nonempty closed set C, and is a singleton when C is in addition convex.

3 Solution method for nonconvex nonsmooth optimization problems

3.1 Successive difference-of-convex approximation method

In this paper, we consider problem (1) and assume that its objective satisfies the assumptions **A1**, **A2** and **A3** in Sect. 1. We will discuss some concrete applications of (1) in more details in Sect. 4. In this section, we present an algorithm for solving (1).

Notice that (1) is in general a nonsmooth nonconvex optimization problem. The nonsmooth nonconvex function $P_0 + \sum_{i=1}^m P_i \circ \mathcal{A}_i$ can be complicated in practice and handling it directly can be challenging. Indeed, although the proximal mappings of $\gamma P_i, i = 0, \ldots, m$, are easy to compute, the proximal mapping of $P_0 + \sum_{i=1}^m P_i \circ \mathcal{A}_i$ may be hard to evaluate and hence the classical proximal gradient algorithm and its variants cannot be adapted directly and efficiently for solving (1). In this paper, we suitably adapt a "smoothing" scheme for solving the above nonconvex nonsmooth problem. In this approach, in each iteration, we minimize the auxiliary function

$$F_{\boldsymbol{\lambda}}(\boldsymbol{x}) := f(\boldsymbol{x}) + P_0(\boldsymbol{x}) + \sum_{i=1}^m e_{\lambda_i} P_i(\mathcal{A}_i \boldsymbol{x}) \tag{5}$$

approximately and then update $\boldsymbol{x}$ and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)$, where $e_{\lambda_i} P_i$ is the Moreau envelope of P_i.

When $P_i, i = 1, \ldots, m$ are all *convex functions*, the corresponding functions $e_{\lambda_i} P_i$ are Lipschitz differentiable [3, Proposition 12.29]. Hence, the function $F_{\boldsymbol{\lambda}}$ becomes the sum of a nonsmooth function P_0 and a smooth function, and can be minimized efficiently using, for example, the proximal gradient algorithm and its variants. This smoothing strategy has been widely used in the literature for convex problems; see [20], and also [4] for a software package for convex optimization problems based on smoothing techniques. However, in our setting, P_i is *not necessarily convex*. Thus, the corresponding Moreau envelope $e_{\lambda_i} P_i$ is *not necessarily smooth* and it is unclear whether $F_{\boldsymbol{\lambda}}$ can be minimized efficiently at first glance.

The key ingredient in our approach (where P_i is possibly nonconvex) is the simple observation that for any nonnegative proper closed function P and any $\mu > 0$,

$$e_\mu P(\boldsymbol{u}) = \frac{1}{2\mu}\|\boldsymbol{u}\|^2 - \underbrace{\sup_{\boldsymbol{y} \in \operatorname{dom} P} \left\{ \frac{1}{\mu}\boldsymbol{u}^\top \boldsymbol{y} - \frac{1}{2\mu}\|\boldsymbol{y}\|^2 - P(\boldsymbol{y}) \right\}}_{D_{\mu,P}(\boldsymbol{u})}. \tag{6}$$

Such a decomposition has been noted in [2] when $P = \delta_C$ for some nonempty closed set C, and in [17, Proposition 3] for the general case. Then $D_{\mu,P}$, as the supreme of affine functions and being finite-valued, is convex continuous. Moreover, using the definition of $e_\mu P(\boldsymbol{u})$, $\mathbf{prox}_{\mu P}(\boldsymbol{u})$ and (6), we see that the supremum in $D_{\mu,P}(\boldsymbol{u})$ is attained at any point in $\mathbf{prox}_{\mu P}(\boldsymbol{u})$. Let $\boldsymbol{y}^* \in \mathbf{prox}_{\mu P}(\mathcal{A}\boldsymbol{x})$. Then $\boldsymbol{y}^* \in \operatorname{dom} P$ and we have for any $\boldsymbol{w}$ that

$$
\begin{aligned}
&D_{\mu,P}(\boldsymbol{w}) - D_{\mu,P}(\mathcal{A}\boldsymbol{x}) \\
&= \sup_{\boldsymbol{y}\in\mathrm{dom}\, P}\left\{\frac{1}{\mu}\boldsymbol{w}^\top\boldsymbol{y} - \frac{1}{2\mu}\|\boldsymbol{y}\|^2 - P(\boldsymbol{y})\right\} - \sup_{\boldsymbol{y}\in\mathrm{dom}\, P}\left\{\frac{1}{\mu}(\mathcal{A}\boldsymbol{x})^\top\boldsymbol{y} - \frac{1}{2\mu}\|\boldsymbol{y}\|^2 - P(\boldsymbol{y})\right\} \\
&\geq \frac{1}{\mu}\boldsymbol{w}^\top\boldsymbol{y}^* - \frac{1}{2\mu}\|\boldsymbol{y}^*\|^2 - P(\boldsymbol{y}^*) - \left(\frac{1}{\mu}(\mathcal{A}\boldsymbol{x})^\top\boldsymbol{y}^* - \frac{1}{2\mu}\|\boldsymbol{y}^*\|^2 - P(\boldsymbol{y}^*)\right) \\
&= \frac{1}{\mu}\boldsymbol{y}^{*\top}(\boldsymbol{w} - \mathcal{A}\boldsymbol{x}).
\end{aligned}
$$

This implies $\frac{1}{\mu}\mathbf{prox}_{\mu P}(\mathcal{A}\boldsymbol{x}) \subseteq \partial D_{\mu,P}(\mathcal{A}\boldsymbol{x})$, from which we deduce further that

$$
\frac{1}{\mu}\mathcal{A}^*\mathbf{prox}_{\mu P}(\mathcal{A}\boldsymbol{x}) \subseteq \mathcal{A}^*\partial D_{\mu,P}(\mathcal{A}\boldsymbol{x}) = \partial(D_{\mu,P}\circ\mathcal{A})(\boldsymbol{x}), \tag{7}
$$

where the last equality follows from [24, Theorem 23.9] because $D_{\mu,P}$ is convex continuous. Thus, (5) is the sum of a smooth function f, a nonsmooth nonconvex function P_0 whose proximal mapping is easy to compute, and a continuous difference-of-convex function such that a subgradient corresponding to its concave part is easy to compute; thanks to (7) and assumption **A2**. Proximal gradient methods with majorization techniques can then be suitably applied to minimizing (5). For instance, one can apply the $\mathrm{NPG}_{\mathrm{major}}$ described in the appendix. Specifically, one can apply $\mathrm{NPG}_{\mathrm{major}}$ with

$$
h(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^m \frac{1}{2\lambda_i}\|\mathcal{A}_i\boldsymbol{x}\|^2,\ P(\boldsymbol{x}) = P_0(\boldsymbol{x}),\ g(\boldsymbol{x}) = \sum_{i=1}^m D_{\lambda_i,P_i}(\mathcal{A}_i\boldsymbol{x}).
$$

It is routine to check that this choice of h, P and g satisfies the assumptions required in the appendix. Moreover, the $F_{\boldsymbol{\lambda}}$ is level-bounded because $f + P_0$ is level-bounded by assumption and $e_{\lambda_i}P_i$ are nonnegative for each $i = 1,\ldots,m$ since P_i are nonnegative. Finally, $F_{\boldsymbol{\lambda}}$ is continuous in its domain because P_0 is. Hence all assumptions required in the appendix for applying $\mathrm{NPG}_{\mathrm{major}}$ are satisfied and the method can be applied to minimizing $F_{\boldsymbol{\lambda}}$ by initializing at *any* point $\boldsymbol{x}^0 \in \mathrm{dom}\, P_0$.

We now describe our method for solving (1) with its update rules below in Algorithm 1. We call this method the successive difference-of-convex approximation method (SDCAM).

We would like to point out that **Step 1** in SDCAM is crucial in our convergence analysis: this strategy was also used in the penalty decomposition method in [15]. As we shall see in the proof of Theorem 2 below, it ensures that (2) can be applied at an accumulation point of $\{\boldsymbol{x}^t\}$.

3.2 Theoretical guarantee for global convergence

In this section, we first discuss how $F_{\boldsymbol{\lambda}_t}$ can be approximately minimized so that (8) is satisfied at the t-th iteration and comment on the computational complexity. Then we prove the convergence of the SDCAM under suitable assumptions.

Algorithm 1 The SDCAM for (1)

Step 0. Pick $m+1$ sequences of positive numbers with $\epsilon_t \downarrow 0$ and $\lambda_{i,t} \downarrow 0$ for $i = 1, \ldots, m$, an $\boldsymbol{x}^{\text{feas}} \in \text{dom } P_0 \cap \bigcap_{i=1}^m \mathcal{A}_i^{-1}(\text{dom } P_i)$, and an $\boldsymbol{x}^0 \in \text{dom } P_0$. Set $t = 0$.

Step 1. If $F_{\boldsymbol{\lambda}_t}(\boldsymbol{x}^t) \le F_{\boldsymbol{\lambda}_t}(\boldsymbol{x}^{\text{feas}})$, set $\boldsymbol{x}^{t,0} = \boldsymbol{x}^t$. Else, set $\boldsymbol{x}^{t,0} = \boldsymbol{x}^{\text{feas}}$.

Step 2. Approximately minimize $F_{\boldsymbol{\lambda}_t}(\boldsymbol{x})$, starting at $\boldsymbol{x}^{t,0}$, and terminating at $\boldsymbol{x}^{t,l_t}$ when

$$\text{dist}\left(0, \nabla f(\boldsymbol{x}^{t,l_t}) + \partial P_0(\boldsymbol{x}^{t,l_t+1}) + \sum_{i=1}^m \frac{1}{\lambda_{i,t}} \mathcal{A}_i^*[\mathcal{A}_i \boldsymbol{x}^{t,l_t} - \mathbf{prox}_{\lambda_{i,t} P_i}(\mathcal{A}_i \boldsymbol{x}^{t,l_t})]\right) \le \epsilon_t, \tag{8}$$
$$\text{and} \quad \|\boldsymbol{x}^{t,l_t+1} - \boldsymbol{x}^{t,l_t}\| \le \epsilon_t, \quad F_{\boldsymbol{\lambda}_t}(\boldsymbol{x}^{t,l_t}) \le F_{\boldsymbol{\lambda}_t}(\boldsymbol{x}^{t,0}).$$

Step 3. Update $\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t,l_t}$ and $t = t+1$. Go to **Step 1.**

As discussed above, $F_{\boldsymbol{\lambda}_t}$ can be minimized by the $\text{NPG}_{\text{major}}$ outlined in the appendix. Moreover, due to (7), one can choose $\boldsymbol{\zeta}^{t,l} = \sum_{i=1}^m \frac{1}{\lambda_{i,t}} \mathcal{A}_i^* \boldsymbol{\zeta}_i^{t,l}$ in the algorithm with

$$\boldsymbol{\zeta}_i^{t,l} \in \mathbf{prox}_{\lambda_{i,t} P_i}(\mathcal{A}_i \boldsymbol{x}^{t,l}) \tag{9}$$

for each $i = 1, \ldots, m$ and $l \ge 0$ so that $\sum_{i=1}^m \frac{1}{\lambda_{i,t}} \mathcal{A}_i^* \boldsymbol{\zeta}_i^{t,l}$ lies in the subdifferential of $\sum_{i=1}^m (D_{\lambda_{i,t}, P_i} \circ \mathcal{A}_i)$ at $\boldsymbol{x}^{t,l}$. Using this special version of $\text{NPG}_{\text{major}}$, we can show that the termination criterion (8) is satisfied after finitely many inner iterations.

Theorem 1 *Suppose that the NPG*$_{\text{major}}$ *is applied with* $\boldsymbol{\zeta}^{t,l} = \sum_{i=1}^m \frac{1}{\lambda_{i,t}} \mathcal{A}_i^* \boldsymbol{\zeta}_i^{t,l}$, *where* $\boldsymbol{\zeta}_i^{t,l}$ *are chosen as in* (9), *to minimizing* $F_{\boldsymbol{\lambda}_t}$ *in the* t*-th iteration of SDCAM. Then the criterion* (8) *is satisfied after finitely many inner iterations.*

Proof According to the convergence properties of the $\text{NPG}_{\text{major}}$, one obtains a sequence $\{\boldsymbol{x}^{t,l}\}_{l \ge 0}$ satisfying

1. $\lim_{l \to \infty} \|\boldsymbol{x}^{t,l+1} - \boldsymbol{x}^{t,l}\| = 0$ (Proposition 2 in the appendix), $F_{\boldsymbol{\lambda}_t}(\boldsymbol{x}^{t,l}) \le F_{\boldsymbol{\lambda}_t}(\boldsymbol{x}^{t,0})$ [thanks to (46)]; and
2. for any $l \ge 0$ [see (45)],

$$\boldsymbol{x}^{t,l+1} \in \underset{\boldsymbol{x}}{\text{Argmin}} \left\{ \left(\nabla f(\boldsymbol{x}^{t,l}) + \sum_{i=1}^m \frac{\boldsymbol{\omega}_i^{t,l}}{\lambda_{i,t}} \right)^\top \boldsymbol{x} + \frac{\bar{L}_{t,l}}{2} \|\boldsymbol{x} - \boldsymbol{x}^{t,l}\|^2 + P_0(\boldsymbol{x}) \right\}, \tag{10}$$

where $\boldsymbol{\omega}_i^{t,l} := \mathcal{A}_i^*[\mathcal{A}_i \boldsymbol{x}^{t,l} - \boldsymbol{\zeta}_i^{t,l}]$. Here, the sequence $\{\bar{L}_{t,l}\}_{l \ge 0}$ can be shown to be bounded; see Proposition 1 in the appendix.

Using [25, Exercise 8.8(c)], the condition (10) implies

$$\begin{aligned}&\mathbf{0} \in \nabla f(\boldsymbol{x}^{t,l}) + \sum_{i=1}^{m} \frac{1}{\lambda_{i,t}} \mathcal{A}_i^*[\mathcal{A}_i \boldsymbol{x}^{t,l} - \boldsymbol{\zeta}_i^{t,l}] + \bar{L}_{t,l}(\boldsymbol{x}^{t,l+1} - \boldsymbol{x}^{t,l}) + \partial P_0(\boldsymbol{x}^{t,l+1}),\\ &\Rightarrow -\bar{L}_{t,l}(\boldsymbol{x}^{t,l+1} - \boldsymbol{x}^{t,l}) \in \nabla f(\boldsymbol{x}^{t,l}) + \sum_{i=1}^{m} \frac{1}{\lambda_{i,t}} \mathcal{A}_i^*[\mathcal{A}_i \boldsymbol{x}^{t,l} - \boldsymbol{\zeta}_i^{t,l}] + \partial P_0(\boldsymbol{x}^{t,l+1}),\end{aligned}$$

from which (8) can be seen to hold with $l_t = l$ when l is sufficiently large because $\lim_{l\to\infty} \|\boldsymbol{x}^{t,l+1} - \boldsymbol{x}^{t,l}\| = 0$ and $\{\bar{L}_{t,l}\}_{l\geq 0}$ is bounded. □

Remark 1 (**Computational complexity**) Suppose that the $\text{NPG}_{\text{major}}$ is applied to minimizing $F_{\boldsymbol{\lambda}_t}$ in each iteration of SDCAM, with the $\boldsymbol{\zeta}^{t,l}$ chosen as in Theorem 1. Then one has to repeatedly solve subproblems of the form (10) for various values of $\boldsymbol{\lambda}_t$ and $\beta > 0$ (in place of $\bar{L}_{t,l}$). These computations are easy under the assumption that the proximal mapping γP_i, $i = 1, \ldots, m$, $\gamma > 0$, is easy to compute. Indeed, the subproblems can be rewritten as

$$\boldsymbol{x}^{t,l+1} \in \mathbf{prox}_{\frac{1}{\beta}P_0}\left(\boldsymbol{x}^{t,l} - \frac{1}{\beta}\left(\nabla f(\boldsymbol{x}^{t,l}) + \sum_{i=1}^{m} \frac{1}{\lambda_{i,t}} \mathcal{A}_i^*[\mathcal{A}_i \boldsymbol{x}^{t,l} - \boldsymbol{\zeta}_i^{t,l}]\right)\right), \tag{11}$$

where $\boldsymbol{\zeta}_i^{t,l} \in \mathbf{prox}_{\lambda_{i,t}P_i}(\mathcal{A}_i \boldsymbol{x}^{t,l})$.

We now state and prove our convergence result for SDCAM. We will comment on (12) in Remark 2 below before proving the theorem.

Theorem 2 (**Convergence of SDCAM**) *Let $\{\boldsymbol{x}^t\}$ be the sequence generated by SDCAM for solving (1). Then $\{\boldsymbol{x}^t\}$ is bounded. Let $\boldsymbol{x}^*$ be an accumulation point of this sequence. Then we have the following results.*

(i) *It holds that $\boldsymbol{x}^* \in \text{dom } P_0 \cap \bigcap_{i=1}^m \mathcal{A}_i^{-1}(\text{dom } P_i)$.*
(ii) *Suppose that*

$$\begin{aligned}&\boldsymbol{y}_0 + \sum_{i=1}^{m} \mathcal{A}_i^* \boldsymbol{y}_i = \mathbf{0} \text{ and } \boldsymbol{y}_0 \in \partial^\infty P_0(\boldsymbol{x}^*), \quad \boldsymbol{y}_i \in \partial^\infty P_i(\mathcal{A}_i \boldsymbol{x}^*) \text{ for } i = 1, \ldots, m\\ &\qquad\qquad \Longrightarrow \ \boldsymbol{y}_i = \mathbf{0} \text{ for } i = 0, \ldots, m.\end{aligned} \tag{12}$$

Then $\boldsymbol{x}^*$ is a stationary point of (1), i.e.,

$$\mathbf{0} \in \nabla f(\boldsymbol{x}^*) + \partial P_0(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mathcal{A}_i^* \partial P_i(\mathcal{A}_i \boldsymbol{x}^*). \tag{13}$$

Remark 2 **[Comments on condition (12)]**

(i) Condition (12) is a classical constraint qualification for nonconvex nonsmooth optimization problems; see [25, Corollary 10.9]. It is satisfied, for example, when $\mathcal{A}_i$ equals the identity map for all i, and all but one P_i are locally Lipschitz so that $\partial^\infty P_i(\boldsymbol{x}^*) = \{\mathbf{0}\}$ for all but one P_i; see [25, Exercise 10.10].

(ii) Under (12), it can be shown using [25, Theorem 10.1], [25, Proposition 10.5] and [25, Theorem 10.6] that any local minimizer $\boldsymbol{x}^*$ of (1) satisfies (13).

Proof Using the nonnegativity of P_i, the last criterion in (8) and the definitions of $F_{\boldsymbol{\lambda}}$ and $\boldsymbol{x}^{t,0}$, we see that

$$f(\boldsymbol{x}^t) + P_0(\boldsymbol{x}^t) \le F_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{x}^t) \le F_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{x}^{\text{feas}}) \le F(\boldsymbol{x}^{\text{feas}}) =: F_{\text{feas}}, \tag{14}$$

where the last inequality follows from the definitions of F, $F_{\boldsymbol{\lambda}}$ and (3). From this, one immediately conclude that $\{\boldsymbol{x}^t\}$ is bounded because $f + P_0$ is level-bounded.

Next, let $\boldsymbol{x}^*$ be an accumulation point of $\{\boldsymbol{x}^t\}$. Then there exists a subsequence $\{\boldsymbol{x}^t\}_{t\in\mathcal{I}}$ so that $\lim_{t\in\mathcal{I}} \boldsymbol{x}^t = \boldsymbol{x}^*$. Using this, (14), and the lower semicontinuity of $f + P_0$, we further see that

$$f(\boldsymbol{x}^*) + P_0(\boldsymbol{x}^*) \le \liminf_{t\in\mathcal{I}} f(\boldsymbol{x}^t) + P_0(\boldsymbol{x}^t) \le F_{\text{feas}} < \infty.$$

This shows that $\boldsymbol{x}^* \in \text{dom}\, P_0$. On the other hand, since P_i is nonnegative, we have

$$\begin{aligned} 0 &\le \frac{1}{2}\text{dist}^2(\mathcal{A}_i\boldsymbol{x}, \text{dom}\, P_i) = \inf_{\boldsymbol{y}\in\text{dom}\, P_i}\left\{\frac{1}{2}\|\mathcal{A}_i\boldsymbol{x} - \boldsymbol{y}\|^2\right\} \\ &\le \inf_{\boldsymbol{y}\in\text{dom}\, P_i}\left\{\frac{1}{2}\|\mathcal{A}_i\boldsymbol{x} - \boldsymbol{y}\|^2 + \lambda_{i,t-1}P_i(\boldsymbol{y})\right\} = \lambda_{i,t-1}e_{\lambda_{i,t-1}}P_i(\mathcal{A}_i\boldsymbol{x}) \end{aligned}$$

for all $\boldsymbol{x}$ and for each $i = 1, \ldots, m$. Using this, the finiteness of $\underline{\ell} := \inf\{f + P_0\}$ (thanks to the level-boundedness of $f + P_0$), and the definition of $F_{\boldsymbol{\lambda}}$, we have for each $i = 1, \ldots, m$ that

$$\underline{\ell} + \frac{1}{2\lambda_{i,t-1}}\text{dist}^2(\mathcal{A}_i\boldsymbol{x}^t, \text{dom}\, P_i) \le \underline{\ell} + e_{\lambda_{i,t-1}}P_i(\mathcal{A}_i\boldsymbol{x}^t) \le F_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{x}^t) \le F_{\text{feas}},$$

where the last inequality follows from (14). Since $\lambda_{i,t-1} \downarrow 0$, we conclude that $\text{dist}^2(\mathcal{A}_i\boldsymbol{x}^*, \text{dom}\, P_i) \le 0$ and hence $\mathcal{A}_i\boldsymbol{x}^* \in \text{dom}\, P_i$ because $\text{dom}\, P_i$ is closed.

We now prove (13) under (12). For notational simplicity, let $\boldsymbol{y}^{t+1} := \boldsymbol{x}^{t,l_t+1}$. Then $\lim_{t\in\mathcal{I}} \boldsymbol{y}^t = \boldsymbol{x}^*$ thanks to the second relation in (8). Moreover, from the first relation in (8), we see that there exist $\boldsymbol{\xi}^t$ with $\|\boldsymbol{\xi}^t\| \le \epsilon_{t-1}$, $\boldsymbol{\eta}^t \in \partial P_0(\boldsymbol{y}^t)$ and $\boldsymbol{\zeta}_i^t \in \mathbf{prox}_{\lambda_{i,t-1}P_i}(\mathcal{A}_i\boldsymbol{x}^t)$ for each $i = 1, \ldots, m$ so that

$$\boldsymbol{\xi}^t = \nabla f(\boldsymbol{x}^t) + \boldsymbol{\eta}^t + \sum_{i=1}^m \frac{1}{\lambda_{i,t-1}}\mathcal{A}_i^*(\mathcal{A}_i\boldsymbol{x}^t - \boldsymbol{\zeta}_i^t). \tag{15}$$

Define

$$r_t := \|\boldsymbol{\eta}^t\| + \sum_{i=1}^{m} \frac{1}{\lambda_{i,t-1}} \|\mathcal{A}_i^*(\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t)\|.$$

We claim that $\{r_t\}_{t\in\mathcal{I}}$ is bounded. Suppose to the contrary that $\{r_t\}_{t\in\mathcal{I}}$ is unbounded and we assume without loss of generality that $\lim_{t\in\mathcal{I}} r_t = \infty$ and $\inf_{t\in\mathcal{I}} r_t > 0$. Then the sequences $\{\frac{1}{r_t}\boldsymbol{\eta}^t\}_{t\in\mathcal{I}}$ and $\{\frac{1}{\lambda_{i,t-1} r_t}\mathcal{A}_i^*(\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t)\}_{t\in\mathcal{I}}$ for $i = 1, \ldots, m$ are bounded. Without loss of generality, we may assume

$$\lim_{t\in\mathcal{I}} \frac{\boldsymbol{\eta}^t}{r_t} = \boldsymbol{\eta}^* \quad \text{and} \quad \lim_{t\in\mathcal{I}} \mathcal{A}_i^* \left(\frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1} r_t} \right) = \boldsymbol{\chi}_i^* \tag{16}$$

for some $\boldsymbol{\eta}^*$ and $\boldsymbol{\chi}_i^*$, $i = 1, \ldots, m$. Notice that

$$1 = \frac{\|\boldsymbol{\eta}^t\| + \sum_{i=1}^m \frac{1}{\lambda_{i,t-1}} \|\mathcal{A}_i^*(\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t)\|}{r_t} \Rightarrow 1 = \|\boldsymbol{\eta}^*\| + \sum_{i=1}^{m} \|\boldsymbol{\chi}_i^*\|. \tag{17}$$

In addition, by dividing r_t from both sides of (15) and passing to the limit along $t \in \mathcal{I}$, we conclude that

$$\mathbf{0} = \boldsymbol{\eta}^* + \sum_{i=1}^{m} \boldsymbol{\chi}_i^*. \tag{18}$$

On the other hand, since $\boldsymbol{\eta}^t \in \partial P_0(\boldsymbol{y}^t)$ and $\lim_{t\in\mathcal{I}} r_t = \infty$, we have from (16), the continuity of P_0 in its domain and (2) that

$$\boldsymbol{\eta}^* \in \partial^\infty P_0(\boldsymbol{x}^*). \tag{19}$$

Next, we prove that $\boldsymbol{\chi}_i^* \in \mathcal{A}_i^* \partial^\infty P_i(\mathcal{A}_i \boldsymbol{x}^*)$ for $i = 1, \ldots, m$. To proceed, we define for each $i = 1, \ldots, m$,

$$w_i^t := \left\| \frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1} r_t} \right\|$$

and claim that $\{w_i^t\}_{t\in\mathcal{I}}$ is bounded for all $i = 1, \ldots, m$. For an arbitrarily fixed $i \in \{1, \ldots, m\}$, suppose to the contrary that $\{w_i^t\}_{t\in\mathcal{I}}$ is unbounded and we assume without loss of generality that $\lim_{t\in\mathcal{I}} w_i^t = \infty$ and that

$$\lim_{t\in\mathcal{I}} \frac{1}{w_i^t} \frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1} r_t} = \boldsymbol{\psi}_i^* \tag{20}$$

for some $\boldsymbol{\psi}_i^*$ with unit norm. Then from the second equation in (16), we have

$$\|\boldsymbol{\psi}_i^*\| = 1 \ \text{ and } \ \mathcal{A}_i^* \boldsymbol{\psi}_i^* = \mathbf{0}. \tag{21}$$

In addition, we observe from (20) that

$$\boldsymbol{\psi}_i^* = \lim_{t\in\mathcal{I}} \frac{1}{w_i^t} \frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1} r_t} \in \left\{ \lim_{t\in\mathcal{I}} \frac{1}{w_i^t r_t} \boldsymbol{u}^t \,:\, \boldsymbol{u}^t \in \partial P_i(\boldsymbol{\zeta}_i^t) \text{ for each } t \right\} \subseteq \partial^\infty P_i(\mathcal{A}_i \boldsymbol{x}^*).$$

where the first inclusion follows from (4) and the second inclusion follows from Lemma 1 (so that $\lim_{t\in\mathcal{I}} \boldsymbol{\zeta}_i^t = \mathcal{A}_i \boldsymbol{x}^*$ and $\{\boldsymbol{\zeta}_i^t\}_{t\in\mathcal{I}} \subseteq \operatorname{dom} P_i$), the continuity of P_i in its domain and (2). These together with the facts $\mathbf{0} \in \partial^\infty P_0(\boldsymbol{x}^*)$, $\mathbf{0} \in \partial^\infty P_i(\mathcal{A}_i \boldsymbol{x}^*)$ $(i = 1, \ldots, m)$[1] and (21) contradict (12). Consequently, $\{w_i^t\}_{t\in\mathcal{I}}$ is bounded for all $i = 1, \ldots, m$. Then, without loss of generality, we assume that $\lim_{t\in\mathcal{I}} \frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1} r_t}$ exists for all $i = 1, \ldots, m$. Then, for each $i = 1, \ldots, m$, we observe from (16) that

$$\boldsymbol{\chi}_i^* = \mathcal{A}_i^* \lim_{t\in\mathcal{I}} \frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1} r_t} \in \mathcal{A}_i^* \left\{ \lim_{t\in\mathcal{I}} \frac{1}{r_t} \boldsymbol{u}^t \,:\, \boldsymbol{u}^t \in \partial P_i(\boldsymbol{\zeta}_i^t) \text{ for each } t \right\} \subseteq \mathcal{A}_i^* \partial^\infty P_i(\mathcal{A}_i \boldsymbol{x}^*),$$

where the first inclusion follows from (4) and the second inclusion follows from Lemma 1 (so that $\lim_{t\in\mathcal{I}} \boldsymbol{\zeta}_i^t = \mathcal{A}_i \boldsymbol{x}^*$ and $\{\boldsymbol{\zeta}_i^t\}_{t\in\mathcal{I}} \subseteq \operatorname{dom} P_i$ for each $i = 1, \ldots, m$), the continuity of P_i in its domain and (2). These together with (17), (18) and (19) contradict (12). Consequently, $\{r_t\}_{t\in\mathcal{I}}$ is bounded.

Since $\{r_t\}_{t\in\mathcal{I}}$ is bounded, we may assume without loss of generality that

$$\lim_{t\in\mathcal{I}} \boldsymbol{\eta}^t = \tilde{\boldsymbol{\eta}}^* \quad \text{and} \quad \lim_{t\in\mathcal{I}} \frac{1}{\lambda_{i,t-1}} \mathcal{A}_i^* (\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t) = \tilde{\boldsymbol{\chi}}_i^* \tag{22}$$

for some $\tilde{\boldsymbol{\eta}}^*$ and $\tilde{\boldsymbol{\chi}}_i^*$, $i = 1, \ldots, m$. Then we have from (2) and the continuity of P_0 in its domain that

$$\tilde{\boldsymbol{\eta}}^* \in \partial P_0(\boldsymbol{x}^*). \tag{23}$$

Next, we prove that $\tilde{\boldsymbol{\chi}}_i^* \in \mathcal{A}_i^* \partial P_i(\mathcal{A}_i \boldsymbol{x}^*)$ for $i = 1, \ldots, m$. To proceed, we define for each $i = 1, \ldots, m$,

$$\nu_i^t := \left\| \frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1}} \right\|$$

and claim that $\{\nu_i^t\}_{t\in\mathcal{I}}$ is bounded for all $i = 1, \ldots, m$. For an arbitrary fixed $i \in \{1, \ldots, m\}$, suppose to the contrary that $\{\nu_i^t\}_{t\in\mathcal{I}}$ is unbounded and we assume without loss of generality that $\lim_{t\in\mathcal{I}} \nu_i^t = \infty$ and that

$$\lim_{t\in\mathcal{I}} \frac{1}{\nu_i^t} \frac{\mathcal{A}_i \boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1}} = \boldsymbol{\phi}_i^* \tag{24}$$

for some $\boldsymbol{\phi}_i^*$ with unit norm. Notice from the second equation of (22) that

$$\|\boldsymbol{\phi}_i^*\| = 1 \quad \text{and} \quad \mathcal{A}_i^* \boldsymbol{\phi}_i^* = \mathbf{0}. \tag{25}$$

[1] These follow from (i) and [25, Corollary 8.10].

In addition, we observe from (24) that

$$\boldsymbol{\phi}_i^* = \lim_{t\in\mathcal{I}} \frac{1}{\nu_i^t}\frac{\mathcal{A}_i\boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1}} \in \left\{\lim_{t\in\mathcal{I}} \frac{1}{\nu_i^t}\boldsymbol{u}^t \,:\, \boldsymbol{u}^t \in \partial P_i(\boldsymbol{\zeta}_i^t) \text{ for each } t\right\} \subseteq \partial^\infty P_i(\mathcal{A}_i\boldsymbol{x}^*).$$

where the first inclusion follows from (4) and the second inclusion follows from Lemma 1 (so that $\lim_{t\in\mathcal{I}} \boldsymbol{\zeta}_i^t = \mathcal{A}_i x^*$ and $\{\boldsymbol{\zeta}_i^t\}_{t\in\mathcal{I}} \subseteq \text{dom } P_i$), the continuity of P_i in its domain and (2). These together with the facts $\mathbf{0} \in \partial^\infty P_0(\boldsymbol{x}^*)$, $\mathbf{0} \in \partial^\infty P_i(\mathcal{A}_i\boldsymbol{x}^*)$ $(i = 1,\ldots,m)$[2] and (25) contradict (12). Consequently, $\{\nu_i^t\}_{t\in\mathcal{I}}$ is bounded for all $i = 1,\ldots,m$. Then, without loss of generality, we assume that $\lim_{t\in\mathcal{I}} \frac{\mathcal{A}_i\boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1}}$ exists for all $i = 1,\ldots,m$. Therefore, for each $i = 1,\ldots,m$, we obtain from (22) that

$$\tilde{\boldsymbol{\chi}}_i^* = \mathcal{A}_i^* \lim_{t\in\mathcal{I}} \frac{\mathcal{A}_i\boldsymbol{x}^t - \boldsymbol{\zeta}_i^t}{\lambda_{i,t-1}} \in \mathcal{A}_i^* \left\{\lim_{t\in\mathcal{I}} \boldsymbol{u}^t \,:\, \boldsymbol{u}^t \in \partial P_i(\boldsymbol{\zeta}_i^t) \text{ for each } t\right\} \subseteq \mathcal{A}_i^*\partial P_i(\mathcal{A}_i\boldsymbol{x}^*), \tag{26}$$

where the first inclusion follows from (4) and the second inclusion follows from Lemma 1 (so that $\lim_{t\in\mathcal{I}} \boldsymbol{\zeta}_i^t = \mathcal{A}_i\boldsymbol{x}^*$ and $\{\boldsymbol{\zeta}_i^t\}_{t\in\mathcal{I}} \subseteq \text{dom } P_i$ for each $i = 1,\ldots,m$), the continuity of P_i in its domain and (2). Passing to the limit in (15) along $t \in \mathcal{I}$ and invoking (22), (23) and (26), we see that

$$0 = \nabla f(\boldsymbol{x}^*) + \tilde{\boldsymbol{\eta}}^* + \sum_{i=1}^m \tilde{\boldsymbol{\chi}}_i^* \in \nabla f(\boldsymbol{x}^*) + \partial P_0(\boldsymbol{x}^*) + \sum_{i=1}^m \mathcal{A}_i^*\partial P_i(\mathcal{A}_i\boldsymbol{x}^*).$$

This completes the proof. □

Remark 3 If, instead of (8), one can guarantee that

$$F_{\boldsymbol{\lambda}_t}(\boldsymbol{x}^{t,l_t}) \le \inf F_{\boldsymbol{\lambda}_t} + \epsilon_t,$$

then one can show that any accumulation point of the sequence $\{\boldsymbol{x}^t\}$ generated by SDCAM is a global minimizer of (1). To see this, recall from [25, Theorem 1.25] that $e_{\lambda_{i,t}} P_i(\mathcal{A}_i\boldsymbol{x}) \to P_i(\mathcal{A}_i\boldsymbol{x})$ for each i and all $\boldsymbol{x}$, and from the discussion on [25, Page 244] that $\{(e_{\lambda_{i,t}} P_i)\circ\mathcal{A}_i\}$ epiconverges to $P_i\circ\mathcal{A}_i$ for each i. Using these together with [25, Theorem 7.46], we further see that $\{F_{\boldsymbol{\lambda}_t}\}$ epiconverges to F. Now, in view of [25, Theorem 7.31(b)], we conclude that any accumulation point of the sequence $\{\boldsymbol{x}^t\}$ generated by SDCAM is a global minimizer of F.

[2] These follow from (i) and [25, Corollary 8.10].

4 Applications to structured optimization problems

4.1 Problems involving sparsity

Consider the following ℓ_0-constrained optimization problem discussed in [30]:

$$\begin{array}{l} \underset{\boldsymbol{x}}{\text{minimize}} \ f(\boldsymbol{x}) \\ \text{subject to } \|\boldsymbol{x}\|_0 \leq k, \ \boldsymbol{x} \in C, \end{array} \tag{27}$$

where f is as in (1) and C is a nonempty closed set. This model includes many important application problems such as sparse principal component analysis, sparse portfolio selection and sparse nonnegative linear regression as special cases. These applications typically involve a closed set C whose projection is easy to compute. For instance, we have $f(\boldsymbol{x}) = -\boldsymbol{x}^\top \boldsymbol{V}\boldsymbol{x}$, defined with a covariance matrix $\boldsymbol{V} \in \mathcal{S}^n$ and $C = \{\boldsymbol{x} : \|\boldsymbol{x}\| = 1\}$ for sparse principal component analysis [27]. As another example, for sparse nonnegative linear regression [26], $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ defined with $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$, and $C = \{\boldsymbol{x} : \boldsymbol{x} \geq 0\}$ are used. For these two examples, the direct projection onto $C \cap \{\boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq k\}$ is easy to compute, and the proximal gradient algorithm can then be applied to solving (27).

We next discuss a specific example where the direct projection onto $C \cap \{\boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq k\}$ might not be easy to compute, and describe how our SDCAM can be applied.

Example 1 **(Sparse portfolio problem)** Given a basket of investable assets, the Markowitz model [19] seeks to find the optimal asset allocation of the portfolio by minimizing the estimated variance with an expected return above a specified level. More recently, [6] has added the ℓ_1-norm to the classical Markowitz model to obtain sparse portfolios, and after that, various types of sparse regularizers such as ℓ_p-norm $(0 < p < 1)$ are incorporated into the Markowitz model (e.g., [8]).

The sparse portfolio selection problem we consider here takes the following form:

$$\begin{array}{l} \underset{\boldsymbol{x}}{\text{minimize}} \ f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} \\ \text{subject to } \|\boldsymbol{x}\|_0 \leq k, \ \boldsymbol{x} \geq 0, \ \boldsymbol{e}^\top \boldsymbol{x} = 1, \ \boldsymbol{r}^\top \boldsymbol{x} = r_0, \end{array} \tag{28}$$

where $\boldsymbol{Q} \in \mathcal{S}^n$ is the estimated covariance matrix of the portfolio, $\boldsymbol{r} \in \mathbb{R}^n$ is the estimated mean return vector of investable assets, $r_0 \in \mathbb{R}$ is a specific return level, and $\boldsymbol{e}$ is the vector of all ones. The constraint $\boldsymbol{x} \geq 0$ is known as the non-shortsale constraint, and model (28) is the formulation of the shorting-prohibited sparse Markowitz model. We assume here that the feasible set of (28) is nonempty.

Notice that the feasible set of (28) is compact and hence (28) has a solution. Let $\boldsymbol{x}^*$ be a solution of (28) and $\tau \gg \max_i |x_i^*|$. Define $\Omega := \{\boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq k, 0 \leq \boldsymbol{x} \leq \tau\}$ and $S := \{\boldsymbol{x} : \boldsymbol{e}^\top \boldsymbol{x} = 1, \boldsymbol{r}^\top \boldsymbol{x} = r_0\}$. Then (28) can be rewritten in the form of (1) (with the same optimal value) as follows

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x}) + \underbrace{\delta_\Omega(\boldsymbol{x})}_{P_0(\boldsymbol{x})} + \underbrace{\delta_S(\boldsymbol{x})}_{P_1(\boldsymbol{x})}, \tag{29}$$

in which $f + P_0$ is level-bounded. Therefore, we can apply SDCAM in Sect. 3 to (29), and in each subproblem of SDCAM we can use $\text{NPG}_{\text{major}}$ to minimize F_{λ_t} as described in Theorem 1. The method involves computing two projections $\mathbf{proj}_\Omega$ and $\mathbf{proj}_S$, which are easy to compute. Indeed, we have $\max\{\min\{\tilde{H}_k(\boldsymbol{y}), \tau\}, 0\} \in \mathbf{proj}_\Omega(\boldsymbol{y})$, where $\tilde{H}_k(\boldsymbol{v})$ keeps any k largest entries of $\boldsymbol{v}$ and sets the rest to zero.[3] □

In statistics, ℓ_1-norm regularizer has been used for inducing sparsity in variable selection problems; see Lasso [28], which is an application of the ℓ_1 penalty to linear regression. A more general model of Lasso, the generalized Lasso [29], has been proposed as

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + c\|\boldsymbol{D}\boldsymbol{x}\|_1,$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a matrix of predictors, $\boldsymbol{b} \in \mathbb{R}^m$ is a response vector, $c \geq 0$ is a tuning parameter and $\boldsymbol{D} \in \mathbb{R}^{d\times n}$ is a specified penalty matrix. The term $\|\boldsymbol{D}\boldsymbol{x}\|_1$ can enforce certain structural sparsity on the coefficients in the solution. For example, with an appropriate $\boldsymbol{D}$, $\|\boldsymbol{D}\boldsymbol{x}\|_1$ can express $\sum_{i=2}^n |x_i - x_{i-1}|$, which penalizes the absolute differences in adjacent coordinates of $\boldsymbol{x}$. This specific $\boldsymbol{D}$ leads to the so-called fused Lasso. A variant of this type of regularizer (anisotropic total variation regularizer) is also used in image processing for minimizing the horizontal or/and vertical differences between pixels. Some other applications which require a non-identity matrix $\boldsymbol{D}$ in the generalized Lasso were discussed in [29]. In the next example, we discuss how our SDCAM can be applied to some nonconvex variants of the generalized Lasso problem.

Example 2 **(Nonconvex fused regularized problem)** Similarly as in [21], we consider the following nonconvex fused regularized problem

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + c_1\phi_1(\boldsymbol{x}) + c_2\phi_2(\boldsymbol{D}\boldsymbol{x}), \tag{30}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{D}\boldsymbol{x} = (x_2 - x_1, \ldots, x_n - x_{n-1})^\top$, $c_1 > 0$ and $c_2 > 0$ are regularization parameters, $\phi_1(\boldsymbol{x}) = \sum_{i=1}^n \varphi_i(|x_i|)$ and ϕ_2 are nonconvex sparsity-inducing regularizers with $\varphi_i : \mathbb{R}_+ \to \mathbb{R}_+$ being closed and nondecreasing, and $\phi_2 : \mathbb{R}^{n-1} \to \mathbb{R}_+$ being closed and level-bounded.

[3] To see this, recall from [15, Proposition 3.1] that an element $\boldsymbol{\zeta}^*$ of $\mathbf{proj}_\Omega(\boldsymbol{y})$ can be obtained as

$$\zeta_i^* = \begin{cases} \tilde{\zeta}_i^* & \text{if } i \in I^*, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{\zeta}_i^* = \text{argmin}\{\frac{1}{2}(\zeta_i - y_i)^2 : 0 \leq \zeta_i \leq \tau\} = \max\{\min\{y_i, \tau\}, 0\}$, and I^* is an index set of size k corresponding to the k largest values of $\{\frac{1}{2}y_i^2 - \frac{1}{2}(\tilde{\zeta}_i^* - y_i)^2\}_{i=1}^n = \{\frac{1}{2}y_i^2 - \frac{1}{2}(\min\{\max\{y_i - \tau, 0\}, y_i\})^2\}_{i=1}^n$. Since the function $t \mapsto \frac{1}{2}t^2 - \frac{1}{2}(\min\{\max\{t - \tau, 0\}, t\})^2$ is nondecreasing, we can let I^* correspond to any k largest entries of $\boldsymbol{y}$.

Note that (30) can be rewritten in the form of

$$\underset{\boldsymbol{x}}{\text{minimize}}\ g(\tilde{\boldsymbol{A}}\boldsymbol{x}-\tilde{\boldsymbol{b}})+\Psi(\boldsymbol{x}), \tag{31}$$

in which $\tilde{\boldsymbol{A}}=\begin{pmatrix}\boldsymbol{A}\\ \boldsymbol{D}\end{pmatrix}$, $\tilde{\boldsymbol{b}}=\begin{pmatrix}\boldsymbol{b}\\ \boldsymbol{0}\end{pmatrix}$, $g(\boldsymbol{y})=\frac{1}{2}\|\boldsymbol{y}_1\|^2+c_2\phi_2(\boldsymbol{y}_2)$ with $\boldsymbol{y}:=(\boldsymbol{y}_1,\boldsymbol{y}_2)\in\mathbb{R}^m\times\mathbb{R}^{n-1}$, and $\Psi(\boldsymbol{x})=c_1\sum_{i=1}^n\varphi_i(|x_i|)$. It is routine to check that g and Ψ satisfy [14, Assumption 2]. Hence, according to [14, Theorem 2.1], we know that (31), and hence (30), has at least one solution.

Notice that we can directly apply the SDCAM in Sect. 3 to (30) when ϕ_1 is level-bounded, e.g., $\phi_1(\boldsymbol{x})=\|\boldsymbol{x}\|^p$: we set $f(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2$, $P_0=c_1\phi_1$ and $P_1=c_2\phi_2$ with $\mathcal{A}_1=\boldsymbol{D}$ in this case. When the $\text{NPG}_{\text{major}}$ is applied as described in Theorem 1 for solving the corresponding subproblems, it involves computing the proximal mappings $\mathbf{prox}_{\mu\phi_1}$ and $\mathbf{prox}_{\mu\phi_2}$ for $\mu>0$. These are easy to compute for many well-known nonconvex sparse regularizers; see [12].

Finally, in the case when ϕ_1 is not level-bounded, let $\boldsymbol{x}^*$ be a solution of (30) and $\tau\gg\max_i|x_i^*|$. We define $\Omega:=\{\boldsymbol{x}:\max_i|x_i|\le\tau\}$ and rewrite (30) in the form of (1) (with the same optimal value) as follows

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \underbrace{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2}_{f(\boldsymbol{x})}+\underbrace{c_1\sum_{i=1}^n\varphi_i(|x_i|)+\delta_\Omega(\boldsymbol{x})}_{P_0(\boldsymbol{x})}+\underbrace{c_2\phi_2(\boldsymbol{D}\boldsymbol{x})}_{P_1(\mathcal{A}_1\boldsymbol{x})}. \tag{32}$$

Then $f+P_0$ is level-bounded and hence the SDCAM in Sect. 3 can be applied. When the $\text{NPG}_{\text{major}}$ is applied in the subproblem of SDCAM as described in Theorem 1, it involves computing the proximal mappings $\mathbf{prox}_{\mu P_0}$ and $\mathbf{prox}_{\mu\phi_2}$ for $\mu>0$. Note that $\mathbf{prox}_{\mu P_0}$ can be obtained from $\mathbf{prox}_{\mu\psi_i}$ with $\psi_i(x_i):=c_1\varphi_i(|x_i|)+\delta_{|\cdot|\le\tau}(x_i)$, $i=1,\ldots,n$, which can be efficiently computed for various nonconvex sparse regularizers such as SCAD, MCP, ℓ_p penalty and Capped-ℓ_1 (see [12]). Finally, the computation of $\mathbf{prox}_{\mu\phi_2}$ is also easy for many of these regularizers. □

4.2 Problems with rank constraints

Our algorithm can also be applied to rank-constrained nonconvex nonsmooth matrix optimization problems. We discuss some concrete examples below.

For notational simplicity, from now on, we let

$$\Xi_k:=\{\boldsymbol{X}:\ \text{rank}(\boldsymbol{X})\le k\}$$

for a given integer k. Note that if $P_1=\delta_{\Xi_k}$, then

$$e_{\lambda_1}P_1(\boldsymbol{X})=\frac{1}{2\lambda_1}\text{dist}^2(\boldsymbol{X},\Xi_k)=\frac{1}{2\lambda_1}(\|\boldsymbol{X}\|_F^2-\|\boldsymbol{X}\|_{k,2}^2),$$

where $\|\boldsymbol{X}\|_{k,2}^2$ denotes the sum of squares of the k largest singular values of $\boldsymbol{X}$. The function $\boldsymbol{X} \mapsto \|\boldsymbol{X}\|_F^2 - \|\boldsymbol{X}\|_{k,2}^2$ is a "rank-related" variant of the so-called k-sparsity functions [1] because the relation $\mathrm{rank}(\boldsymbol{X}) \leq k$ can be equivalently expressed as $\|\boldsymbol{X}\|_F^2 - \|\boldsymbol{X}\|_{k,2}^2 = 0$. A variant of this function was used in [30] as a penalty function for inducing sparsity. It is interesting to note that this function falls out naturally from the Moreau envelope of the indicator function of Ξ_k.

Example 3 **(Matrix completion)** The problem of recovering a low-rank data matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ from a sampling of its entries is known as the matrix completion problem [7]. This problem can be formulated as

$$\begin{array}{ll}\underset{\boldsymbol{X}}{\text{minimize}} & \mathrm{rank}(\boldsymbol{X}) \\ \text{subject to} & P_\Omega(\boldsymbol{X}) = P_\Omega(\boldsymbol{M}),\end{array}$$

where Ω is the index set of known entries of $\boldsymbol{M}$, and P_Ω is the sampling map defined as

$$[P_\Omega(\boldsymbol{Y})]_{ij} = \begin{cases} Y_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.}\end{cases}$$

When the entries of the data matrix are noisy, one can consider the following variants of the above model:

$$\begin{array}{ll}\underset{\boldsymbol{X}}{\text{minimize}} & \|P_\Omega(\boldsymbol{X}) - P_\Omega(\boldsymbol{M})\|_F^2 \\ \text{subject to} & \mathrm{rank}(\boldsymbol{X}) \leq k,\end{array} \quad \text{or} \quad \underset{\boldsymbol{X}}{\text{minimize}} \ \|P_\Omega(\boldsymbol{X}) - P_\Omega(\boldsymbol{M})\|_F^2 + \mu\,\mathrm{rank}(\boldsymbol{X}),$$

where $\mu > 0$ is tuning parameter, and k is a positive integer. Since these problems are nonconvex in general, some popular convex relaxation approaches have been proposed, where the rank function is replaced by the nuclear norm function [22]. The convex relaxations can be shown to be equivalent to the original nonconvex problems under suitable conditions [7].

Here we consider the following variation of the matrix completion problem:

$$\begin{array}{ll}\underset{\boldsymbol{X}}{\text{minimize}} & \dfrac{1}{2}\|P_\Omega(\boldsymbol{X}) - P_\Omega(\boldsymbol{M})\|_F^2 \\ \text{subject to} & P_\Theta(\boldsymbol{X}) = P_\Theta(\boldsymbol{M}),\ \mathrm{rank}(\boldsymbol{X}) \leq k,\end{array} \tag{33}$$

where Ω is an index set corresponding to possibly *noisy* known entries of $\boldsymbol{M}$, and Θ is another index set corresponding to *noiseless* known entries of $\boldsymbol{M}$. Suppose that (33) has a solution $\boldsymbol{X}^*$, and take $\tau \gg \max\{\max_{i,j} |X_{ij}^*|, \sigma_{\max}(\boldsymbol{X}^*)\}$.

Let $S := \{\boldsymbol{X} : P_\Theta(\boldsymbol{X}) = P_\Theta(\boldsymbol{M})\}$, $\tilde{S} := \{\boldsymbol{X} \in S : \max_{i,j} |X_{ij}| \leq \tau\}$ and $\tilde{\Xi}_k := \{\boldsymbol{X} \in \Xi_k : \sigma_{\max}(\boldsymbol{X}) \leq \tau\}$. Then (33) can be rewritten in the form of (1) (with the same optimal value) in the following two ways:

$$\underset{\boldsymbol{X}}{\text{minimize}} \ \underbrace{\frac{1}{2}\|P_{\Omega}(\boldsymbol{X}-\boldsymbol{M})\|_F^2}_{f(\boldsymbol{X})} + \underbrace{\delta_S(\boldsymbol{X})}_{P_1(\boldsymbol{X})} + \underbrace{\delta_{\tilde{\Xi}_k}(\boldsymbol{X})}_{P_0(\boldsymbol{X})}, \tag{34}$$

$$\underset{\boldsymbol{X}}{\text{minimize}} \ \underbrace{\frac{1}{2}\|P_{\Omega}(\boldsymbol{X}-\boldsymbol{M})\|_F^2}_{f(\boldsymbol{X})} + \underbrace{\delta_{\tilde{S}}(\boldsymbol{X})}_{P_0(\boldsymbol{X})} + \underbrace{\delta_{\Xi_k}(\boldsymbol{X})}_{P_1(\boldsymbol{X})}. \tag{35}$$

Note that in both cases, $f + P_0$ is level-bounded and hence the SDCAM in Sect. 3 can be applied.

Suppose that SDCAM is applied to (34). Then when the $\text{NPG}_{\text{major}}$ is applied as described in Theorem 1 for solving the subproblems, it requires computing $\mathbf{proj}_S$ and $\mathbf{proj}_{\tilde{\Xi}_k}$. Both of these are easy to compute. In particular, let $\boldsymbol{U}\text{Diag}(\boldsymbol{\sigma})\boldsymbol{V}^\top$ be a singular value decomposition of $\boldsymbol{W}$. Then an element $\boldsymbol{Y} \in \mathbf{proj}_{\tilde{\Xi}_k}(\boldsymbol{W})$ can be computed as $\boldsymbol{Y} = \boldsymbol{U}\text{Diag}(\boldsymbol{\zeta}^*)\boldsymbol{V}^\top$ with $\boldsymbol{\zeta}^* = \min\{H_k(\boldsymbol{\sigma}), \tau\boldsymbol{e}\}$, where $\boldsymbol{e}$ is the vector of all ones, the minimum is taken componentwise, and $H_k(\boldsymbol{v})$ is the hard thresholding operator that keeps any k largest entries of $\boldsymbol{v}$ in magnitude and sets the rest to zero.[4]

On the other hand, when applying SDCAM to (35) with the $\text{NPG}_{\text{major}}$ as described in Theorem 1 applied to the subproblems, one needs to compute $\mathbf{proj}_{\tilde{S}}$ and $\mathbf{proj}_{\Xi_k}$. Again, both of these are easy to compute. In particular, let $\boldsymbol{U}\text{Diag}(\boldsymbol{\sigma})\boldsymbol{V}^\top$ be a singular value decomposition of $\boldsymbol{W}$. Then an element $\boldsymbol{Y} \in \mathbf{proj}_{\Xi_k}(\boldsymbol{W})$ can be computed as $\boldsymbol{Y} = \boldsymbol{U}\text{Diag}(H_k(\boldsymbol{\sigma}))\boldsymbol{V}^\top$. □

Example 4 **(Nearest low-rank correlation matrix)** Finding the nearest low-rank correlation matrix has important applications in finance; see [5,11]. The problem is often formulated as

$$\begin{array}{ll} \underset{\boldsymbol{X}\in\mathcal{S}^n}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{H}\circ(\boldsymbol{X}-\boldsymbol{M})\|_F^2 \\ \text{subject to} & \text{diag}(\boldsymbol{X}) = \boldsymbol{e}, \\ & \boldsymbol{X} \succeq 0, \ \text{rank}(\boldsymbol{X}) \le k, \end{array} \tag{36}$$

where $\mathcal{S}^n$ is the space of $n \times n$ symmetric matrices, $\boldsymbol{H}$ is a given nonnegative weight matrix, $\boldsymbol{M}$ is a given symmetric matrix and $\boldsymbol{e}$ is the vector of all ones, $k \ge 1$. In [11], the constraint $\text{rank}(\boldsymbol{X}) \le k$ was rewritten equivalently as requiring the sum of the $n-k$ smallest eigenvalues equal zero. A penalty approach was then adopted to handle this latter equality constraint.

[4] To see this, recall from [16, Corollary 2.3] and [15, Proposition 3.1] that an element $\boldsymbol{Y} \in \mathbf{proj}_{\tilde{\Xi}_k}(\boldsymbol{W})$ can be computed as $\boldsymbol{Y} = \boldsymbol{U}\text{Diag}(\boldsymbol{\zeta}^*)\boldsymbol{V}^\top$, where

$$\zeta_i^* = \begin{cases} \tilde{\zeta}_i^* & \text{if } i \in I^*, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{\zeta}_i^* = \text{argmin}\{\frac{1}{2}(\zeta_i - \sigma_i)^2 : |\zeta_i| \le \tau\} = \min\{\sigma_i, \tau\}$, and I^* is an index set of size k corresponding to the k largest values of $\{\frac{1}{2}\sigma_i^2 - \frac{1}{2}(\tilde{\zeta}_i^* - \sigma_i)^2\}_{i=1}^n = \{\frac{1}{2}\sigma_i^2 - \frac{1}{2}(\max\{0, \sigma_i - \tau\})^2\}_{i=1}^n$. Since $t \mapsto \frac{1}{2}t^2 - \frac{1}{2}(\max\{0, t-\tau\})^2$ is nondecreasing for nonnegative t, we can take I^* to correspond to any k largest singular values.

In the following, we describe how to solve (36) by the SDCAM in Sect. 3. Notice that for any $\boldsymbol{X} \in \mathcal{S}^n$ satisfying $\mathrm{diag}(\boldsymbol{X}) = \boldsymbol{e}$ and $\boldsymbol{X} \succeq 0$, we have $\boldsymbol{X} \preceq n\,\boldsymbol{I}$. Thus, the feasible set of (36) is compact and hence (36) has a solution. Let $\boldsymbol{X}^*$ be a solution of (36) and $\tau \gg \max\{\max_{i,j} |X^*_{ij}|, \lambda_{\max}(\boldsymbol{X}^*)\}$. Define

$$S := \left\{\boldsymbol{X} \in \mathcal{S}^n : \mathrm{diag}(\boldsymbol{X}) = \boldsymbol{e}\right\}, \quad \tilde{S} := \left\{\boldsymbol{X} \in S : \max_{i,j} |X_{ij}| \le \tau\right\},$$
$$\Pi_k := \{\boldsymbol{X} \succeq 0 : \mathrm{rank}(\boldsymbol{X}) \le k\}, \quad \tilde{\Pi}_k := \{\boldsymbol{X} \in \Pi_k : \lambda_{\max}(\boldsymbol{X}) \le \tau\}.$$

Then (36) can be rewritten in the form of (1) (with the same optimal value) in the following two ways:

$$\underset{\boldsymbol{X}\in\mathcal{S}^n}{\text{minimize}} \ \underbrace{\frac{1}{2}\|\boldsymbol{H} \circ (\boldsymbol{X} - \boldsymbol{M})\|_F^2}_{f(\boldsymbol{X})} + \underbrace{\delta_S(\boldsymbol{X})}_{P_1(\boldsymbol{X})} + \underbrace{\delta_{\tilde{\Pi}_k}(\boldsymbol{X})}_{P_0(\boldsymbol{X})}, \tag{37}$$

$$\underset{\boldsymbol{X}\in\mathcal{S}^n}{\text{minimize}} \ \underbrace{\frac{1}{2}\|\boldsymbol{H} \circ (\boldsymbol{X} - \boldsymbol{M})\|_F^2}_{f(\boldsymbol{X})} + \underbrace{\delta_{\tilde{S}}(\boldsymbol{X})}_{P_0(\boldsymbol{X})} + \underbrace{\delta_{\Pi_k}(\boldsymbol{X})}_{P_1(\boldsymbol{X})}. \tag{38}$$

Notice that in both cases, $f + P_0$ is level-bounded and hence we can apply the SDCAM in Sect. 3.

We first look at (37). When the $\mathrm{NPG}_{\mathrm{major}}$ as described in Theorem 1 is applied to the subproblems, one has to compute $\mathbf{proj}_S$ and $\mathbf{proj}_{\tilde{\Pi}_k}$. Both projections can be easily computed. In particular, let $\boldsymbol{U}\mathrm{Diag}(\boldsymbol{\lambda})\boldsymbol{U}^\top$ be an eigenvalue decomposition of $\boldsymbol{W} \in \mathcal{S}^n$. Then an element $\boldsymbol{Y} \in \mathbf{proj}_{\tilde{\Pi}_k}(\boldsymbol{W})$ can be computed as $\boldsymbol{Y} = \boldsymbol{U}\mathrm{Diag}(\boldsymbol{\zeta}^*)\boldsymbol{V}^\top$ with $\boldsymbol{\zeta}^* = \max\{\min\{\tilde{H}_k(\boldsymbol{\lambda}), \tau\}, 0\}$, where $\tilde{H}_k(\boldsymbol{v})$ keeps any k largest entries of $\boldsymbol{v}$ and sets the rest to zero.[5]

We next turn to (38). In this case, in each $\mathrm{NPG}_{\mathrm{major}}$ iteration, one has to compute $\mathbf{proj}_{\tilde{S}}$ and $\mathbf{proj}_{\Pi_k}$. Again, both projections can be easily computed. In particular, let $\boldsymbol{U}\mathrm{Diag}(\boldsymbol{\lambda})\boldsymbol{U}^\top$ be an eigenvalue decomposition of $\boldsymbol{W} \in \mathcal{S}^n$. Then an element $\boldsymbol{Y} \in \mathbf{proj}_{\Pi_k}(\boldsymbol{W})$ can be computed as $\boldsymbol{Y} = \boldsymbol{U}\mathrm{Diag}(\max\{\tilde{H}_k(\boldsymbol{\lambda}), \boldsymbol{0}\})\boldsymbol{U}^\top$. □

[5] To see this, recall from [16, Proposition 2.8] and [15, Proposition 3.1] that an element $\boldsymbol{Y} \in \mathbf{proj}_{\tilde{\Pi}_k}(\boldsymbol{W})$ can be computed as $\boldsymbol{Y} = \boldsymbol{U}\mathrm{Diag}(\boldsymbol{\zeta}^*)\boldsymbol{V}^\top$, where

$$\zeta_i^* = \begin{cases} \tilde{\zeta}_i^* & \text{if } i \in I^*, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{\zeta}_i^* = \mathrm{argmin}\{\frac{1}{2}(\zeta_i - \lambda_i)^2 : \ 0 \le \zeta_i \le \tau\} = \max\{\min\{\lambda_i, \tau\}, 0\}$, and I^* is an index set of size k corresponding to the k largest values of $\{\frac{1}{2}\lambda_i^2 - \frac{1}{2}(\tilde{\zeta}_i^* - \lambda_i)^2\}_{i=1}^n = \{\frac{1}{2}\lambda_i^2 - \frac{1}{2}(\min\{\max\{\lambda_i - \tau, 0\}, \lambda_i\})^2\}_{i=1}^n$. Since the function $t \mapsto \frac{1}{2}t^2 - \frac{1}{2}(\min\{\max\{t - \tau, 0\}, t\})^2$ is nondecreasing, we can let I^* correspond to any k largest entries of $\boldsymbol{\lambda}$.

Example 5 **(Simultaneously sparse and low rank matrix optimization problem)** The following problem was considered in [23]:

$$\underset{\boldsymbol{X}}{\text{minimize}}\ f(\boldsymbol{X}) + \gamma\|\text{vec}(\boldsymbol{X})\|_1 + \tau\|\boldsymbol{X}\|_*,$$

where f is as in (1), γ and τ are positive numbers. This problem aims at finding solutions which are both sparse and low-rank, and can be applied to identifying clusters in social networks; see [23, Section 6.2]. This model relaxes and penalizes the sparsity index $\|\text{vec}(\boldsymbol{X})\|_0$ and the low-rank index $\text{rank}(\boldsymbol{X})$ by two convex functions $\|\text{vec}(\boldsymbol{X})\|_1$ and $\|\boldsymbol{X}\|_*$, respectively.

Here, we consider the following variant that explicitly incorporates the sparsity and rank constraints:

$$\begin{array}{ll}\underset{\boldsymbol{X}}{\text{minimize}} & f(\boldsymbol{X}) \\ \text{subject to} & \|\text{vec}(\boldsymbol{X})\|_0 \le s,\ \text{rank}(\boldsymbol{X}) \le k.\end{array} \tag{39}$$

Suppose that (39) has a solution $\boldsymbol{X}^*$, and let $\tau \gg \max\{\max_{i,j}|X^*_{ij}|, \sigma_{\max}(\boldsymbol{X}^*)\}$. Define $S := \{\boldsymbol{X} : \|\text{vec}(\boldsymbol{X})\|_0 \le s\}$, $\tilde{S} := \{\boldsymbol{X} \in S : \max_{i,j}|X_{ij}| \le \tau\}$ and $\tilde{\Xi}_k := \{\boldsymbol{X} \in \Xi_k : \sigma_{\max}(\boldsymbol{X}) \le \tau\}$. Then (39) can be rewritten in the form of (1) (with the same optimal value) in the following two ways:

$$\underset{\boldsymbol{X}}{\text{minimize}}\ f(\boldsymbol{X}) + \underbrace{\delta_S(\boldsymbol{X})}_{P_1(\boldsymbol{X})} + \underbrace{\delta_{\tilde{\Xi}_k}(\boldsymbol{X})}_{P_0(\boldsymbol{X})}, \tag{40}$$

$$\underset{\boldsymbol{X}}{\text{minimize}}\ f(\boldsymbol{X}) + \underbrace{\delta_{\tilde{S}}(\boldsymbol{X})}_{P_0(\boldsymbol{X})} + \underbrace{\delta_{\Xi_k}(\boldsymbol{X})}_{P_1(\boldsymbol{X})}. \tag{41}$$

Note that in both cases, $f + P_0$ is level-bounded and hence the SDCAM in Sect. 3 can be applied. When the $\text{NPG}_{\text{major}}$ as described in Theorem 1 is applied to the corresponding subproblems, one has to compute $\mathbf{proj}_S$ and $\mathbf{proj}_{\tilde{\Xi}_k}$ for (40), and $\mathbf{proj}_{\tilde{S}}$ and $\mathbf{proj}_{\Xi_k}$ for (41). All these projections can be computed efficiently; see Examples 1 and 3. □

5 Numerical experiments

In this section, we apply our SDCAM in Sect. 3 with subproblems solved by $\text{NPG}_{\text{major}}$ as described in Theorem 1 to an instance of Examples 2 and 5: the nonconvex fused regularized problem and the simultaneously sparse and low rank matrix optimization problem. All numerical experiments are performed in Matlab R2016a on a 64-bit PC with an Intel(R) Core(TM) i7-6700 CPU (3.41 GHz) and 32 GB of RAM.

5.1 Nonconvex fused regularized problem: comparison against a solution method based on smoothing

We consider the following special instance of nonconvex fused regularized problem:

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{b}\|^2 + c_1\|\boldsymbol{x}\|_1 + c_2\|\boldsymbol{D}\boldsymbol{x}\|_p^p, \tag{42}$$

where $c_1 > 0$, $c_2 > 0$, $p = 0.5$, $\boldsymbol{D}\boldsymbol{x} = (x_2 - x_1, \ldots, x_n - x_{n-1})^\top$, and $\boldsymbol{b} \in \mathbb{R}^n$ is the noisy measurement of a piecewise constant sparse signal. Notice that the function $\|\cdot\|_1$ is level-bounded. We can directly apply SDCAM as described in Example 2 and solve the subproblems by $\text{NPG}_{\text{major}}$. On the other hand, a commonly used technique for handling optimization problems involving ℓ_p penalty functions ($0 < p < 1$) is smoothing. Thus, in our experiments below, we compare SDCAM with a method based on smoothing, the smoothing nonmonotone proximal gradient method (sNPG), for solving (42). In sNPG, we solve the following sequence of subproblems approximately by NPG (this is $\text{NPG}_{\text{major}}$ applied to (44) when $g = 0$):

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \underbrace{\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{b}\|^2 + c_2\sum_{i=1}^{n-1}\left((\boldsymbol{D}\boldsymbol{x})_i^2 + \lambda_t^2\right)^{\frac{p}{2}}}_{f_t(\boldsymbol{x})} + \underbrace{c_1\|\boldsymbol{x}\|_1}_{Q(\boldsymbol{x})},$$

where $\lambda_t \downarrow 0$ is the smoothing parameter. The approximate stationary point of $f_t + Q$ obtained is then used as initialization for minimizing $f_{t+1} + Q$.

Data generation: We first randomly generate a piecewise constant signal $\boldsymbol{x} \in \mathbb{R}^n$ using the following Matlab code:

```
J = randperm(10);I = sort(J(1:6),'ascend');x = zeros(n,1);
for i = 1:r
    if randn > 0
        x(n*I(i)/10 - 3*n/50 - randi(3) : n*I(i)/10) = randi(3);
    else
        x(n*I(i)/10 - 3*n/50 - randi(3) : n*I(i)/10) = -randi(3);
    end
end
```

Then we let $\boldsymbol{b} = \boldsymbol{x} + \sigma\boldsymbol{\xi}$, where $\sigma > 0$ is a noise factor and $\boldsymbol{\xi}$ has i.i.d. standard Gaussian entries. In our experiments, motivated by [21], we choose $c_1 = c_2 = \sigma\sqrt{n}/40$. We shall see that this choice leads to reasonable recovery results in Fig. 1. We also set $\sigma = 0.1$, $n = 2000, 4000, 6000, 8000, 10000$.

Parameter setting: In SDCAM, we set $\lambda_t = 1/10^{t+1}$ and $\boldsymbol{x}^{\text{feas}}$ to be the vector of all ones. In the $\text{NPG}_{\text{major}}$ for solving the subproblems, we set $M = 4$, $L_{\max} = 10^8$, $L_{\min} = 10^{-8}$, $\tau = 2$, $c = 10^{-4}$, $L_{t,0}^0 = 1$ and for $l \geq 1$,

$$L_{t,l}^0 = \max\left\{\min\left\{\frac{\boldsymbol{s}^{l\top}\boldsymbol{y}^l}{\|\boldsymbol{s}^l\|^2}, L_{\max}\right\}, L_{\min}\right\},$$

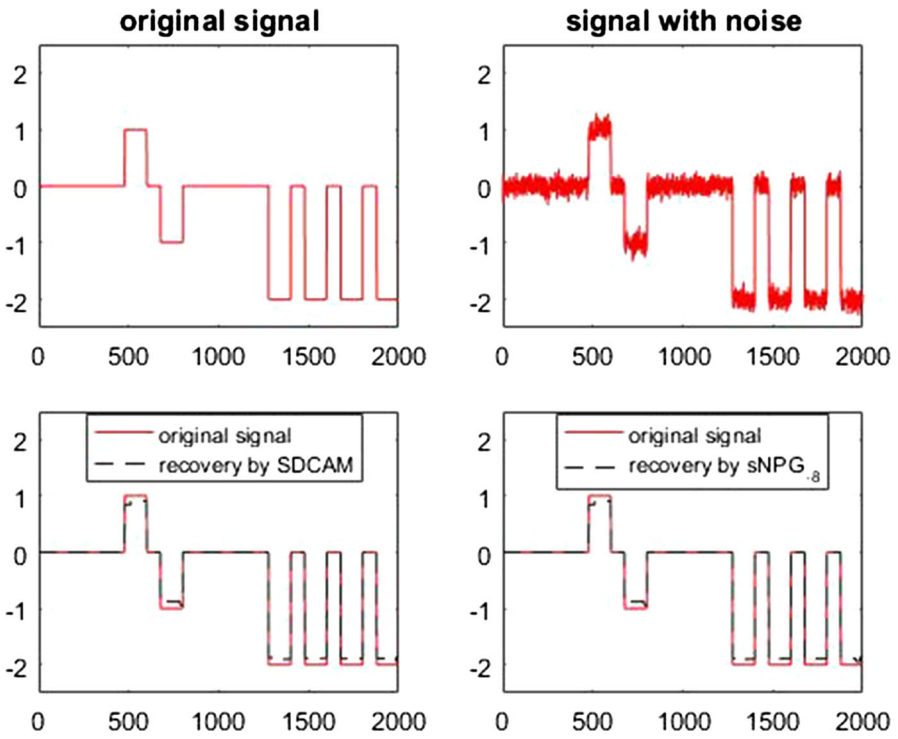

Fig. 1 Recovery comparison for noisy signal

(which is the inverse of the so-called Barzilai-Borwein stepsize) where $s^l = x^{t,l} - x^{t,l-1}$ and $y^l = \nabla h(x^{t,l}) - \nabla h(x^{t,l-1})$. We initialize $\text{NPG}_{\text{major}}$ at x^{feas} and terminate it when the maximum number of iterations exceeds 10000 or

$$\frac{\|x^{t,l} - x^{t,l-1}\|}{\max(\|x^{t,l}\|, 1)} < \epsilon_t / \bar{L}_{t,l-1} \text{ or } \frac{|F_{\lambda_t}(x^{t,l}) - F_{\lambda_t}(x^{t,l-1})|}{\max\{1, |F_{\lambda_t}(x^{t,l})|\}} < 10^{-12},$$

where $\epsilon_0 = 10^{-5}$ and $\epsilon_t = \max\{\epsilon_{t-1}/1.5, 10^{-6}\}$. On the other hand, in sNPG, we also let $\lambda_t = 1/10^{t+1}$ and solve the subproblems using NPG (i.e., $\text{NPG}_{\text{major}}$ applied to (44) with $g = 0$) with the same setting as described above, except that the F_{λ_t} above is replaced by $f_t + Q$ and for $l \geq 1$,

$$L_{t,l}^0 = \begin{cases} \max\left\{\min\left\{\frac{s^{l\top} y^l}{\|s^l\|^2}, L_{\max}\right\}, L_{\min}\right\} & \text{if } s^{l\top} y^l > 10^{-12}, \\ \max\left\{\min\left\{\bar{L}_{t,l-1}/2, L_{\max}\right\}, L_{\min}\right\} & \text{otherwise.} \end{cases}$$

Finally, we terminate SDCAM when $\lambda_t < 10^{-9}$. And for a fair comparison, we consider two different termination criteria for sNPG: $\lambda_t < 10^{-7}$ (sNPG_{-7}) and $\lambda_t < 10^{-8}$ (sNPG_{-8}).

Numerical results: In Table 1, we compare SDCAM, sNPG$_{-7}$ and sNPG$_{-8}$ in terms of the number of iterations (iter),[6] CPU time (CPU) and the terminating function values (fval), averaged over 10 randomly generated instances. One can see that the terminating function values are comparable, and SDCAM is in general faster than sNPG$_{-8}$ and slower than sNPG$_{-7}$. Moreover, SDCAM outperforms the sNPG's slightly in terms of function values when the dimension is relatively low (≤ 4000). To illustrate the ability to recover the original signal, we also plot the original signal, the noisy measurement $\boldsymbol{b}$ and the signals recovered by SDCAM and sNPG$_{-8}$ for a random instance with $n = 2000$ in Fig. 1.

To illustrate intuitively the approximation used in our SDCAM and sNPG, we plot the function $f(x) = |x|^{1/2}$ (in dashed lines), its Moreau envelope and its smoothing function in Fig. 2. One can see that the envelope smooths the original nonsmooth point by a quadratic function. It is a lower approximation of f, while the smoothing function is an upper approximation of f.

5.2 Simultaneously sparse and low rank matrix optimization problem: which constraint should be modeled by P_1?

We consider the following special instance of simultaneously sparse and low rank matrix optimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{X}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{M}\|_F^2 \\ \text{subject to} & \|\text{vec}(\boldsymbol{X})\|_0 \leq s,\ \text{rank}(\boldsymbol{X}) \leq k, \end{array} \tag{43}$$

where $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ is a given noisy matrix, s and k are positive integers. Note that $f(\boldsymbol{X}) := \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{M}\|_F^2$ is level-bounded. Therefore, (43) has at least one solution. Then, as discussed in Example 5, we can apply SDCAM to solving (43) in two different ways by considering, respectively, the two formulations in (40) and (41)[7]: the indicator function $\delta_{|\cdot|_0\leq s}(\cdot)$ is approximated by the Moreau envelope in (40) and the function $\delta_{\text{rank}(\cdot)\leq k}(\cdot)$ is approximated by its Moreau envelope in (41). We call the method based on (40) SDCAM$_r$ and the method based on (41) SDCAM$_s$. In the following experiments, we compare these two methods.

Data generation: We first randomly generate $\boldsymbol{M}_1 \in \mathbb{R}^{m\times k}$ and $\boldsymbol{M}_2 \in \mathbb{R}^{k\times n}$ to have i.i.d. standard Gaussian entries. Then we set $m/10$ random rows of $\boldsymbol{M}_1$ to zero and let $\boldsymbol{M} = \boldsymbol{M}_1\boldsymbol{M}_2 + \sigma\boldsymbol{\Delta}$, where $\sigma > 0$ is a noise factor and $\boldsymbol{\Delta}$ has i.i.d. standard Gaussian entries. We fix $n = 500$, $k = 10$ and $s = mn/10$, and we experiment with $\sigma = 0.005, 0.01, 0.02$ and $m = 1000, 2000, 3000$ below.

Parameter setting: In both SDCAM$_r$ and SDCAM$_s$, we set $\lambda_t = 1/10^{t+1}$ and $\boldsymbol{X}^{\text{feas}} = \boldsymbol{0}$. In the NPG$_{\text{major}}$ for solving the subproblems, we use the same parameter setting as in Sect. 5.1. We initialize both algorithms at $\boldsymbol{X}^{\text{feas}}$ and terminate them when

[6] This refers to the total number of inner iterations.

[7] We would like to point out that we are indeed using Ξ_k in place of $\tilde{\Xi}_k$ in (40) and using S in place of $\tilde{S}$ in (41) in our experiments below. Notice that **A3** is still satisfied because f is level-bounded.

Table 1 Results for SDCAM, sNPG$_{-7}$ and sNPG$_{-8}$ for solving (42)

n	iter			CPU			fval		
	SDCAM	sNPG$_{-7}$	sNPG$_{-8}$	SDCAM	sNPG$_{-7}$	sNPG$_{-8}$	SDCAM	sNPG$_{-7}$	sNPG$_{-8}$
2000	27,796	18,498	23,700	5.8	5.2	8.5	1.77278e+02	1.77294e+02	1.77290e+02
4000	41,686	33,465	43,465	17.0	16.5	29.1	4.95918e+02	4.95929e+02	4.95923e+02
6000	45,573	34,113	44,113	25.6	22.2	38.2	8.49430e+02	8.49420e+02	8.49398e+02
8000	49,089	28,984	38,984	34.7	23.7	43.5	1.32155e+03	1.32160e+03	1.32153e+03
10,000	45,320	37,379	47,379	45.5	38.1	63.1	1.65874e+03	1.65870e+03	1.65864e+03

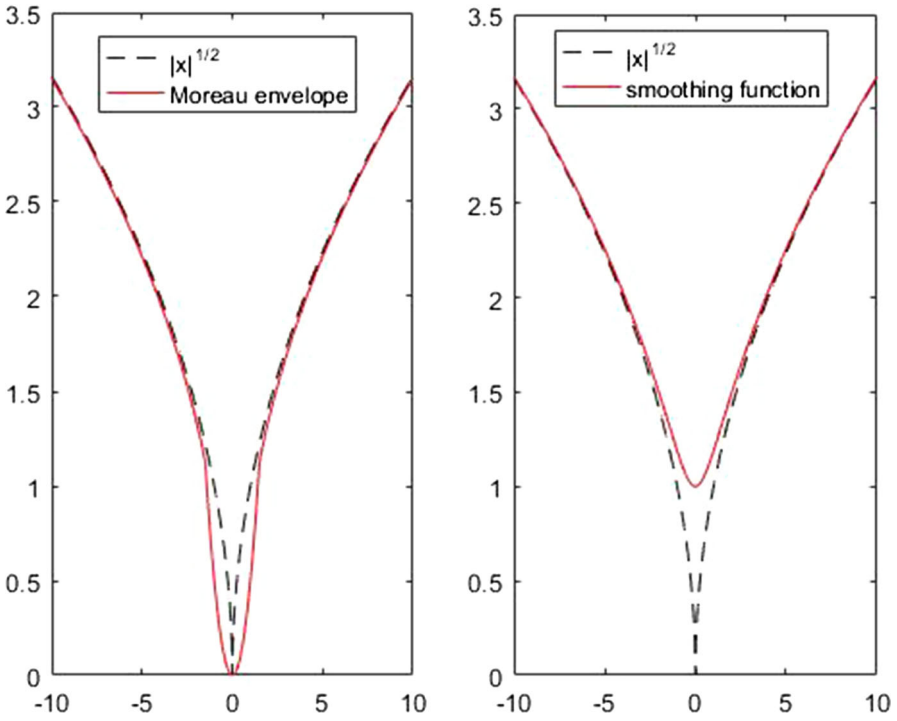

Fig. 2 $|x|^{1/2}$ with its Moreau envelope and smoothing function

$$\mathrm{dist}(\boldsymbol{X}^t, S) \le 10^{-6} \cdot \|\boldsymbol{X}^t\|_F \quad \text{and} \quad \mathrm{dist}(\boldsymbol{X}^t, \Xi_k) \le 10^{-6} \cdot \|\boldsymbol{X}^t\|_F,$$

respectively.

Numerical results: In Table 2, we compare SDCAM$_r$ and SDCAM$_s$ in terms of the number of iterations (iter),[8] CPU time (CPU) and the feasibility violation (vio) (i.e., $\mathrm{dist}(\boldsymbol{X}^t, S)$ and $\mathrm{dist}(\boldsymbol{X}^t, \Xi_k)$, respectively) at termination, averaged over 10 randomly generated instances. One can see that SDCAM$_r$ takes fewer iterations and less time. An intuitive explanation could be that the rank constraint is a more complicated constraint than the sparsity constraint to approximate via "subgradients". Thus, the algorithm SDCAM$_r$ that maintains all its iterates in the rank constraint and then attempts to approximately satisfy the sparsity constraint as the algorithm progresses ends up converging more quickly.

6 Conclusions

In this paper, we propose a successive difference-of-convex approximation method for solving (1). The key idea of this method is to approximate the nonsmooth functions in the objective of (1) by their Moreau envelopes. The approximation function can then

[8] This refers to the total number of inner iterations.

Table 2 Comparison of SDCAM$_r$ and SDCAM$_s$ for solving (43)

σ	m	iter		CPU		vio	
		SDCAM$_r$	SDCAM$_s$	SDCAM$_r$	SDCAM$_s$	SDCAM$_r$	SDCAM$_s$
0.005	1000	41	5597	4.7	378.1	4.7569e−04	1.0515e−04
	2000	12	5298	4.0	647.0	6.7084e−04	1.5247e−04
	3000	12	4618	6.0	862.8	8.2038e-04	1.8857e−04
0.010	1000	4508	7900	379.3	529.2	9.4347e−05	2.1032e−04
	2000	4453	7526	653.6	912.6	1.3412e−04	3.0580e−04
	3000	4428	5721	969.5	1080.6	1.6434e−04	3.7701e−04
0.020	1000	4922	11631	413.7	769.2	1.8985e−04	4.2222e−04
	2000	4634	10267	675.5	1251.3	2.6849e−04	6.1136e−04
	3000	4580	10859	1003.5	2043.0	3.2804e−04	7.5510e−04

be minimized by various proximal gradient methods with majorization techniques such as NPG$_{\text{major}}$ in the appendix, thanks to (6). We prove that the sequence generated by our method is bounded and any accumulation point is a stationary point of (1) under suitable conditions. We also discuss how to apply our method to concrete applications and conduct numerical experiments to illustrate its efficiency.

A Convergence of an NPG method with majorization

In this appendix, we consider the following optimization problem:

$$\min_{\boldsymbol{x}} \quad F(\boldsymbol{x}) = h(\boldsymbol{x}) + P(\boldsymbol{x}) - g(\boldsymbol{x}), \tag{44}$$

where h is an L_h-smooth function, P is a proper closed function with $\inf P > -\infty$ and g is a continuous convex function. We assume in addition that there exists $\boldsymbol{x}^0 \in \operatorname{dom} P$ so that F is continuous in $\Omega(\boldsymbol{x}^0) := \{\boldsymbol{x} : \ F(\boldsymbol{x}) \le F(\boldsymbol{x}^0)\}$ and the set $\Omega(\boldsymbol{x}^0)$ is compact. As a consequence, it holds that $\inf F > -\infty$.

In Algorithm 2 below, we describe an algorithm, the nonmonotone proximal gradient method with majorization (NPG$_{\text{major}}$), for solving (44). We first show that the line-search criterion is well-defined.

Proposition 1 *For each t, the condition (46) is satisfied after at most*

$$\tilde{n} := \max\left\{\left\lceil \frac{\log(L_h + c) - \log(L_{\min})}{\log \tau} \right\rceil, 1\right\}$$

inner iterations, which is independent of t. Consequently, $\{\bar{L}_t\}$ is bounded.

Algorithm 2 The $\mathrm{NPG}_{\mathrm{major}}$ for (44)

Step 0. Choose $\boldsymbol{x}^0 \in \mathrm{dom}\, P$ so that $\Omega(\boldsymbol{x}^0)$ is compact and F is continuous in it. Pick $L_{\max} \geq L_{\min} > 0$, $\tau > 1$, $c > 0$ and an integer $M \geq 0$ arbitrarily. Set $t = 0$.

Step 1. Choose any $L_t^0 \in [L_{\min}, L_{\max}]$ and set $L_t = L_t^0$.

1a) Pick any $\boldsymbol{\zeta}^t \in \partial g(\boldsymbol{x}^t)$. Solve the subproblem

$$\boldsymbol{u} \in \operatorname*{Argmin}_{\boldsymbol{x}} \left\{ (\nabla h(\boldsymbol{x}^t) - \boldsymbol{\zeta}^t)^\top (\boldsymbol{x} - \boldsymbol{x}^t) + \frac{L_t}{2} \|\boldsymbol{x} - \boldsymbol{x}^t\|^2 + P(\boldsymbol{x}) \right\}. \tag{45}$$

1b) If

$$F(\boldsymbol{u}) \leq \max_{[t-M]_+ \leq i \leq t} F(\boldsymbol{x}^i) - \frac{c}{2} \|\boldsymbol{u} - \boldsymbol{x}^t\|^2 \tag{46}$$

is satisfied, then go to **step 2)**.

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

Step 2. If a termination criterion is not met, set $\bar{L}_t = L_t$, $\boldsymbol{x}^{t+1} = \boldsymbol{u}$, $t = t + 1$. Go to **Step 1.**

Proof For each t and $L > 0$, let $\boldsymbol{u}_L^t$ be an arbitrarily fixed element in

$$\operatorname*{Argmin}_{\boldsymbol{x}} \left\{ (\nabla h(\boldsymbol{x}^t) - \boldsymbol{\zeta}^t)^\top (\boldsymbol{x} - \boldsymbol{x}^t) + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}^t\|^2 + P(\boldsymbol{x}) \right\}.$$

Then we have

$$\begin{aligned} F(\boldsymbol{u}_L^t) &\leq h(\boldsymbol{x}^t) + \nabla h(\boldsymbol{x}^t)^\top (\boldsymbol{u}_L^t - \boldsymbol{x}^t) + \frac{L_h}{2} \|\boldsymbol{u}_L^t - \boldsymbol{x}^t\|^2 + P(\boldsymbol{u}_L^t) - g(\boldsymbol{x}^t) \\ &\quad - \boldsymbol{\zeta}^{t\top} (\boldsymbol{u}_L^t - \boldsymbol{x}^t) \\ &= h(\boldsymbol{x}^t) - g(\boldsymbol{x}^t) + (\nabla h(\boldsymbol{x}^t) - \boldsymbol{\zeta}^t)^\top (\boldsymbol{u}_L^t - \boldsymbol{x}^t) + \frac{L_h}{2} \|\boldsymbol{u}_L^t - \boldsymbol{x}^t\|^2 + P(\boldsymbol{u}_L^t) \\ &\leq F(\boldsymbol{x}^t) + \frac{L_h - L}{2} \|\boldsymbol{u}_L^t - \boldsymbol{x}^t\|^2, \end{aligned}$$

where the first inequality holds because of the L_h-smoothness of h, the convexity of g and the fact that $\boldsymbol{\zeta}^t \in \partial g(\boldsymbol{x}^t)$, and the last inequality follows from the definition of $\boldsymbol{u}_L^t$ as a minimizer. Thus, at the t-th iteration, the criterion (46) is satisfied by $\boldsymbol{u} = \boldsymbol{u}_L^t$ whenever $L \geq L_h + c$. Since we have

$$\tau^{\tilde{n}} L_t^0 \geq \tau^{\tilde{n}} L_{\min} \geq L_h + c,$$

we conclude that (46) must be satisfied at or before the $\tilde{n}$-th inner iteration. Consequently, we have $\bar{L}_t \leq \tau^{\tilde{n}} L_{\max}$ for all t. □

The convergence of $\mathrm{NPG}_{\mathrm{major}}$ can now be proved similarly as in [31, Lemma 4].

Proposition 2 *Let $\{\boldsymbol{x}^t\}$ be the sequence generated by $\mathrm{NPG}_{\mathrm{major}}$. Then $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\| \to 0$.*

References

1. Ahn, M., Pang, J.S., Xin, J.: Difference-of-convex learning: directional stationarity, optimality, and sparsity. SIAM J. Optim. **27**, 1637–1665 (2017)
2. Asplund, E.: Differentiability of the metric projection in finite dimensional Euclidean space. Proc. Am. Math. Soc. **38**, 218–219 (1973)
3. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, Berlin (2011)
4. Becker, S., Candès, E.J., Grant, M.: Templates for convex cone problems with applications to sparse signal recovery. Math. Progr. Comput. **3**, 165–218 (2011)
5. Borsdorf, R., Higham, N.J., Raydan, M.: Computing a nearest correlation matrix with factor structure. SIAM J. Matrix Anal. Appl. **31**, 2603–2622 (2010)
6. Brodie, J., Daubechies, I., De Mol, C., Giannone, D., Loris, I.: Sparse and stable Markowitz portfolios. Proc. Natl. Acad. Sci. **106**, 12267–12272 (2009)
7. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**, 717–772 (2009)
8. Chen, C., Li, X., Tolman, C., Wang, S., Ye, Y.: Sparse portfolio selection via quasi-norm regularization. arXiv:1312.6350, (2013)
9. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Progr. **55**, 293–318 (1992)
10. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
11. Gao, Y., Sun, D.: A majorized penalty approach for calibrating rank constrained correlation matrix problems, Technical report, National University of Singapore (2010)
12. Gong, P., Zhang, C., Lu, Z., Huang, J., Ye, J.: A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In: Proceedings of the 30th International Conference on Machine Learning, 37–45 (2013)
13. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**, 2434–2460 (2015)
14. Lu, Z., Li, X.: Sparse recovery via partial regularization: models, theory and algorithms. arXiv:1511.07293 (2015)
15. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM J. Optim. **23**, 2448–2478 (2013)
16. Lu, Z., Zhang, Y., Li, X.: Penalty decomposition methods for rank minimization. Optim. Methods Softw. **30**, 531–558 (2015)
17. Lucet, Y.: Fast Moreau envelope computation I: numerical algorithms. Numer. Algorithms **43**, 235–249 (2006)
18. Markovsky, I.: Structured low-rank approximation and its applications. Automatica **44**, 891–909 (2008)
19. Markowitz, H.: Portfolio selection. J. Financ. **7**, 77–91 (1952)
20. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Progr. **103**, 127–152 (2005)
21. Parekh, A., Selesnick, I.W.: Convex fused Lasso denoising with non-convex regularization and its use for pulse detection. In: Proceedings of IEEE Signal Processing in Medicine and Biology Symposium, 1–6 (2015)
22. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions for linear matrix equations via nuclear norm minimization. SIAM Rev **52**, 471–501 (2010)
23. Richard, E., Savalle, P.-A., Vayatis, N.: Estimation of simultaneously sparse and low rank matrices. arXiv:1206.6474 (2012)
24. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
25. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
26. Slawski, M., Hein, M.: Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization. Electron. J. Stat. **7**, 3004–3056 (2013)
27. Thiao, M., Pham, D.T., Le Thi, H.A.: A DC programming approach for sparse eigenvalue problem. In: Proceedings of the 27th International Conference on Machine Learning, 1063–1070 (2010)
28. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)
29. Tibshirani, R., Taylor, J.: The solution path of the generalized Lasso. Ann. Stat. **39**, 1335–1371 (2011)
30. Tono, K., Takeda, A., Gotoh, J.: Efficient DC algorithm for constrained sparse optimization. arXiv:1701.08498 (2017)

31. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**, 2479–2493 (2009)
32. Yu, Y.L.: Better approximation and faster algorithm using the proximal average. Adv. Neural Inf. Process. Syst. **26**, 458–466 (2013)
33. Yu, Y.L., Zheng, X., Marchetti-Bowick, M., Xing, E.: Minimizing nonconvex non-separable functions. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics **38**, 1107–1115 (2015)

Affiliations

Tianxiang Liu[1] · Ting Kei Pong[1] · Akiko Takeda[2,3]

Tianxiang Liu
tiskyliu@polyu.edu.hk

Ting Kei Pong
tk.pong@polyu.edu.hk

[1] Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong

[2] Department of Creative Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

[3] RIKEN Center for Advanced Intelligence Project, 1-4-1, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan