Computational Optimization and Applications (2019) 72:45–85
https://doi.org/10.1007/s10589-018-0036-9

On monotone and primal-dual active set schemes for ℓ^p-type problems, $p \in (0, 1]$

Daria Ghilli[1] · **Karl Kunisch**[1,2]

Received: 14 May 2017 / Published online: 4 October 2018

Abstract
Nonsmooth nonconvex optimization problems involving the ℓ^p quasi-norm, $p \in (0, 1]$, of a linear map are considered. A monotonically convergent scheme for a regularized version of the original problem is developed and necessary optimality conditions for the original problem in the form of a complementary system amenable for computation are given. Then an algorithm for solving the above mentioned necessary optimality conditions is proposed. It is based on a combination of the monotone scheme and a primal-dual active set strategy. The performance of the two algorithms is studied by means of a series of numerical tests in different cases, including optimal control problems, fracture mechanics and microscopy image reconstruction.

Keywords Nonsmooth nonconvex optimization · Active-set method · Monotone algorithm · Optimal control problems · Image reconstruction · Fracture mechanics

Mathematics Subject Classification 49K99 · 49M05 · 65K10

Supported by the ERC advanced Grant 668998 (OCLOC) under the EU's H2020 research programme.

✉ Daria Ghilli
daria.ghilli@uni-graz.at

Karl Kunisch
karl.kunisch@uni-graz.at

[1] Institute of Mathematics and Scientific Computing, Karl-Franzens-Universität, Heinrichstrasse 36, 8010 Graz, Austria

[2] Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, 4040 Linz, Austria

1 Introduction

We consider the following nonconvex nonsmooth optimization problem

$$\min_{x\in\mathbb{R}^d} J(x) = \frac{1}{2}|Ax - b|_2^2 + \beta|\Lambda x|_p^p, \tag{1.1}$$

where $A \in \mathbb{M}^{m\times d}$, $\Lambda \in \mathbb{M}^{r\times d}$, $b \in \mathbb{R}^m$, $p \in (0, 1]$ and $\beta \in \mathbb{R}^+$. Here

$$|x|_p = \left(\sum_{k=1}^{d} |x_k|^p\right)^{\frac{1}{p}},$$

which is a norm for $p = 1$ and a quasi-norm for $0 < p < 1$.

Optimization of problems as (1.1) arises frequently in many applications as an efficient way to extract the essential features of generalized solutions. In particular, many problems in sparse learning and compressed sensing can be written as (1.1) with $\Lambda = I$, I being the identity (see e.g. [11,41] and the references therein). In image analysis, ℓ^p-regularizers as in (1.1) have recently been proposed as nonconvex extensions of the total generalized variation (TGV) regularizer used to reconstruct piecewise smooth functions (e.g. in [24,42]). Also, the use of ℓ^p-functionals with $p \in (0, 1)$ is of particular importance in fracture mechanics (see [43]). Recently, sparsity techniques have been investigated also by the optimal control community, see e.g. [9,22,27,32,48]. The literature on sparsity optimization problems as (1.1) is rapidly increasing, here we mention also [1,6,18,44].

The nonsmoothness and nonconvexity make the study of problems as (1.1) both an analytical and a numerical challenge. Many numerical techniques have been developed when $\Lambda = I$ (e.g. in [20,27,29,30]) and attention has recently been given to the case of more general operators, here we mention e.g. [24,36,42] and we refer to the end of the introduction for further details. However, the presence of the matrix inside the ℓ^p-term combined with the nonconvexity and nonsmoothness remains one main issue in the development of numerical schemes for (1.1).

In the present work, we first propose a monotone algorithm to solve a regularized version of (1.1). The scheme is based on an iterative procedure solving a modified problem where the singularity at the origin is regularized. The convergence of this algorithm and the monotone decay of the cost during the iterations are proved. Then its performance is successfully tested in four different situations, a time-dependent control problem, a fracture mechanics example for cohesive fracture models, an M-matrix example, and an elliptic control problem.

We also focus on the investigation of suitable necessary optimality conditions for solving the original problem. Relying on an augmented Lagrangian formulation, optimality conditions of complementary type are derived. For this purpose we consider the case where Λ is a regular matrix, since in the general case the optimality conditions of complementary type are not readily obtainable. An active set primal-dual strategy which exploits the particular form of these optimality conditions is developed. A new particular feature of our method is that at each iteration level the monotone scheme

is used in order to solve the nonlinear equation satisfied by the non zero components. The convergence of the active set primal-dual strategy is proved in the case $\Lambda = I$ under a diagonal dominance condition. Finally the algorithm was tested on the same time-dependent control problem as the one analysed for the monotone scheme as well as for a microscopy image reconstruction example. In all the above mentioned examples the matrix inside the ℓ^p-term appears as a discretized gradient with very different purposes, e.g. as a regularization term in imaging and with modelling purposes in fracture mechanics.

Similar type of algorithms were proposed in [27] and [20] for problems as (1.1) in case of no matrix inside the ℓ^p-term and in the infinite dimensional sequence spaces ℓ^p, with $p \in [0, 1]$. In particular in [27] a primal-dual active set strategy has been studied for the case $p = 0$, its convergence was proved under a diagonal dominance condition and its performance was tested in three different situations. In the case $p \in (0, 1]$ a monotone convergent scheme was proposed but no numerical tests were made. Then inspired by [27], the two authors of the present paper developed in [20] a primal-dual active set strategy for $p \in (0, 1]$ and tested its performance in diverse test cases. Note that in [20] the convergence of the primal-dual strategy was not investigated. The monotone and primal-dual active set monotone algorithm studied in the present paper are inspired by the schemes proposed respectively in [27] and [20], but with the main novelties that now we treat the case of a regular matrix in the ℓ^p-term and we provide diverse numerical tests for both the schemes. Moreover, we prove the convergence of the primal-dual active set strategy. Note that the monotone scheme has not been tested in the earlier papers.

Let us recall some further literature concerning ℓ^p, $p \in (0, 1]$ sparse regularizers. Iteratively reweighted least-squares algorithms with suitable smoothing of the singularity at the origin were analysed in [13,34,35]. In [37] a unified convergence analysis was given and new variants were also proposed. An iteratively reweighted ℓ_1 algorithm ([8]) was developed in [14] for a class of nonconvex ℓ^2-ℓ^p problems, with $p \in (0, 1)$. A generalized gradient projection method for a general class of nonsmooth nonconvex functionals and a generalized iterated shrinkage algorithm are analysed respectively in [6] and in [52]. Also, in [44] a surrogate functional approach combined with a gradient technique is proposed. However, all the previous works do not investigate the case of a linear operator inside the ℓ^p-term.

Then in [42] an iteratively reweighted convex majorization algorithm is proposed for a class of nonconvex problems including the ℓ^p, $p \in (0, 1]$ regularizer acting on a linear map. However, an additional assumption of Lipschitz continuity of the objective functional is required to establish convergence of the whole sequence generated by the algorithm. Nonconvex TV^p-models with $p \in (0, 1)$ for image restoration are studied in [24] by a Newton-type solution algorithm for a regularized version of the original problem.

We mention also [30], where a primal-dual active set method is studied for problems as in (1.1) with $\Lambda = I$ for a large class of penalties including also the ℓ^p, with $p \in [0, 1)$. A continuation strategy with the respect to the regularization parameter β is proposed and the convergence of the primal-dual active set strategy coupled with the continuation strategy is proved. However, in [30], differently from the present work, the nonlinear problem arising at each iteration level of the active set scheme is not

investigated. Moreover, in [30] the matrix A has normalized column vectors, whereas in the present work A is a general matrix.

Finally, in [36] an alternating direction method of multipliers (ADMM) is studied in the case of a regular matrix inside the ℓ^p-term, optimality conditions were derived and convergence was proved. We underline that in [36] the linear map inside the ℓ^p-term has to be surjective, which can be restrictive per applications. Although the ADMM in [36] is also deduced from an augmented Lagrangian formulation, we remark that the optimality conditions of [36] are of a different nature than ours and hence the two approaches cannot readily be compared. We refer to Remark 4 for a more detailed explanation.

Concerning the general importance of ℓ^p-functionals with $p \in (0, 1)$, numerical experience has shown that their use can promote sparsity better than the ℓ^1-norm (see [10,19,49]), e.g. allowing possibly a smaller number of measurements in feature selection and compressed sensing (see also [11,12,40]). Moreover, many works demonstrated empirically that nonconvex regularization terms in total variation-based image restoration provide better edge preservation than the ℓ^1-regularization (see [5,39,40,45]). Also, the use of nonconvex optimization can be considered from natural image statistics [26] and it appears to be more robust with respect to heavy-tailed distributed noise (see e.g. [51]).

The paper is organized as follows. In Sect. 2 we present our proposed monotone algorithm and we prove its convergence. In Sect. 3 we report our numerical results for the four test cases mentioned above. In Sect. 4 we derive the necessary optimality conditions for (1.1), we describe our primal-dual active set strategy and prove convergence in the case $\Lambda = I$. Finally in Sect. 5 we report the numerical results obtained by testing the active set monotone algorithm in the two situations mentioned above.

2 Existence and monotone algorithm for a regularized problem

For convenience of exposition, we recall the problem under consideration

$$\min_{x\in\mathbb{R}^d} J(x) = \frac{1}{2}|Ax - b|_2^2 + \beta|\Lambda x|_p^p, \tag{2.1}$$

where $A \in \mathbb{M}^{m\times d}$, $\Lambda \in \mathbb{M}^{r\times d}$, $b \in \mathbb{R}^m$, $p \in (0, 1]$ and $\beta \in \mathbb{R}^+$.

Throughout this section we assume

$$\mathrm{Ker}(A) \cap \mathrm{Ker}(\Lambda) = \{0\}. \tag{2.2}$$

The first result is an existence result for (2.1).

Theorem 1 *For any $\beta > 0$, there exists a solution to 2.1.*

Proof Since J is bounded from below, existence will follow from the continuity and coercivity of J. Thus we prove that J is coercive, that is, $|J(x_k)|_2 \to +\infty$ whenever $|x_k|_2 \to +\infty$ for some sequence $\{x_k\} \subset \mathbb{R}^d$. By contradiction, suppose that $|x_k|_2 \to$

$+\infty$ and $J(x_k)$ is bounded. For each k, let $x_k = t_k z_k$ be such that $t_k \geq 0$, $x_k \in \mathbb{R}^d$ and $|z_k|_2 = 1$. Since $t_k \to +\infty$, $p < 2$, we have for k sufficiently large

$$0 \leq \frac{1}{2t_k^2}|Ax_k|_2^2 + \beta\frac{1}{t_k^p}|\Lambda x_k|_p^p \leq \left(\frac{1}{2} + \beta\right)\frac{1}{t_k^p}\left(|Ax_k|_2^2 + |\Lambda x_k|_p^p\right) \to 0$$

and hence

$$\lim_{k\to+\infty} \frac{1}{2}|Az_k|_2^2 + \beta|\Lambda z_k|_p^p = 0.$$

By compactness, the sequence $\{z_k\}$ has an accumulation point $\bar{z}$ such that $|\bar{z}| = 1$ and $\bar{z} \in \mathrm{Ker}(A) \cap \mathrm{Ker}(\Lambda)$, which contradicts 2.2. □

Following [27], in order to overcome the singularity of $(|s|^p)' = \frac{ps}{|s|^{2-p}}$ near $s = 0$, we consider for $\varepsilon > 0$ the following regularized version of (2.1)

$$\min_{x\in\mathbb{R}^d} J_\varepsilon(x) = \frac{1}{2}|Ax - b|_2^2 + \beta\Psi_\varepsilon(|\Lambda x|^2), \tag{2.3}$$

where for $t \geq 0$

$$\Psi_\varepsilon(t) = \begin{cases} \frac{p}{2}\frac{t}{\varepsilon^{2-p}} + (1 - \frac{p}{2})\varepsilon^p & \text{for } \ 0 \leq t \leq \varepsilon^2 \\ t^{\frac{p}{2}} & \text{for } \ t \geq \varepsilon^2, \end{cases} \tag{2.4}$$

and $\Psi_\varepsilon(|\Lambda x|^2)$ is short for $\sum_{i=1}^r \Psi_\varepsilon(|(\Lambda x)_i|^2)$.

Remark 1 Notice that by the coercivity of the functional J in (2.1), the coercivity of J_ε and hence existence for (2.3) follow as well.

The necessary optimality condition for (2.3) is given by

$$A^*Ax + \Lambda^*\frac{\beta p}{\max(\varepsilon^{2-p}, |\Lambda x|^{2-p})}\Lambda x = A^*b,$$

where the max-operation is interpreted coordinate-wise. We set $y = \Lambda x$. Then

$$A^*Ax + \Lambda^*\frac{\beta p}{\max(\varepsilon^{2-p}, |y|^{2-p})}y = A^*b. \tag{2.5}$$

In order to solve (2.5), the following iterative procedure is considered:

$$A^*Ax^{k+1} + \Lambda^*\frac{\beta p}{\max(\varepsilon^{2-p}, |y^k|^{2-p})}y^{k+1} = A^*b, \tag{2.6}$$

where we denote $y^k = \Lambda x^k$, and the second addend is short for the vectors with l^{th}-component $\sum_{i=1}^d (\Lambda^*)_{li}\frac{\beta p}{\max(\varepsilon^{2-p}, |y_i^k|^{2-p})}y_i^{k+1}$.

We have the following convergence result.

Theorem 2 *For $\varepsilon > 0$, let $\{x_k\}$ be generated by (2.6). Then, $J_\varepsilon(x_k)$ is strictly monotonically decreasing, unless there exists some k such that $x^k = x^{k+1}$ and x^k satisfies the necessary optimality condition (2.5). Moreover every cluster point of x^k, of which there exists at least one, is a solution of (2.5).*

Proof The proof follows similar arguments to that of Theorem 4.1, [27]. Multiplying (2.6) by $x^{k+1} - x^k$, we get

$$\begin{aligned}&\frac{1}{2}|Ax^{k+1}|^2 - \frac{1}{2}|Ax^k|^2 + \frac{1}{2}|A(x^{k+1} - x^k)|^2 \\ &\quad + \beta p \left\langle \frac{1}{\max(\varepsilon^{2-p}, |y^k|^{2-p})} y^{k+1}, y^{k+1} - y^k \right\rangle \\ &= \langle A^* b, x^{k+1} - x^k \rangle.\end{aligned}$$

Note that

$$\left\langle \frac{1}{\max(\varepsilon^{2-p}, |y^k|^{2-p})} y^{k+1}, y^{k+1} - y^k \right\rangle = \frac{1}{2} \sum_{i=1}^{d} \frac{\left(|y_i^{k+1}|^2 - |y_i^k|^2 + |y_i^{k+1} - y_i^k|^2\right)}{\max(\varepsilon^{2-p}, |y_i^k|^{2-p})} \tag{2.7}$$

and

$$\frac{1}{\max(\varepsilon^{2-p}, |y_i^k|^{2-p})} \frac{p}{2} (|y_i^{k+1}|^2 - |y_i^k|^2) = \Psi_\varepsilon'(|y_i^k|^2)(|y_i^{k+1}|^2 - |y_i^k|^2). \tag{2.8}$$

Since $t \to \Psi_\varepsilon(t)$ is concave, we have

$$\Psi_\varepsilon(|y_i^{k+1}|^2) - \Psi_\varepsilon(|y_i^k|^2) - \frac{1}{\max(\varepsilon^{2-p}, |y_i^k|^{2-p})} \frac{p}{2} (|y_i^{k+1}|^2 - |y_i^k|^2) \le 0. \tag{2.9}$$

Then, using (2.7), (2.8), (2.9), we get

$$\begin{aligned}J_\varepsilon(x^{k+1}) + \frac{1}{2}|A(x^{k+1} - x^k)|_2^2 + \frac{1}{2} \sum_{i=1}^{d} \frac{\beta p}{\max(\varepsilon^{2-p}, |y_i^k|^{2-p})} |y_i^{k+1} - y_i^k|^2 \\ \le J_\varepsilon(x^k).\end{aligned} \tag{2.10}$$

From (2.10) and the coercivity of J_ε it follows that $\{x^k\}_{k=1}^\infty$ and thus $\{y^k\}_{k=1}^\infty$ are bounded. Then, from (2.10), there exists a constant $\kappa > 0$ such that

$$J_\varepsilon(x^{k+1}) + \frac{1}{2}|A(x^{k+1} - x^k)|_2^2 + \kappa |y^{k+1} - y^k|_2^2 \le J_\varepsilon(x^k), \tag{2.11}$$

from which we conclude the first part of the theorem. From (2.11), we conclude that

$$\sum_{k=0}^{\infty} |A(x^{k+1} - x^k)|_2^2 + |y^{k+1} - y^k|_2^2 < \infty. \tag{2.12}$$

Since $\{x^k\}_{k=1}^{\infty}$ is bounded, there exists a subsequence and $\bar{x} \in \mathbb{R}^d$ such that $x^{k_l} \to \bar{x}$. By (2.12) and (2.2) we have that $x^{k_l+1} \to \bar{x}$. Then, passing to the limit with respect to k in (2.6), we get that $\bar{x}$ is a solution to (2.5). □

In the following proposition we establish the convergence of (2.3)–(2.1) as ε goes to zero.

Proposition 1 *Let $\{x_\varepsilon\}_{\varepsilon>0}$ be solution to (2.3). Then any cluster point of $\{x_\varepsilon\}_{\varepsilon>0}$ as $\varepsilon \to 0^+$, of which there exists at least one, is a solution of (2.1).*

Proof First note that $J_\varepsilon(x_\varepsilon) \le J_\varepsilon(0) = \frac{1}{2}|b|_2^2 + \left(1 - \frac{p}{2}\right)\varepsilon^p$, and hence $J_\varepsilon(x_\varepsilon)$ is uniformly bounded for $\varepsilon \in (0, 1]$. Let us argue that $\{x_\varepsilon\}$ is uniformly bounded for $\varepsilon \to 0^+$. Assume to the contrary that this is not the case and that $\lim_{k\to\infty} |x_{\varepsilon_k}|_2 \to \infty$, where $\varepsilon_k \to 0$ as $k \to \infty$. We show that this leads to a contradiction to (2.2). Proceeding similarly as in the proof of Theorem 1 we define $\{t_k\}$ and $\{z_k\}$ such that $x_{\varepsilon_k} = t_k z_k$ with $t_k \ge 0$, and $|z_k| = 1$ and $z = \lim_k z_k$. Note that $\lim_k t_k = \infty$. In the following computation k is sufficiently large, so that $t_k \ge 1$. Then

$$\begin{aligned}\frac{1}{t_k^p}\left(|Ax_{\varepsilon_k}|_2^2 + \beta\Psi_{\varepsilon_k}(|\Lambda x_{\varepsilon_k}|^2)\right) &\ge |Az_k|_2^2 + \beta\sum_{i=1}^{r}\frac{1}{t_k^p}\Psi_{\varepsilon_k}(|(\Lambda x_{\varepsilon_k})_i|^2) \\ &\ge |Az_k|_2^2 + \beta\sum_{i=1}^{r}\frac{1}{t_k^p}|(\Lambda x_{\varepsilon_k})_i|^p, \end{aligned} \tag{2.13}$$

where in the last inequality we used that $\Psi_\varepsilon(t) \ge t^{\frac{p}{2}}$.

Since the left-hand side of (2.13) converges to 0 for $k \to \infty$ we have

$$0 = \lim_k |Az_k|_2^2 + \beta\sum_{i=1}^{r}|(\Lambda z_k)_i|^p = |Az_k|_2^2 + \beta|\Lambda z|_p^p.$$

Hence $z \in \ker(A) \cap \ker(\Lambda) = \{0\}$, which contradicts that $|z|_2 = 1$.

Then by the uniform boundedness of $\{x_\varepsilon\}_\varepsilon$ there exists a subsequence and $\bar{x} \in \mathbb{R}^d$ such that $x_{\varepsilon_l} \to \bar{x}$. Since $\{x_\varepsilon\}_\varepsilon$ solves (2.3), by letting $\varepsilon \to 0$ and using the definition of Ψ_ε, we easily get that $\bar{x}$ is a solution of (2.1). □

3 Monotone algorithm: numerical results

The focus of this section is to investigate the performance of the monotone algorithm in practice. For this purpose we choose four problems with matrices A of very different

structure: a time-dependent optimal control problem, a fracture mechanics example, the M matrix and a stationary optimal control problem.

3.1 The numerical scheme

For further references it is convenient to recall the algorithm in the following form (see Algorithm 1). Note that a continuation strategy with respect to the parameter ε is performed. The initialization and range of ε-values is described for each class of problems below.

The system in (3.2) is solved though the MATLAB function *mldivive* (that is, the *backslash* command). The algorithm stops when the ℓ^∞-norm of the residue of (2.5) is $O(10^{-3})$ in all the examples, except the fracture problem, where it is $O(10^{-15})$. At this instance, the ℓ^2-residue is typically much smaller. Thus, we find an approximate solution of the ε-regularized optimality condition (2.5). The initialization x^0 is chosen in the following way

$$x^0 = (A^*A + 2\beta\Lambda^*\Lambda)^{-1}A^*b, \tag{3.1}$$

that is, x^0 is chosen as the solution of the problem (2.1) where the ℓ^p-term is replaced by the ℓ^2-norm. Our numerical experience shows that for some values of β the previous initialization is not suitable, that is, the residue obtained is too big. In order to get a lower residue, we successfully tested a continuation strategy with respect to increasing β-values.

Algorithm 1 Monotone algorithm + ε-continuation strategy

1: Initialize ε^0, x^0 and set $y^0 = \Lambda x^0$. Set $k = 0$;
2: **repeat**
3: Solve for x^{k+1}

$$A^*Ax^{k+1} + \Lambda^* \frac{\beta p}{\max(\varepsilon^{2-p}, |y^k|^{2-p})}\Lambda x^{k+1} = A^*b. \tag{3.2}$$

4: Set $y^{k+1} = \Lambda x^{k+1}$.
5: Set $k = k + 1$.
6: **until** the stopping criterion is fulfilled.
7: Reduce ε and repeat 2.

In the presentation of our numerical results, the total number of iterations shown in the tables takes into account the continuation strategy with respect to ε. However, it does not take into account the continuation with respect to β. We remark that in all the experiments presented in the following sections, the value of the functional for each iterations was checked to be monotonically decreasing accordingly to Theorem 2.

The following notation will hold for the rest of the paper. For $x \in \mathbb{R}^d$ we will denote $|x|_0 = \#\{i \,:\, |x_i| > 10^{-10}\}$, $|x|_0^c = \#\{i \,:\, |x_i| \le 10^{-10}\}$, and by $|x|_2$ the euclidean norm of x.

3.2 Time-dependent control problem

We consider the linear control system

$$\frac{d}{dt}y(t) = \mathcal{A}y(t) + Bu(t), \quad y(0) = 0,$$

that is,

$$y(T) = \int_0^T e^{\mathcal{A}(T-s)}Bu(s)ds, \tag{3.3}$$

where the linear closed operator $\mathcal{A}$ generates a C_0-semigroup $e^{\mathcal{A}t}$, $t \geq 0$ on the state space X. More specifically, we consider the one-dimensional controlled heat equation for $y = y(t, x)$:

$$y_t = y_{xx} + b_1(x)u_1(t) + b_2(x)u_2(t), \quad x \in (0, 1), \tag{3.4}$$

with homogeneous boundary conditions $y(t, 0) = y(t, 1) = 0$ and thus $X = L^2(0, 1)$. The differential operator $\mathcal{A}y = y_{xx}$ is discretized in space by the second order finite difference approximation with $n = 49$ interior spatial nodes ($\Delta x = \frac{1}{50}$). We use two time dependent controls $\overrightarrow{u} = (u_1, u_2)$ with corresponding spatial control distributions b_i chosen as step functions:

$$b_1(x) = \chi_{(.2,.3)}, \quad b_2(x) = \chi_{(.6,.7)}.$$

The control problem consists in finding the control function $\overrightarrow{u}$ that steers the state $y(0) = 0$ to a neighbourhood of the desired state y_d at the terminal time $T = 1$. We discretize the problem in time by the mid-point rule, i.e.

$$A\overrightarrow{u} = \sum_{k=1}^{m} e^{\mathcal{A}\left(T-t_k-\frac{\Delta t}{2}\right)}(B\overrightarrow{u})_k \Delta t, \tag{3.5}$$

where $\overrightarrow{u} = (u_1^1, \ldots, u_1^m, u_2^1, \ldots u_2^m)$ is a discretized control vector whose coordinates represent the values at the mid-point of the intervals (t_k, t_{k+1}). Note that in (3.5) we denote by B a suitable rearrangement of the matrix B in (3.3) with some abuse of notation. A uniform step-size $\Delta t = \frac{1}{50}$ ($m = 50$) is utilized. The solution of the control problem is based on the sparsity formulation (2.1), where Λ is the backward difference operator acting independently on each component of the control, that is, $\Lambda = m(I_2 \otimes D)$ where I_2 is the 2×2 identity matrix and $D : \mathbb{R}^m \to \mathbb{R}^m$ is as follows

$$D = \begin{pmatrix} 1 & 0 & 0 \ldots 0 \\ -1 & 1 & 0 \ldots 0 \\ & & \\ 0 & \ldots 0 & -1 \; 1 \end{pmatrix}. \tag{3.6}$$

Table 1 Sparsity in a time-dependent control problem, $p = .5$, mesh size $h = \frac{1}{50}$. Results obtained by Algorithm 1

β	10^{-3}	10^{-2}	10^{-1}	1
No. of iterates	630	635	29	19
$\lvert Du_2\rvert_0^c$	97	99	100	100
$\lvert Du_2\rvert_p^p$	158	16.7	$6 * 10^{-5}$	10^{-4}
Residue	$3 * 10^{-3}$	$2 * 10^{-3}$	$1.2 * 10^{-3}$	$2.5 * 10^{-10}$
Sp	97	99	100	100

Also, b in (2.1) is the discretized target function chosen as the Gaussian distribution $y_d(x) = 0.4 \exp(-70(x - .7)^2))$ centered at $x = .7$. That is, we apply our algorithm for the discretized optimal control problem in time and space where x from (2.1) is the discretized control vector $u \in \mathbb{R}^{2m}$ which is mapped by A to the discretized output y at time 1 by means of (3.5). Moreover b from (2.1) is the discretized state y_d with respect to the spatial grid Δx. The parameter ε was initialized with 10^{-3} and decreased down to 10^{-8}. Note that, since the second control distribution is well within the support of the desired state y_d, we expect the authority of this control to be stronger than that of the first one, which is away from the target.

In Table 1 we report the results of our tests for $p = .5$ for β incrementally increasing by factor of 10 from 10^{-3} to 1. We report only the values for the second control u_2 since the first control u_1 is always zero. In the third row we see that $(|Du_2|_0)^c$ increases with β, consistent with our expectation. Note also that the quantity $|Du_2|_p^p$ decreases for β increasing.

For any $i = 1, \ldots, m$, we say that i is a singular component of the vector Du_2 if $i \in \{i : |(Du_2)_i| < \varepsilon\}$. In particular, note that the singular components are the ones where the ε-regularization is most influential. In the sixth row of Table 1 we show their number at the end of the ε-path following scheme (denoted by Sp) and we observe that it coincides with the quantity $|Du_2|_0^c$, which is reassuring the validity of our ε-strategy.

The algorithm was also tested for values of p near to 1, e.g. for $p = .9$. The results obtained show a less piecewise constant behaviour of the solution with respect to the ones for $p = .5$. Finally, we remark that if we change the initialization (3.1), the method converges to the same solution with no remarkable modifications in the number of iterations.

3.3 Quasi-static evolution of cohesive fracture models

In this section we focus on a modelling problem for quasi-static evolutions of cohesive fractures. This kind of problems requires the minimization of an energy functional, which has two components: the elastic energy and the cohesive fracture energy. The underlying idea is that the fracture energy is released gradually with the growth of the crack opening. The cohesive energy, denoted by θ, is assumed to be a monotonic non-decreasing function of the jump amplitude of the displacement, denoted by $[\![u]\!]$.

Cohesive energies were introduced independently by Dugdale [16] and Barenblatt [3], we refer to [43] for more details on the models. Let us just remark that the two models differ mainly in the evolution of the derivative $\theta'([\![u]\!])$, that is, the *bridging force*, across a crack amplitude $[\![u]\!]$. In Dugdale's model this force keeps a constant value up to a critical value of the crack opening and then drops to zero. In Barenblatt's model, the dependence of the force on $[\![u]\!]$ is continuous and decreasing.

In this section we test the ℓ^p-term $0 < p < 1$ as a model for the cohesive energy. In particular, the cohesive energy is not differentiable in zero and the bridging force goes to infinity when the jump amplitude goes to zero. Note also that the bridging force goes to zero when the jump amplitude goes to infinity.

Let us introduce all the elements that we need for the rest of the section. We consider the one-dimensional domain $\Omega = [0, 2l]$ with $l > 0$ and we denote by $u : \Omega \to \mathbb{R}$ the displacement function. The deformation of the domain is given by an external force which we express in terms of an external displacement function $g : \Omega \times [0, T] \to \mathbb{R}$. We require that the displacement u coincides with the external deformation, that is

$$u|_{\partial\Omega} = g|_{\partial\Omega}.$$

We denote by Γ the point of the (potential) crack, which we chose as the midpoint $\Gamma = l$ and by $\theta([\![u]\!])_\Gamma$ the value of the cohesive energy θ on the crack amplitude of the displacement $[\![u]\!]$ on Γ. Since we are in a quasi-static setting, we introduce the time discretization $0 = t_0 < t_1 < \cdots < t_T = T$ and look for the equilibrium configurations which are minimizers of the energy of the system. This means that for each $i \in \{0, \dots, T\}$ we need to minimize the energy of the system

$$J(u) = \frac{1}{2}\int_{\Omega\setminus\Gamma} |\nabla u|^2 dx + \beta\theta([\![u]\!])_\Gamma$$

with respect to a given boundary datum g:

$$u^* \in \operatorname*{argmin}_{u=g(t_i) \text{ on } \partial\Omega} J(u),$$

where $\beta > 0$ in $J(u)$ is a material parameter. In particular, we consider the following type of cohesive energy

$$\theta([\![u]\!]) = |[\![u]\!]|^p,$$

for $p \in (0, 1)$. We divide Ω into $2N$ intervals and approximate the displacement function with a function u_h that is piecewise linear on $\Omega\setminus\Gamma$ and has two degrees of freedom on Γ to represent correctly the two lips of the fracture, denoting with u_N^- the degree on $[0, l]$ and u_N^+ the one on $[l, 1]$. We discretize the problem in the following way

$$J_h(u_h) = \frac{1}{2}\sum_{i=1}^{2N} \frac{N}{l}|u_i - u_{i-1}|^2 + \beta|[\![u_N]\!]|^p, \tag{3.7}$$

where if $i \leq N$ we identify $u_N = u_N^-$ while for $i > N$, $u_N = u_N^+$. We remark that the jump of the displacement is not taken into account in the sum, and the gradient of u is approximated with finite difference of first order. The Dirichlet condition is applied on $\partial\Omega = \{0, 2l\}$ and the external displacement is chosen as

$$u(0,t) = 0, \quad u(2l,t) = 2lt.$$

To enforce the boundary condition in the minimization process, we add it to the energy functional as a penalization term. Hence, we solve the following unconstrained minimization problem

$$\min \frac{N}{2l}|Au_h - g|_2^2 + \beta|[\![u_N]\!]|^p, \tag{3.8}$$

where the operator $A \in \mathbb{R}^{(2N+1)\times(2N+1)}$ is given by

$$A = \begin{bmatrix} & \bar{D} & \\ 0 & \dots & 0 \ \gamma \end{bmatrix}.$$

Here $\bar{D} \in \mathbb{R}^{2N\times(2N+1)}$ denotes the backward finite difference operator $D : \mathbb{R}^{2N+1} \to \mathbb{R}^{2N+1}$ without the $N+1$ row, where D is defined in (3.6). Moreover $g \in \mathbb{R}^{2N+1}$ in (3.8) is given by $g = (0, \dots, \gamma 2lt_i)'$ and γ is the penalization parameter. To compute the jump between the two lips of the fracture, we introduce the operator $D_f : \mathbb{R}^{2N+1} \to \mathbb{R}$ defined as $D_f = (0, \dots, -1, 1, 0, \dots, 0)$ where -1 and 1 are respectively in the N and $N+1$ positions. Then we write the functional (3.8) as follows

$$\min \frac{N}{2l}|Au_h - g|_2^2 + \beta|D_f u|^p, \tag{3.9}$$

Note that KerA $= 0$, hence assumption (2.2) is satisfied and existence of a minimizer for (3.9) is guaranteed.

Our numerical experiments were conducted with a discretization in $2N$ intervals with $N = 100$ and a prescribed potential crack $\Gamma = 0.5$. The time step in the time discretization of $[0, T]$ with $T = 3$ is set to $dt = 0.01$. The parameters of the energy functional $J_h(u_h)$ are set to $\beta = 1$, $\gamma = 50$. The parameter ε is decreased from 10^{-1} to 10^{-12}.

In Fig. 1 we report three time frames to represent the evolutions of the crack obtained with Algorithm 1 for two different values of p, that is, $p = .01, .1$ respectively. Each time frame consists of three different time steps (t_1, t_2, t_3), where t_2, t_3 are chosen as the first instant where the prefracture and the fracture appear. The evolution presents the three phases that we expect from a cohesive fracture model:

- *Pure elastic deformation* in this case the jump amplitude is zero and the gradient of the displacement is constant in $\Omega\backslash\Gamma$;
- *Prefracture* the two lips of the fracture do not touch each other, but they are not free to move. The elastic energy is still present.

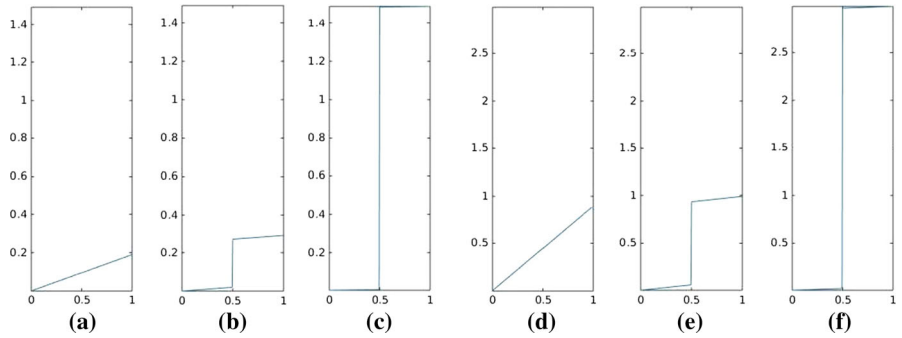

Fig. 1 Three time-step evolution of the displacement for $p = .01$, $t = .2, .3, 1.5$ (left), $p = .1$, $t = .9, 1, 3$ (right). Results obtained by Algorithm 1. **a** $t = 0.2$, **b** $t = 0.3$, **c** $t = 1.5$, **d** $t = 0.9$, **e** $t = 1$, **f** $t = 3$

– *Fracture* the two parts are free to move. In this final phase the gradient of the displacement (and then the elastic energy) is zero.

Moreover we remark that the formation of the crack is anticipated for smaller values of p. As we see in Fig. 1, for $p = .01$ prefracture and fracture are reached at $t = .3$ and $t = 1.5$ respectively. As p is increased to $p = .1$, prefracture and fracture occur at $t = 1$ and $t = 3$ respectively. Finally we remark that in our experiments the residue is $O(10^{-16})$ and the number of iterations is small, e.g. 12, 15 for $p = .01, .1$ respectively.

3.4 M-matrix

We consider

$$\min_{x\in\mathbb{R}^{d^2}} \frac{1}{2}|Ax - b|_2^2 + \beta|\Lambda x|_p^p, \tag{3.10}$$

where A is the backward finite difference gradient

$$A = (d+1)\begin{pmatrix} G_1 \\ G_2 \end{pmatrix}, \tag{3.11}$$

with $G_1 \in \mathbb{R}^{d(d+1)\times d^2}$, $G_2 \in \mathbb{R}^{d(d+1)\times d^2}$ given by

$$G_1 = I \otimes D, \quad G_2 = D \otimes I.$$

Here I is the $d \times d$ identity matrix, $\otimes$ denotes the tensor product, and $D \in \mathbb{R}^{(d+1)\times d}$ is given by

$$D = \begin{pmatrix} 1 & 0 & 0 \dots & 0 \\ -1 & 1 & 0 \dots & 0 \\ & & & \\ 0 & \dots & 0 & -1 & 1 \\ 0 & \dots & 0 & 0 & -1 \end{pmatrix}. \tag{3.12}$$

Table 2 M-matrix example, $\Lambda = (d+1)[D_1; D_2]$, $p = .1$, mesh size $h = \frac{1}{64}$. Results obtained by Algorithm 1

β	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10
No. of iterates	1701	2469	3929	4254	14	7
$\lvert\Lambda x\rvert_0^c$	16	103	791	5384	7938	7938
$\lvert\Lambda x\rvert_p^p$	$6 * 10^3$	$5.8 * 10^3$	$5 * 10^3$	$2.4 * 10^3$	584	464
Residue	$2.7 * 10^{-7}$	$5.5 * 10^{-6}$	$9 * 10^{-5}$	$9 * 10^{-4}$	$3 * 10^{-12}$	$2.7 * 10^{-12}$
Sp	247	696	2097	5599	7938	7938

Then A^*A is an M matrix coinciding with the 5-point star discretization on a uniform mesh on a square of the Laplacian with Dirichlet boundary conditions. Note that (3.10) can be equivalently expressed as

$$\min_{x \in \mathbb{R}^{d \times d}} \frac{1}{2} |Ax|_2^2 - \langle x, f \rangle + \beta |\Lambda x|_p^p, \tag{3.13}$$

where $f = A^*b$. If $\beta = 0$ this is the discretized variational form of the elliptic equation

$$-\Delta y = f \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega. \tag{3.14}$$

For $\beta > 0$ the variational problem (3.13) gives a solution piecewise constant enhancing behaviour.

Our tests were conducted with f chosen as discretization of $f = 10x_1 \sin(5x_2) \cos(7x_1)$ and

$$\Lambda = (d+1) \begin{pmatrix} D_1 \\ D_2 \end{pmatrix}, \tag{3.15}$$

where $D_1 \in \mathbb{R}^{d^2 \times d^2}$, $D_2 \in \mathbb{R}^{d^2 \times d^2}$ are defined as follows

$$D_1 = I \otimes D, \quad D_2 = D \otimes I, \tag{3.16}$$

and $D \in \mathbb{R}^{d \times d}$ is the backward difference operator defined in (3.12) without the $d+1$-row. The parameter ε was initialized with 10^{-1} and decreased to 10^{-6}.

In Tables 2 we show the performance of Algorithm 1 for $p = .1$, $h = 1/64$ as mesh size and β incrementally increasing by factor of 10 from 10^{-4} to 10. In Fig. 2 we report the graphics of the solutions for different values of β between .01 and .3 where most changes occur in the graphics.

We observe significant differences in the results with respect to different values of β. Consistently with our expectations, $|\Lambda x|_0^c$ increases with β (see the third row of Table 2). For example, for $\beta = 1, 10$, we have $|\Lambda x|_0^c = 7938$, or equivalently, $|\Lambda x|_0 = 0$, that is, the solution to (3.13) is constant. Moreover the fourth row shows that $|\Lambda x|_p^p$ decreases when β increases.

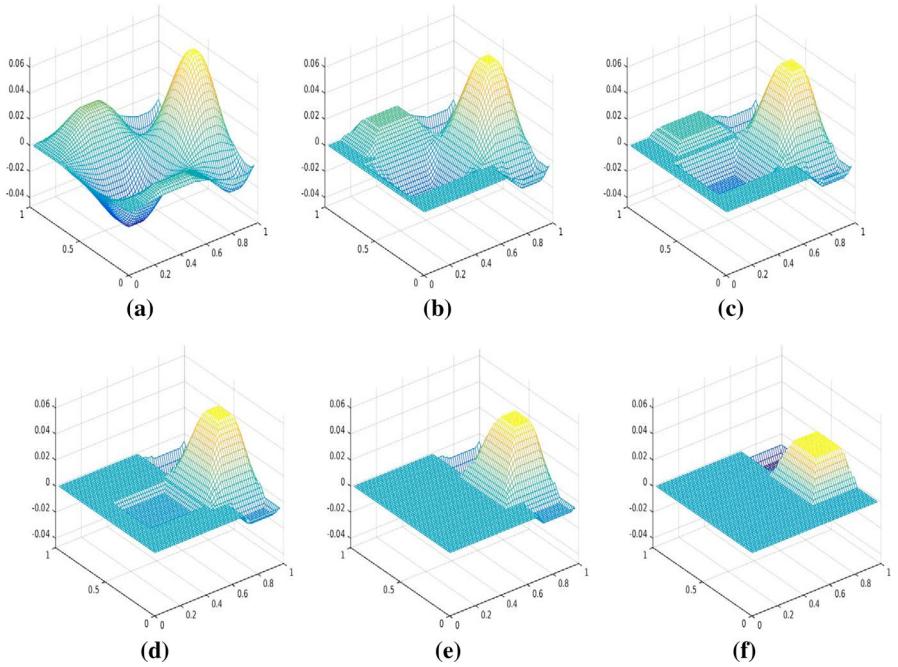

Fig. 2 Solution of the M-matrix problem, $p = .1$, $\Lambda = (d+1)[D_1; D_2]$, mesh size $h = \frac{1}{64}$. Results obtained by Algorithm 1. **a** $\beta = 0.01$, **b** $\beta = 0.05$, **c** $\beta = 0.08$, **d** $\beta = 0.12$, **e** $\beta = 0.15$, **f** $\beta = 0.3$

The fifth row exhibits the ℓ^∞ norm of the residue, which is $O(10^{-4})$ for all the considered β. We remark that the number of iterations is sensitive with respect to β, in particular it increases when β is increasing from 10^{-4} to 10^{-1} and then it decreases significantly for $\beta = 1, 10$.

The algorithm was also tested for different values of p. The results obtained show dependence on p, in particular $|\Lambda x|_0^c$ decreases as p is increasing. For example, for $p = .5$ and $\beta = .1$ we have $|\Lambda x|_0^c = 188$, $|\Lambda x|_p^p = 528$.

In the sixth row of Table 2 we show the number of singular components of the vector Λx at the end of the ε-path following scheme, that is, $Sp := \#\{i \mid |(\Lambda x)_i| < \varepsilon\}$. For most values of β, we note that Sp is comparable to $|\Lambda x|_0^c$. This again confirms that the ε-strategy is effective.

Finally, we remark that if we modify the initialization (3.1), the method converges to the same solution with no remarkable modifications in the number of iterations, which is a sign for the global nature of the algorithm.

Remark 2 The algorithm was also tested in the following two particular cases: $\Lambda = I$, where I is the identity matrix of size d^2, and $\Lambda = (d+1)D_1$, where D_1 is as in (3.16).

In the case $\Lambda = I$ the variational problem (3.13) for $\beta > 0$ gives a sparsity enhancing solution for the elliptic equation (3.14), that is, the displacement y will be 0 when the forcing f is small. Indeed, in this case we have sparsity of the solution increasing with β. Also, the residue is $O(10^{-8})$ and the number of iterations is considerably smaller than in the case Λ is as in (3.15).

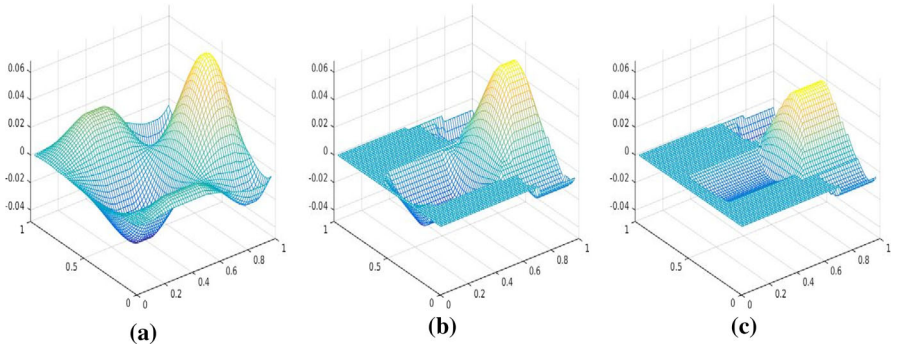

Fig. 3 Solution of the M-matrix problem, $p = .1$, $\Lambda = (d+1)D_1$, mesh size $h = \frac{1}{64}$. Results obtained by Algorithm 1. **a** $\beta = 0.01$, **b** $\beta = 0.1$, **c** $\beta = 0.3$

For $\Lambda = (d+1)D_1$ we show the graphics in Fig. 3. Comparing the graphs for $\beta = .3$ in Figs. 2 and 3, we can find subdomains where the solution is only unidirectionally piecewise constant in Fig. 3 and piecewise constant in Fig. 2. The number of iterations, $|\Lambda x|_0^c$, $|\Lambda x|_p^p$ and the residue are comparable to the ones of Table 2.

3.5 Elliptic control problem

We consider the following two-dimensional control problem

$$\inf \frac{1}{2}|y - y_d|_2^2 + \beta|\nabla u|_p^p, \quad p \in (0, 1], \tag{3.17}$$

where we minimize over $u \in L^p(\Omega)$ such that $\nabla u \in L^p(\Omega)$, Ω is the unit square, $y_d \in L^2(\Omega)$ is a given target function, and $y \in L^2(\Omega)$ satisfies

$$\begin{cases} -\Delta y = u & \text{in } \Omega \\ y = 0 & \text{in } \partial\Omega. \end{cases} \tag{3.18}$$

We discretize (3.17) by the following $\frac{1}{d}$-mesh size discretized minimization problem

$$\min_{u \in \mathbb{R}^{d^2}} \frac{1}{2}|Eu - b|_2^2 + \beta|\Lambda u|_p^p, \tag{3.19}$$

where $E = (A^*A)^{-1}$, A is as in (3.11) (that is, A^*A is the 5-point star discretization on a uniform mesh on a square of the Laplacian with Dirichlet boundary condition), $\Lambda = (d+1)\begin{pmatrix} D_1 \\ D_2 \end{pmatrix}$ is as in Sect. 3.4 and b is the discretized target function.

Table 3 Sparsity in an elliptic control problem, $p = .1$, mesh size $h = \frac{1}{64}$. Results obtained by Algorithm 1

β	10^{-3}	10^{-2}	10^{-1}	1
No. of iterates	102	119	5204	10440
$\lvert\Lambda u\rvert_0^c$	799	1486	1673	2376
$\lvert\Lambda u\rvert_p^p$	$3.2 * 10^4$	$2.6 * 10^4$	$2.6 * 10^4$	$1.2 * 10^4$
Residue	$1.6 * 10^{-5}$	$2.4 * 10^{-4}$	$2 * 10^{-3}$	$7 * 10^{-3}$

For numerical reasons, in order to avoid the inversion of the matrix A^*A, we multiply the necessary optimality condition (2.5) by $(E^{-1})^*$ and we get

$$Eu + (E^{-1})^*\Lambda^* \frac{\beta p}{\max(\varepsilon^{2-p}, |y|^{2-p})} y^1 = b, \tag{3.20}$$

where $y = \Lambda u$. We introduce

$$z = Eu, \quad p = (\Lambda^* N \Lambda)u,$$

where we denote by N the diagonal matrix with i-entry $(N)_{ii} = \frac{\beta p}{\max(\varepsilon^{2-p}, |y_i|^{2-p})}$, $i = 1, \dots d^2$. Since $E^{-1} = A^*A$, we can express (3.20) in the form

$$\begin{cases} A^*Az = u \\ A^*Ap = b - z \\ (\Lambda^* N \Lambda)u = p. \end{cases} \tag{3.21}$$

To solve (3.21) the following iteration procedure is used

$$\begin{pmatrix} I & 0 & A^*A \\ 0 & \Lambda^* N^k \Lambda & -I \\ A^*A & -I & 0 \end{pmatrix} \begin{pmatrix} z^{k+1} \\ u^{k+1} \\ p^{k+1} \end{pmatrix} = \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix} \tag{3.22}$$

where we denote by N^k the diagonal matrix with i-entry $(N^k)_{ii} = \frac{\beta p}{\max(\varepsilon^{2-p}, |y_i^k|^{2-p})}$ for $i = 1, \dots, d^2$ and $y^k = \Lambda u^k$. Note that the system matrix (3.22) is symmetric.

In our tests the target b is chosen as the image through E of the linear interpolation inside $[.2, .8] \times [.2, .8] \setminus [.3, .7] \times [.3, .7]$ of the step function $1000\chi_{[.3,.7]\times[.3,.7]}$. The parameter ε was initialized with 10^{-1} and decreased to 10^{-6}. The system in (3.22) is solved though the MATLAB function *mldivive* (that is, the *backslash* command). In Table 3 we report the results of our test for $h = \frac{1}{64}$, $p = .1$ and β incrementally increasing by factor of 10 from 10^{-3} to 1. As expected, when β increases, $\lvert\Lambda u\rvert_0^c$ increases and $\lvert\Lambda u\rvert_p^p$ decreases. In Fig. 4 we show the graphics of the solution for different values of β, thus showing the enhancing piecewise constant behaviour of the solution.

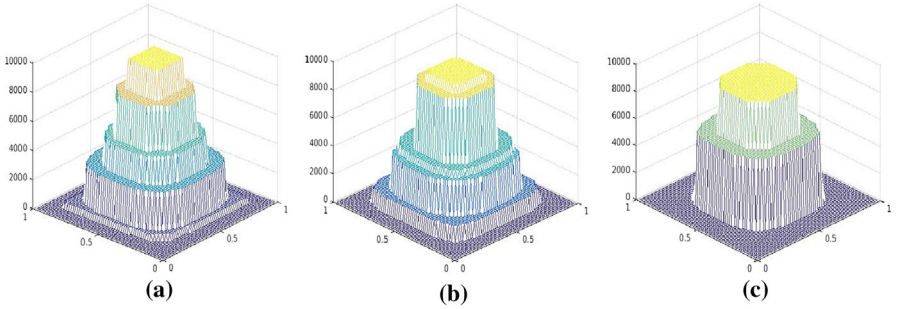

Fig. 4 Solution of the elliptic control problem, $p = .1$, mesh size $h = \frac{1}{64}$. Results obtained by Algorithm 1. **a** $\beta = 0.01$, **b** $\beta = 0.1$, **c** $\beta = 1$

From our tests we conclude that the monotone algorithm is reliable to find a solution of the ε-regularized optimality condition (2.5) for a diverse spectrum of problems. It is also stable with respect to the choice of initial conditions. According to the last rows of Tables 1, 2 and 3 we have that $\#\{i \,|\, |(\Lambda x)_i| \leq 10^{-10}\}$ is typically very close to the number of singular components at the end of the ε-path following scheme. Depending on the choice of β the algorithm requires on the order of $O(10^2)$ to $O(10^3)$ iterations to reach convergence. In the following sections we aim at analysing an alternative algorithm for which the iteration number is smaller, despite the fact that the convergence can be proved only in special cases.

4 The active set monotone algorithm for the optimality conditions

In the following we discuss an algorithm which aims at finding a solution of the original unregularized problem

$$\min_{x \in \mathbb{R}^d} J(x) = \frac{1}{2}|Ax - b|_2^2 + \beta |\Lambda x|_p^p, \tag{4.1}$$

where $A \in \mathbb{M}^{m \times d}$, $b \in \mathbb{R}^m$, $p \in (0, 1]$ and $\beta \in \mathbb{R}^+$ are as in Sect. 2. Differently from the previous sections we now assume throughout that

$$\Lambda \in \mathbb{M}^{d \times d} \text{ is a regular matrix.} \tag{4.2}$$

Existence for the problem (4.1) follows from Theorem 1.

First necessary optimality conditions for problem (4.1) in the form of a complementary systems are derived and a sufficient condition for the uniqueness for solutions to this system are established. Subsequently an active-set strategy is proposed relying on this form of the optimality conditions.

4.1 Necessary optimality conditions

For any matrix $A \in \mathbb{M}^{m\times d}$, we denote by A_i the i-th column of A. We have the following necessary optimality conditions for a global minimizer of (4.1).

Theorem 3 *Assume that (4.2) holds, let $\bar{x}$ be a global minimizer of (4.1), and denote $\bar{y} = \Lambda\bar{x}$. Then we have*

$$\begin{cases} A^*(Ax-b)+\Lambda^*\lambda=0, & \\ (\Lambda\bar{x})_i=0 & \text{if } \left||\tilde{A}_i|_2^2\bar{y}_i+\lambda_i\right| < \mu_i, \\ |(\Lambda x)_i|>0 \text{ and } \lambda_i=\frac{\beta p(\Lambda\bar{x})_i}{|(\Lambda\bar{x})_i|^{2-p}} & \text{if } \left||\tilde{A}_i|_2^2\bar{y}_i+\lambda_i\right| > \mu_i, \end{cases} \tag{4.3}$$

where $\tilde{A}=A\Lambda^{-1}$, $\mu_i=\beta^{\frac{1}{2-p}}(2-p)(2(1-p))^{-\frac{1-p}{2-p}}|\tilde{A}_i|_2^{1-\frac{p}{2-p}}$. *If* $\left||\tilde{A}_i|_2^2\bar{y}_i+\lambda_i\right| = \mu_i$, *then* $(\Lambda\bar{x})_i=0$ *or* $(\Lambda\bar{x})_i=\left(\frac{2\beta(1-p)}{|\tilde{A}_i|_2^2}\right)^{\frac{1}{2-p}}$ *sgn* $(|\tilde{A}_i|_2^2\bar{y}_i+\lambda_i)$.

Proof Note that if $\bar{x}$ is a global minimizer of (4.1), then $\bar{y}=\Lambda\bar{x}$ is a global minimizer of

$$\min_{y\in\mathbb{R}^d}\frac{1}{2}|\tilde{A}y-b|_2^2+\beta|y|_p^p, \tag{4.4}$$

where $\tilde{A}=A\Lambda^{-1}$. Then, by the same arguments as in [27], Theorem 2.2 applied to the functional (4.4), we get the following property of global minimizers

$$\begin{cases} \bar{y}_i=0 & \text{if } |\langle\tilde{A}_i,f_i\rangle|<\mu_i, \\ |y_i|>0 \text{ and } \langle\tilde{A}_i,\tilde{A}\bar{y}-b\rangle+\frac{\beta p\bar{y}_i}{|\bar{y}_i|^{2-p}}=0 & \text{if } |\langle\tilde{A}_i,f_i\rangle|>\mu_i, \end{cases} \tag{4.5}$$

where $f_i=b-\tilde{A}y+\tilde{A}_i\bar{y}_i$ and $\mu_i=\beta^{\frac{1}{2-p}}(2-p)(2(1-p))^{-\frac{1-p}{2-p}}|\tilde{A}_i|_2^{1-\frac{p}{2-p}}$. Moreover, if $|\langle\tilde{A}_i,f_i\rangle|=\mu_i$, then $\bar{y}_i=0$ or $\bar{y}_i=\left(\frac{2\beta(1-p)}{|\tilde{A}_i|_2^2}\right)^{\frac{1}{2-p}}$ sgn $(\langle\tilde{A}_i,f_i\rangle)$. We introduce the multiplier λ and we write (4.5) in the following way

$$\begin{cases} \tilde{A}^*(\tilde{A}y-b)+\lambda=0, & \\ \bar{y}_i=0 & \text{if } \left||\tilde{A}_i|_2^2\bar{y}_i+\lambda_i\right|<\mu_i, \\ |y_i|>0 \text{ and } \lambda_i=\frac{\beta p\bar{y}_i}{|\bar{y}_i|^{2-p}} & \text{if } \left||\tilde{A}_i|_2^2\bar{y}_i+\lambda_i\right|>\mu_i. \end{cases} \tag{4.6}$$

Then the optimality conditions (4.3) follows from (4.6) with $\bar{y}=\Lambda\bar{x}$. The equality conditions follow similarly by $\bar{y}=\Lambda\bar{x}$ and the first equation in (4.6). □

Remark 3 We remark that Theorem 3 still hold when considering (4.1) in the infinite dimensional sequence spaces ℓ^p in the case $\Lambda=I$.

From Theorem 3 we can deduce a lower bound for the nonzero components of Λx following the arguments in [27], Corollary 2.1. The result of Corollary 1 are of the same order with respect to β, p, and A, compared to the existing lower bound for ℓ^2-ℓ^p problems derived in the literature, see e.g. [15]. However, the proof of [15] is different from the one presented in [27] and relies mainly on second order necessary optimality conditions.

Corollary 1 *If* $(\Lambda\bar{x})_i \neq 0$, *then* $|(\Lambda\bar{x})_i| \geq \left(\frac{2\beta(1-p)}{|(A\Lambda^{-1})_i|_2^2}\right)^{\frac{1}{2-p}}$.

4.2 The augmented Lagrangian formulation and the primal-dual active set strategy

The active set strategy can be motivated by the following augmented Lagrangian formulation for problem (4.1). Let P be a nonnegative self-adjoint matrix P, satisfying

$$\langle (A^T A + \eta P)x, x\rangle \geq \xi |x|_2^2 \tag{4.7}$$

for some $\eta, \xi > 0$, independent of $x \in \mathbb{R}^d$. We set

$$B_i = |(\bar{A}\Lambda^{-1})_i|_2^2, \text{ where } \bar{A} = \begin{pmatrix} A \\ (\eta P)^{\frac{1}{2}} \end{pmatrix}, \tag{4.8}$$

and let B denote the diagonal invertible operator with entries B_i. Thus, if A is nearly singular, we use $\eta > 0$ and the functional $\frac{\eta}{2}\langle x, Px\rangle$ to regularize (4.1). Consider the associated augmented Lagrangian functional

$$\begin{aligned} L(x, y, \lambda) = \frac{1}{2}|Ax - b|_2^2 + \frac{\eta}{2}\langle Px, x\rangle + \beta \sum_{i=1}^{d} |y_i|^p \\ + \sum_{i=1}^{d} \left[\frac{B_i}{2}|y_i - (\Lambda x)_i|^2 + \lambda_i((\Lambda x)_i - y_i) \right]. \end{aligned} \tag{4.9}$$

Given x, λ, we first minimize L coordinate-wise with respect to y. For this purpose we consider

$$\begin{aligned} &\beta|y_i|^p + \frac{B_i}{2}|y_i - (\Lambda x)_i|^2 + \lambda_i((\Lambda x)_i - y_i) \\ &= \beta|y_i|^p + \frac{B_i}{2}\left(y_i^2 - 2y_i\left((\Lambda x)_i + \frac{\lambda_i}{B_i}\right)\right) + \frac{B_i(\Lambda x)_i^2}{2} + \lambda_i(\Lambda x)_i \\ &= \beta|y_i|^p + \frac{B_i}{2}\left[y_i - \left((\Lambda x)_i + \frac{\lambda_i}{B_i}\right)\right]^2 - \frac{B_i}{2}\left[(\Lambda x)_i + \frac{\lambda_i}{B_i}\right]^2 + \frac{B_i(\Lambda x)_i^2}{2} \\ &\quad + \lambda_i(\Lambda x)_i = \beta|y_i|^p + \frac{1}{2}\left[B_i^{\frac{1}{2}} y_i - \left(B_i^{\frac{1}{2}}(\Lambda x)_i + \frac{\lambda_i}{B_i^{\frac{1}{2}}}\right)\right]^2 - \frac{\lambda_i^2}{2B_i}. \end{aligned} \tag{4.10}$$

Then, by Theorem 3, the augmented Lagrangian L can be minimized coordinate-wise with respect to y by considering the expressions $\beta|y_i|^p + \frac{1}{2}\left[B_i^{\frac{1}{2}} y_i - \left(B_i^{\frac{1}{2}}(\Lambda x)_i + \frac{\lambda_i}{B_i^{\frac{1}{2}}}\right)\right]^2$ to obtain

$$y_i = \Phi(x,\lambda)_i = \begin{cases} |y_i| > 0 \text{ and } B_i y_i + \frac{\beta p y_i}{|y_i|^{2-p}} = B_i(\Lambda x)_i + \lambda_i & \text{if } |B_i(\Lambda x)_i + \lambda_i| > \mu_i, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\mu_i = \beta^{\frac{1}{2-p}}(2-p)(2(1-p))^{-\frac{1-p}{2-p}} B_i^{\frac{1-p}{2-p}}. \tag{4.11}$$

Given y, λ, we minimize L at x to obtain

$$A^*(Ax-b) + \eta P x + \Lambda^* B(\Lambda x - y) + \Lambda^*\lambda = 0,$$

where B is the diagonal operator with entries B_i. Thus, the augmented Lagrangian method [28] uses the updates:

$$\begin{cases} A^*(Ax^{n+1}-b) + \eta P x^{n+1} + \Lambda^* B(\Lambda x^{n+1} - y^n) + \Lambda^*\lambda^n = 0, \\ y^{n+1} = \Phi(x^{n+1}, \lambda^n), \\ \lambda^{n+1} = \lambda^n + B(\Lambda x^{n+1} - y^{n+1}). \end{cases} \tag{4.12}$$

If it converges, i.e. $x^n \to x$, $y^n \to \Lambda x^n$, and $\lambda^n \to \lambda$, then

$$\begin{cases} A^*(Ax-b) + \eta P x + \Lambda^*\lambda = 0, \\ (\Lambda x)_i = 0 & \text{if } |B_i y_i + \lambda_i| \le \mu_i, \\ |(\Lambda x)_i| > 0 \text{ and } \lambda_i = \frac{\beta p (\Lambda x)_i}{|(\Lambda x)_i|^{2-p}} & \text{if } |B_i y_i + \lambda_i| > \mu_i, \end{cases} \tag{4.13}$$

which is the optimality condition for $J_P(x) = \min_{x\in\mathbb{R}^d} \frac{1}{2}|Ax-b|_2^2 + \beta|\Lambda x|_p^p + \frac{\eta}{2}\langle x, Px\rangle$, compare (4.3).

Motivated by the form of the optimality conditions (4.13) obtained by the augmented Lagrangian formulation, we formulate in Algorithm 2 a primal-dual active set strategy for the following system

$$\begin{cases} A^*(Ax-b) + \eta P x + \Lambda^*\lambda = 0, \\ (\Lambda x)_i = 0 & \text{if } |B_i y_i + \lambda_i| \le \mu_i, \\ \lambda_i = \frac{\beta p (\Lambda x)_i}{\max(\varepsilon^{2-p}, |(\Lambda x)_i|^{2-p})} & \text{if } |B_i y_i + \lambda_i| > \mu_i, \end{cases} \tag{4.14}$$

where $\varepsilon > 0$ in the third equation is a fixed small parameter. Note that according to Corollary 1 system (4.14) coincides with (4.13) if $\varepsilon < \left(\frac{2\beta(1-p)}{|(A\Lambda^{-1})_i|_2^2}\right)^{\frac{1}{2-p}}$ for all i with $|B_i y_i + \lambda_i| > \mu_i$. The motivation for the parameter ε is to avoid the expression $\frac{\beta p(\Lambda x^{n+1})_i}{|(\Lambda x^{n+1})_i|^{2-p})|}$, where computing the update x^{n+1}, λ^{n+1} is specified next.

Algorithm 2 Primal-dual active set strategy

1: Initialize λ^0, x^0. Set $y^0 = \Lambda x^0$. Set $n = 0$.
2: **repeat**
3: Solve for (x^{n+1}, λ^{n+1})

$$A^*(Ax^{n+1} - b) + \eta P x^{n+1} + \Lambda^* \lambda^{n+1} = 0, \tag{4.15}$$

where

$$(\Lambda x^{n+1})_i = 0 \quad \text{if } i \in \{i : |B_i y_i^n + \lambda_i^n| \le \mu_i\} \tag{4.16}$$

$$\lambda_i^{n+1} = \frac{\beta p(\Lambda x^{n+1})_i}{\max(\varepsilon^{2-p}, |(\Lambda x^{n+1})_i|^{2-p})} \quad \text{if } i \in \{i : |B_i y_i^n + \lambda_i^n| > \mu_i\}. \tag{4.17}$$

4: Set $y^{n+1} = \Lambda x^{n+1}$, $n = n + 1$.
5: **until** the stopping criterion is fulfilled.

System (4.15)–(4.17) in Algorithm 2 is implicit. It can be solved, for example, by an iterative procedure based on Algorithm 1. The final scheme resulting from the combination of Algorithm 1 and 2 is described in Algorithm 3 in the following section.

Remark 4 The specific choice (4.8) of the penalization constant B_i in the augmented Lagrangian method (4.9) is crucial to ensure the convergence of the method to the optimality condition (4.3). Indeed, contrary to what would be tempting following the standard penalization technique procedure, here B_i has to be taken exactly as in (4.8) and in particular it cannot be taken too large. To make this evident, we propose to look at the following one-dimensional example. Suppose we want to minimize

$$\frac{1}{2}|x - b|_2^2 + \beta |x|_p^p \tag{4.18}$$

for $p \in (0, 1]$, $\beta > 0$. By Theorem 3, the optimality condition is

$$\begin{cases} x = 0 & \text{if } b < \mu, \\ |x| > 0 \text{ and } x - b + \beta p \frac{x}{|x|^{2-p}} = 0 & \text{if } b > \mu, \end{cases} \tag{4.19}$$

where we denote $\mu := d_{\beta,p}$ and $d_{\beta,p} = \beta^{\frac{1}{2-p}}(2-p)(2(1-p))^{-\frac{1-p}{2-p}}$ is given in (4.3). Consider for $c > 0$ the augmented Lagrangian

$$L(x, y) = \frac{1}{2}|x - b|_2^2 + \beta|y|^p + \frac{c}{2}|x - y|_2^2 + \langle \lambda, y - x\rangle. \tag{4.20}$$

For y fixed, we minimize with respect to x to obtain

$$x - b + c(x - y) - \lambda = 0.$$

Then, given x fixed, we minimize with respect to y the expression $\beta|y|^p + \frac{c}{2}|x-y|^2 + \lambda y$. First note that

$$\beta|y|^p + \frac{c}{2}|x - y|^2 + \lambda y = \beta|y|^p + \frac{c}{2}\left|\left(x - \frac{\lambda}{c}\right) - y\right|^2 - \frac{\lambda^2}{2c} + \lambda x.$$

Then by Theorem 3 we obtain

$$\begin{cases} y = 0 & \text{if } c|(x - \frac{\lambda}{c})| < \mu_c, \\ c(y - \left(x - \frac{\lambda}{c}\right)) + \beta p \frac{y}{|y|^{2-p}} = 0 & \text{if } c|(x - \frac{\lambda}{c})| > \mu_c, \end{cases} \tag{4.21}$$

where $\mu_c = d_{\beta,p}\sqrt{c}^{\frac{(2-2p)}{2-p}}$. Finally we minimize with respect to λ and we get $x = y$. Then we obtain the following optimality conditions of complementary type

$$\begin{cases} x - b + c(x - y) - \lambda = 0, & \\ y = 0, & \text{if } c\left(x - \frac{\lambda}{c}\right) < \mu_c, \\ c(y - \left(x - \frac{\lambda}{c}\right)) + \beta p \frac{y}{|y|^{2-p}} = 0, & \text{if } c\left(x - \frac{\lambda}{c}\right) > \mu_c, \\ x = y. & \end{cases} \tag{4.22}$$

Note that if we use the formula (4.8) for the penalization term of the augmented Lagrangian formulation (4.9), we get $c = 1$ and then (4.22) coincides with (4.19), as desired. For $c > 1$, on the other hand, the convergence to the solution of the optimality conditions (4.19) is not guaranteed. Indeed, if $c > 1$ and $b \in (\mu, \mu_c)$ then $\mu < \mu_c$, and from the augmented Lagrangian formulation (4.20), we get that $y = 0$, $x = 0$, $\lambda = -b$ is a solution to (4.22). But $b > \mu$ implies that $x = 0$ is not a solution of (4.19). This shows that, if the penalization constant c does not satisfy (4.8), then the augumented Lagragian method based on the formulation (4.20) might not converge to the original problem (4.18).

Remark 5 Now consider the following optimality condition for the problem (4.20)

$$0 \in \partial\left(\frac{1}{2}|x - b|_2^2 + \beta|y|^p + \frac{c}{2}|x - y|_2^2 + \langle \lambda, y - x\rangle)\right), \tag{4.23}$$

where ∂ denotes the *limiting subdifferential*. For a definition of the limiting subdifferential we refer to [36], where similar problems to the ones studied in the present paper are solve through optimality conditions as in (4.23), see also the end of this remark for more details.

Since

$$\partial |y|^p = \begin{cases} py|y|^{p-2} & \text{if } y \neq 0, \\ \mathbb{R} & \text{otherwise,} \end{cases} \tag{4.24}$$

the optimality condition (4.23) explicitly reads

$$x - b = \lambda, \quad -\lambda \in \partial |y|^p, \quad y = x. \tag{4.25}$$

It is straightforward to notice that (4.25) differs from (4.22), since in (4.25) there is no complementary condition.

For our problem (4.18) from (4.24) and further easy computations, we see that the triple $x = 0$, $y = 0$, $\lambda = -b$ is a solution of (4.25), whereas, as already remarked, it does not satisfy (4.22). This is strictly connected to the fact that (4.22) is a primal-dual optimality condition of complementary type, which is equivalent to say that (4.22) is a necessary condition for global minimizers. On the contrary (4.25) is a necessary condition for local ones.

Finally we mention that in [36] an alternate direction method of multipliers is proposed to find stationary points of the type (4.25) (see Eq. 4 of [36]). The method is derived from an augmented Lagrangian formulation, where however, differently from the current paper, the penalization term in [36] is chosen "large enough".

4.3 Uniqueness and convergence

The goals of this subsection are to provide sufficient conditions for uniqueness of solutions to (4.14) and convergence of Algorithm 2. This algorithm proved to be very efficient in our numerical tests which we shall report on further below. Its analysis, however, is quite challenging and will require additional assumptions. Before specifying them we introduce some additional notation.

For any pair x, λ let

$$\mathcal{I}(x, \lambda) = \{i \,:\, |B_i(\Lambda x)_i + \lambda_i| > \mu_i\} \text{ and } \mathcal{A}(x, \lambda) = \{i \,:\, |B_i(\Lambda x)_i + \lambda_i| \leq \mu_i\}.$$

We define the square matrix

$$Q = (\Lambda^*)^{-1} A^* A (\Lambda)^{-1} + (\Lambda^*)^{-1} \eta P \Lambda^{-1}, \tag{4.26}$$

and the constants

$$\alpha := \frac{1-p}{2-p} \geq 0, \quad \gamma = \frac{1}{2-p} > 0, \tag{4.27}$$

where $p \in (0, 1]$. We note that $1 = \gamma + \alpha$.

We shall employ a diagonal dominance condition expressed as

$$\|B^{-\alpha}(Q - B)B^{-\gamma}\|_\infty \leq \rho \text{ for some } \rho \in (0, 1). \tag{4.28}$$

Additionally we make use of the following strict complementarity conditions for a solution x, λ to (4.14)

$$\min_{\mathcal{I}(x,\lambda)} |B^{-\alpha}(\lambda + B\Lambda x)| \geq (1+\delta)\beta^\gamma C_1, \tag{4.29}$$

$$\max_{\mathcal{A}(x,\lambda)} |B^{-\alpha}(\lambda + B\Lambda x)| \leq (1-\delta)\beta^\gamma C_1, \tag{4.30}$$

for some $\delta \in [0,1]$ large enough. Here $\min_{\mathcal{I}(x,\lambda)} |B^{-\alpha}(\lambda + B\Lambda x)|$ stands for $\min_{i\in\mathcal{I}(x,\lambda)} |B_i^{-\alpha}(\lambda_i + B_i(\Lambda x)_i)|$.

We shall choose ε such that

$$\varepsilon \geq \max_i \left(\frac{\beta(1-p)pB_i^{-1}}{1-\tilde{\rho}} \right)^\gamma, \quad \text{where } \tilde{\rho} \in (\rho, 1). \tag{4.31}$$

Finally we set

$$\begin{aligned} C_1 = (2-p)(2(1-p))^{-\alpha}, \quad C_2 = ||B^{-\alpha}||_\infty ||B^\alpha||_\infty p^\gamma \left(\tfrac{1-p}{1-\tilde{\rho}}\right)^{-\alpha}, \\ C_3 = \max(C_2, C_1), \quad \text{which results in } \mu_i = \beta^\gamma C_1 B_i^\alpha, \end{aligned} \tag{4.32}$$

where we recall the definition of μ_i in (4.11).

Let us make some remarks on these specifications. In the case that Q is a diagonal matrix $Q - B = 0$ and (4.28) is trivially satisfied. We observe that for $p \to 0$, we have $\alpha = \gamma = \frac{1}{2}$. In particular (4.28) coincides when $\Lambda = I$ with the diagonal dominance condition considered in [27] to prove the convergence of the primal-dual active set strategy in the case $p = 0$ and $\Lambda = I$. Note we note that the admissible range of ε in (4.31) decreases with β and $p < 1$.

We first prove the following lemma, which is a key ingredient for the proof of Theorem 4 and will be used also in the proof of the convergence result Theorem 5.

Lemma 1 *Assume that (4.2), (4.28) and (4.31) hold.*

(i) *Let $y \in \mathbb{R}$, $y \neq 0$, and $\lambda = \frac{\beta p y}{\max\{\varepsilon^{2-p}, |y|^{2-p}\}}$. Then for any $i = 1, \ldots, d$ it holds*

$$|B_i^{-\alpha}\lambda| \leq \beta^\gamma C_2. \tag{4.33}$$

(ii) *Let $y^1, y^2 \in \mathbb{R}^d$ be such that for all $i = 1, \ldots, d$ we have $y_i^1 \neq 0$, $y_i^2 \neq 0$, and*

$$\begin{aligned} &B_i^\gamma (y_i^1 - y_i^2) + B_i^{-\alpha} \left(\frac{\beta p y_i^1}{\max(\varepsilon^{2-p}, |y_i^1|^{2-p})} - \frac{\beta p y_i^2}{\max(\varepsilon^{2-p}, |y_i^2|^{2-p})} \right) \\ &= \left(B^{-\alpha}(B-Q)B^{-\gamma}B^\gamma(y^1 - y^2) \right)_i. \end{aligned} \tag{4.34}$$

Then $y^1 = y^2$.

Proof We first prove (i). From the definition of λ we have

$$|B_i^{-\alpha}\lambda| \leq ||B^{-\alpha}||_\infty \frac{\beta p|y|}{\max(\varepsilon^{2-p}, |y|^{2-p})}| \leq ||B^{-\alpha}||_\infty \beta p \varepsilon^{p-1}$$

and by (4.31) we conclude (4.33).

Now we prove (ii). For $i = 1, \dots, d$, let us define $h_i : \mathbb{R} \to \mathbb{R}$ and $h_i^\varepsilon : \mathbb{R} \to \mathbb{R}$ by

$$h_i(z) = z + B_i^{-1} \frac{\beta p z}{|z|^{2-p}}, \quad h_i^\varepsilon(z) = z + B_i^{-1} \frac{\beta p z}{\max(\varepsilon^{2-p}, |z|^{2-p})}. \tag{4.35}$$

A short computation shows that h_i has a local minimum at $y_i^+ = \left(\beta(1-p)pB_i^{-1}\right)^{\frac{1}{2-p}}$, respectively a local maximum at $y_i^- = -\left(\beta(1-p)pB_i^{-1}\right)^{\frac{1}{2-p}}$. By the choice of ε in (4.31), considering the cases $\varepsilon \geq |z|$ and $\varepsilon < |z|$, it can be argued that

$$(h_i^\varepsilon)' \geq \tilde{\rho}, \quad \text{on } \mathbb{R}. \tag{4.36}$$

Together with $h_i^\varepsilon(0) = 0$, it follows that

$$(h_i^\varepsilon(y) - h_i^\varepsilon(x))(y - x) \geq \tilde{\rho}|y - x|^2, \quad \text{for all } x, y \in \mathbb{R}. \tag{4.37}$$

By (4.34) and recalling that $\alpha + \gamma = 1$, we obtain

$$B_i^\gamma \left(h_i^\varepsilon(y_i^1) - h_i^\varepsilon(y_i^2)\right) = \left(B^{-\alpha}(B - Q)B^{-\gamma}B^\gamma(y^1 - y^2)\right)_i.$$

Multiplying by $(y_i^1 - y_i^2)$, using (4.37) and (4.28) we obtain

$$\tilde{\rho} B_i^\gamma |y_i^1 - y_i^2|^2 \leq \rho |y_i^1 - y_i^2| |B^\gamma(y^1 - y^2)|_\infty.$$

Thus $\tilde{\rho}|B_i^\gamma(y_i^1 - y_i^2)|_\infty \leq \rho |B^\gamma(y^1 - y^2)|_\infty$ and since $\tilde{\rho} \in (\rho, 1)$ we get that $y^1 = y^2$. □

Theorem 4 *(Uniqueness) Assume that (4.2), (4.28), and (4.31) hold. Then there exists at most one solution to (4.14) satisfying (4.29) with $\delta > \frac{2\rho C_3}{(1-\rho)C_1}$. An analogous statement holds with (4.29) replaced by (4.30).*

Proof Assume that there exist two pairs x, λ and $\hat{x}, \hat{\lambda}$ satisfying (4.14) and (4.29). Let $y = \Lambda x$ and $\hat{y} = \Lambda \hat{x}$. By multiplying (4.14) by $(\Lambda^*)^{-1}$ and using the definition of Q, we have

$$Q(y - \hat{y}) + \lambda - \hat{\lambda} = 0. \tag{4.38}$$

Multiplying (4.38) by $B^{-\alpha}$ and using that $\alpha + \gamma = 1$, we have

$$B^\gamma y + B^{-\alpha}\lambda - (B^\gamma \hat{y} + B^{-\alpha}\hat{\lambda}) = B^{-\alpha}(B - Q)B^{-\gamma}B^\gamma(y - \hat{y}). \tag{4.39}$$

Case 1 First consider the case $y_i \neq 0$ if and only if $\hat{y}_i \neq 0$ for each i. From (4.14) we have

$$\lambda_i = \frac{\beta p y_i}{\max(\varepsilon^{2-p}, |y_i|^{2-p})}, \quad \hat{\lambda}_i = \frac{\beta p \hat{y}_i}{\max(\varepsilon^{2-p}, |\hat{y}_i|^{2-p})}. \tag{4.40}$$

By (4.39) we have

$$\begin{aligned} &B_i^\gamma(y_i - \hat{y}_i) + B_i^{-\alpha}\left(\frac{\beta p y_i}{\max(\varepsilon^{2-p}, |y_i|^{2-p})} - \frac{\beta p \hat{y}_i}{\max(\varepsilon^{2-p}, |\hat{y}_i|^{2-p})}\right) \\ &= \left(B^{-\alpha}(B - Q)B^{-\gamma}B^\gamma(y - \hat{y})\right)_i, \end{aligned}$$

that is, y and $\hat{y}$ satisfy (4.34) with $y = y^1$ and $\hat{y} = y^2$. Then by Lemma 1 (i), we conclude that $y = \hat{y}$.

Case 2 Here we suppose that there exists j such that $y_j \neq 0$ and $\hat{y}_j = 0$. The case $\hat{y}_j \neq 0$ and $y_j = 0$ can be treated analogously. Before using these settings some considerations are necessary. From (4.14) and since $\mu_i = \beta^\gamma C_1 B_i^\alpha$ we deduce that

$$|B_i^{-\alpha}\lambda_i| \leq \beta^\gamma C_1 \quad \text{if } y_i = 0, \quad |B_i^{-\alpha}\hat{\lambda}_i| \leq \beta^\gamma C_1 \quad \text{if } \hat{y}_i = 0, \tag{4.41}$$

and by (4.33)

$$|B_i^{-\alpha}\hat{\lambda}_i| \leq \beta^\gamma C_2 \quad \text{if } \hat{y}_i \neq 0, \quad |B_i^{-\alpha}\hat{\lambda}_i| \leq \beta^\gamma C_2 \quad \text{if } \hat{y}_i \neq 0. \tag{4.42}$$

By (4.39) and (4.28) we have

$$|B_i^\gamma(y_i - \hat{y}_i)| \leq |B_i^{-\alpha}(\hat{\lambda}_i - \lambda_i)| + \rho|B^\gamma(y - \hat{y})|_\infty$$

and combined with (4.41), (4.42) and the definition of C_3

$$|B^\gamma(y - \hat{y})|_\infty \leq 2C_3\beta^\gamma + \rho|B^\gamma(y - \hat{y})|_\infty$$

and thus

$$|B^\gamma(y - \hat{y})|_\infty \leq \frac{2C_3\beta^\gamma}{1 - \rho}. \tag{4.43}$$

For j chosen as above we estimate using (4.28), (4.39), and (4.43)

$$|B_j^{-\alpha}(\lambda_j + B_j y_j)| - |B_j^{-\alpha}(\hat{\lambda}_j + B_j \hat{y}_j)| \le |B^{-\alpha}(\lambda - \hat{\lambda}) + B^{\gamma}(y - \hat{y})|_\infty$$
$$= |B^{-\alpha}(B - Q)B^{-\gamma} B^{\gamma}(y - \hat{y})|_\infty$$
$$\le \rho |B^{\gamma}(y - \hat{y})|_\infty \le \frac{2\rho C_3 \beta^{\gamma}}{1 - \rho}. \quad (4.44)$$

Now we utilize (4.29) and (4.42) in the last estimate to obtain

$$(1 + \delta)C_1 \beta^{\gamma} - \beta^{\gamma} C_1 \le \frac{2\rho C_3 \beta^{\gamma}}{1 - \rho}$$

or equivalently

$$\delta \le \frac{2\rho C_3}{(1 - \rho)C_1}, \quad (4.45)$$

which contradicts the choice of δ.

Case 3 The only remaining case consists in $y_i = \hat{y}_i = 0$ for all i. But then $\lambda = \hat{\lambda}$ by (4.14) and $x = \hat{x}$ by the invertibility of Λ as desired. This concludes the proof under assumption (4.29).

The case (4.30) can be treated analogously. Note that condition (4.29) (or respectively (4.30)) arises only in *Case 2* and therefore the proof of *Case 1* and *Case 3* can be carried out exactly as above. In *Case 2* we proceed as above to get (4.44). Since $y_j \neq 0$ by the third equation in (4.3) and (4.8), we have

$$|\lambda_j + B_j y_j| \ge \mu_j$$

and recalling that $\mu_j = \beta^{\gamma} C_1 B_j^{\alpha}$, we get

$$|B_j^{-\alpha}(\lambda_j + B_j y_j)| \ge \beta^{\gamma} C_1. \quad (4.46)$$

By (4.30) we get

$$|B_j^{-\alpha}(\hat{\lambda}_j + B_j \hat{y}_j)| \le (1 - \delta)\beta^{\gamma} C_1. \quad (4.47)$$

Putting (4.46) and (4.47) in (4.44), we obtain

$$\beta^{\gamma} C_1 - (1 - \delta)\beta^{\gamma} C_1 \le \frac{2\rho C_3 \beta^{\gamma}}{(1 - \rho)},$$

and thus $\delta \le \frac{2\rho C_3}{(1-\rho)C_1}$, which gives the desired contradiction to the choice of δ. □

4.3.1 Convergence

Here we give a sufficient condition for the convergence of the primal-dual active set method. As in [27] we utilize the diagonal dominance condition (4.28) and consider

a solution x, λ to (4.14) which satisfies the strict complementary conditions. As such it is unique according to Theorem 4.

At first let us emphasize that the sequence $\{x^n\}_{n\in\mathbb{N}}$ is bounded uniformly in n. Indeed since $\langle x^{n+1}, \lambda^{n+1}\rangle \geq 0$ for all n, we have from the first equation in (4.14)

$$\langle (A^*A + \eta P)x^{n+1}, x^{n+1}\rangle \leq \langle Ax^{n+1}, b\rangle,$$

which coupled with (4.7) gives

$$\xi |x^{n+1}|_2 \leq \|A\|_2 |b|_2.$$

Setting $M := \max(\|A\|_2|b|_2\xi^{-1}, |x_0|_2)$, where ξ is defined in (4.7) we conclude that

$$|x^n|_2 \leq M \quad \text{for all } i = 1, 2, \ldots. \tag{4.48}$$

Note that the same bound as (4.48) holds also for any solution x of the necessary optimality condition (4.14).

We will use the following notation

$$F := ||B^\gamma||_\infty ||\Lambda||_\infty M, \tag{4.49}$$

where γ and M are defined respectively in (4.27) and (4.48).

Theorem 5 *Suppose that (4.2), (4.28), and (4.31) hold. Let $\bar{x}, \bar{\lambda}$ be a solution to (4.14) satisfying the strict complementary condition (4.29)–(4.30), with $\delta > \frac{\rho}{(1-\rho)C_1}\left(2\rho\beta^{-\gamma}F + 2C_3\right)$, and define the sets*

$$\mathcal{S}^n = \left\{ i \in \mathcal{I}(\bar{x}, \bar{\lambda}) \,:\, \lambda_i^n = \frac{\beta\, p(\Lambda x^n)_i}{\max(\varepsilon^{2-p}, |(\Lambda x^n)_i|^{2-p})} \right\},$$
$$\mathcal{T}^n = \{ i \in \mathcal{A}(\bar{x}, \bar{\lambda}) \,:\, (\Lambda x^n)_i = 0 \}.$$

Then $\mathcal{S}^n = \mathcal{I}(\bar{x}, \bar{\lambda})$ for $n \geq 2$ and $\mathcal{T}^n \subset \mathcal{T}^{n+1}$ for $n \geq 1$. As soon as $\mathcal{S}^n = \mathcal{S}^{n+1}$ and $\mathcal{T}^n = \mathcal{T}^{n+1}$, we have $x^{n+1}, \lambda^{n+1} = \bar{x}, \bar{\lambda}$.

Note that $\mathcal{S}^n$ is the set of all indices which are inactive both for the $\bar{x}, \bar{\lambda}$ and for x^n, λ^n, and analogously $\mathcal{S}^n$ is the set of active indices for both pairs. Note also that due to the finite dimensionality the case $\mathcal{T}^n = \mathcal{T}^{n+1}$ for some n must occur.

Proof We divide the proof into three steps. In *Step* (i) we verify a bound on $x^n - \bar{x}$ which will be used throughout the rest of the proof, in *Step* (ii) we prove the claimed properties of $\mathcal{S}^n$ and $\mathcal{T}^n$, and in *Step* (iii) we conclude the proof of convergence.
Step (*i*) Let $\bar{y} = \Lambda\bar{x}$ and $y^n = \Lambda x^n$, for $n \geq 1$. Multiplying (4.14) by $(\Lambda^*)^{-1}$ and using the definition of Q in (4.26), we have

$$Q(y^n - \bar{y}) + \lambda^n - \bar{\lambda} = 0. \tag{4.50}$$

Multiplying (4.50) by $B^{-\alpha}$ and using that $\alpha + \gamma = 1$ we get

$$B^{\gamma}(y^n - \bar{y}) + B^{-\alpha}(\lambda^n - \bar{\lambda}) = B^{-\alpha}(B - Q)B^{-\gamma}B^{\gamma}(y^n - \bar{y}). \tag{4.51}$$

Suppose first that $y_i^n \neq 0$ and $\bar{y}_i \neq 0$ for all i. By (4.51), (4.14), (4.16), and (4.17) we have that y^n and $\bar{y}$ satisfy the assumptions of Lemma 1 (ii) with $y^1 = y^n$ and $y^2 = \bar{y}$. Then by Lemma 1 (ii) we obtain that $y^n = \bar{y}$.

Next consider the case $y_i^n = 0$, $\bar{y}_i \neq 0$. For two consecutive iterates we have

$$Q(y^n - y^{n-1}) + \lambda^n - \lambda^{n-1} = 0 \tag{4.52}$$

and thus, multiplying the equation by $B^{-\alpha}$ and using that $\alpha + \gamma = 1$ we get

$$B^{-\alpha}(\lambda^n + By^n) - B^{-\alpha}(\lambda^{n-1} + By^{n-1}) = B^{-\alpha}(B - Q)B^{-\gamma}B^{\gamma}(y^n - y^{n-1}). \tag{4.53}$$

Since $y_i^n = 0$, by (4.16) and (4.17) we have $|B_i y_i^{n-1} + \lambda_i^{n-1}| \leq \mu_i = \beta^{\gamma} C_1 B_i^{\alpha}$, and by (4.53) and (4.28) we get

$$\begin{aligned} |B_i^{-\alpha}\lambda_i^n| &\leq |[B^{-\alpha}(B - Q)B^{-\gamma}B^{\gamma}(y^n - y^{n-1})]_i| + |B_i^{-\alpha}(\lambda_i^{n-1} + B_i y_i^{n-1})| \\ &\leq \rho |B^{\gamma}(y^n - y^{n-1})|_{\infty} + \beta^{\gamma} C_1 \leq 2\rho F + \beta^{\gamma} C_1, \end{aligned}$$

where F is defined in (4.49). Since $\bar{y}_i \neq 0$, by (4.14) and (4.33), we have

$$|B_i^{-\alpha}\bar{\lambda}_i| \leq \beta^{\gamma} C_2. \tag{4.54}$$

By (4.51), (4.54), (4.54) and (4.28) we get

$$|B_i^{\gamma}(y_i^n - \bar{y}_i)| \leq 2\rho F + \beta^{\gamma} C_1 + \beta^{\gamma} C_2 + \rho |B^{\gamma}(y^n - \bar{y})|_{\infty}. \tag{4.55}$$

Similarly if $y_i^n \neq 0$ and $\bar{y}_i = 0$, we get

$$|B_i^{\gamma}(y_i^n - \bar{y}_i)| \leq \beta^{\gamma} C_1 + \beta^{\gamma} C_2 + \rho |B^{\gamma}(y^n - \bar{y})|_{\infty}. \tag{4.56}$$

Combining (4.55) and (4.56), we have

$$|B_i^{\gamma}(y_i^n - \bar{y}_i)| \leq 2\rho F + 2\beta^{\gamma} C_3 + \rho |B^{\gamma}(y^n - \bar{y})|_{\infty}.$$

Summarizing these estimates we have

$$|B^{\gamma}(y^n - \bar{y})|_{\infty} \leq \frac{2\rho F + 2\beta^{\gamma} C_3}{1 - \rho} \quad \text{for } n \geq 1. \tag{4.57}$$

Step (ii) We first consider the relationship between $\mathcal{S}^n$ and $\mathcal{I}(\bar{x}, \bar{\lambda})$. Note that if $\mathcal{I}(\bar{x}, \bar{\lambda}) = \emptyset$, then $\mathcal{S}^n = \emptyset$ for all $n \geq 0$. Henceforth we consider the case $\mathcal{I}(\bar{x}, \bar{\lambda}) \neq \emptyset$. We have

$$|B_i^{-\alpha}\lambda_i^n + B_i^{\gamma} y_i^n| \geq |B_i^{-\alpha}\bar{\lambda}_i + B_i^{\gamma}\bar{y}_i| - |B_i^{-\alpha}\lambda_i^n + B_i^{\gamma} y_i^n - B_i^{-\alpha}\bar{\lambda}_i - B_i^{\gamma}\bar{y}_i|. \tag{4.58}$$

By (4.51), (4.28), and (4.57) we deduce for $n \geq 1$

$$\begin{aligned} |B_i^{-\alpha}\lambda_i^n + B_i^{\gamma} y_i^n - B_i^{-\alpha}\bar{\lambda}_i - B_i^{\gamma}\bar{y}_i| &= |(B^{-\alpha}(B-Q)B^{-\gamma}B^{\gamma}(y^n-\bar{y}))_i| \\ &\leq \rho |B^{\gamma}(y^n-\bar{y})|_\infty \leq \frac{\rho}{1-\rho}(2\rho F + 2\beta^{\gamma} C_3). \end{aligned} \tag{4.59}$$

Then by (4.58), (4.59), and (4.29) we obtain for $i \in \mathcal{I}(\bar{x}, \bar{\lambda})$

$$|B_i^{-\alpha}\lambda_i^n + B_i^{\gamma} y_i^n| \geq (1+\delta)\beta^{\gamma} C_1 - \frac{\rho}{1-\rho}(2\rho F + 2\beta^{\gamma} C_3).$$

Since by assumption

$$\delta > \frac{\rho}{(1-\rho)C_1}(2\rho\beta^{-\gamma} F + 2C_3), \tag{4.60}$$

we have

$$|B_i^{-\alpha}\lambda_i^n + B_i^{\gamma} y_i^n| > \beta^{\gamma} C_1 = \mu_i B_i^{-\alpha}.$$

From (4.17) we deduce that $\lambda_i^{n+1} = \frac{\beta p(\Lambda x^{n+1})_i}{\max(\varepsilon^{2-p}, |(\Lambda x^{n+1})_i|^{2-p})}$, and thus $i \in \mathcal{S}^{n+1}$. Since $\mathcal{S}^{n+1} \subset \mathcal{I}(\bar{x}, \bar{\lambda})$, it follows that $\mathcal{S}^{n+1} = \mathcal{I}(\bar{x}, \bar{\lambda})$ for $n \geq 1$.

For $i \in \mathcal{T}^n$ by (4.51), (4.28) and (4.57) we have

$$|B_i^{-\alpha}(\lambda_i^n - \bar{\lambda}_i)| \leq \rho |B^{\gamma}(y^n - \bar{y})|_\infty \leq \frac{\rho}{1-\rho}(2\rho F + 2\beta^{\gamma} C_3). \tag{4.61}$$

By the definition of $\mathcal{T}^n$, (4.61) and the strict complementary condition (4.30), we get

$$\begin{aligned} |B_i^{-\alpha}(\lambda_i^n + B_i y_i^n)| = |B^{-\alpha}\lambda_i^n| &\leq |B_i^{-\alpha}(\lambda_i^n - \bar{\lambda}_i)| + |B_i^{-\alpha}\bar{\lambda}_i| \\ &\leq \frac{\rho}{1-\rho}(2\rho F + 2\beta^{\gamma} C_3) + (1-\delta)\beta^{\gamma} C_1. \end{aligned} \tag{4.62}$$

By taking δ as in (4.60) we have

$$|B_i^{-\alpha}(\lambda_i^n + B_i y_i^n)| < \beta^{\gamma} C_1$$

and hence $y_i^{n+1} = 0$ and $i \in \mathcal{T}^{n+1}$. Thus $\mathcal{T}^n \subseteq \mathcal{T}^{n+1}$.

Step (iii) Assume now that $\mathcal{S}^n = \mathcal{S}^{n+1}$ and $\mathcal{T}^n = \mathcal{T}^{n+1} \subset \mathcal{A}(\bar{x}, \bar{\lambda})$. Then $\mathcal{S}^n = \mathcal{I}(\bar{x}, \bar{\lambda})$. This is proved in *Step 2* for the case $n \geq 2$ and in case $n = 1$ we have $\mathcal{S}^1 = \mathcal{S}^2 = \mathcal{I}(\bar{x}, \bar{\lambda})$.

Assume now that $\mathcal{T}^n = \mathcal{T}^{n+1} \subsetneq \mathcal{A}(\bar{x}, \bar{\lambda})$ and $i \in \mathcal{A}(\bar{x}, \bar{\lambda}) \backslash \mathcal{T}^n$. Then

$$y_i^{n+1} \neq 0, \quad y_i^n \neq 0, \quad \bar{y}_i = 0, \quad \lambda_i^{n+1} = \frac{\beta p y_i^{n+1}}{\max(\varepsilon^{2-p}, |y_i^{n+1}|^{2-p})},$$
$$\lambda_i^n = \frac{\beta p y_i^n}{\max(\varepsilon^{2-p}, |y_i^n|^{2-p})}. \tag{4.63}$$

By (4.17), the first and third equations in (4.63), (4.51), and (4.30) we have

$$\begin{aligned} \beta^\gamma C_1 = B_i^{-\alpha} \mu_i &\leq |B_i^\gamma y_i^n + B_i^{-\alpha} \lambda_i^n| \\ &\leq |B^\gamma y_i^n + B_i^{-\alpha} \lambda_i^n - B_i^\gamma \bar{y}_i - B_i^{-\alpha} \bar{\lambda}_i| + |B_i^{-\alpha} \bar{\lambda}_i| \\ &\leq |B^{-\alpha}(B - Q) B^{-\gamma} B^\gamma (y^n - \bar{y})|_\infty + (1 - \delta)\beta^\gamma C_1. \end{aligned} \tag{4.64}$$

By (4.28) and (4.57) we obtain

$$\beta^\gamma C_1 \leq \rho |B^\gamma (y^n - \bar{y})|_\infty + (1 - \delta)\beta^\gamma C_1 \leq \tfrac{\rho}{1-\rho}(2\rho F + 2\beta^\gamma C_3) + (1 - \delta)\beta^\gamma C_1, \tag{4.65}$$

from which it follows that

$$\delta \leq \frac{\rho}{(1 - \rho) C_1} \left(2\rho \beta^{-\gamma} F + 2C_3\right),$$

leading to a contradiction to the choice of δ in (4.60).

It follows that $\mathcal{S}^n = \mathcal{I}(\bar{x}, \bar{\lambda})$ and $\mathcal{T}^n = \mathcal{A}(\bar{x}, \bar{\lambda})$. Comparing (4.14), which is satisfied by $\bar{x}, \bar{\lambda}$, and (4.15)–(4.17), which holds with $y_i^n = y_i^{n+1}, \lambda_i^n = \lambda_i^{n+1}$, we conclude that $x^{n+1} = \bar{x}, \lambda^{n+1} = \bar{\lambda}$. □

Remark 6 If $\bar{x}_i \neq 0$ for all i we have $\mathcal{A}(\bar{x}, \bar{\lambda}) = \emptyset$. Then by the definition of $\mathcal{T}^n$ and Step (ii) of the proof of Theorem 5 we obtain $\mathcal{T}^1 = \mathcal{T}^2 = \emptyset$ and therefore Algorithm 2 is a 2-step algorithm, in this case.

5 Active set monotone algorithm: numerical results

Here we describe the active set monotone scheme (see Algorithm 3) and discuss the numerical results for two different test cases. The first one is the time-dependent control problem from Sect. 3.2, the second one is an example in microscopy image reconstruction. Typically the active set monotone scheme requires fewer iterations and achieves a lower residue than the monotone scheme of Sect. 2.

5.1 The numerical scheme

The proposed active set monotone algorithm consists of an *outer loop* based on the primal-dual active set strategy and an *inner loop* which uses the monotone algorithm to solve the nonlinear part of the optimality condition.

In order to achieve a better numerical performance, we write the optimality condition as explained in the following. At each iteration of the active set strategy (Algorithm 2) we solve the following system in x^{n+1}, λ^{n+1}

$$\begin{cases} A^*(Ax^{n+1} - b) + \eta P x^{n+1} + \Lambda^* \lambda^{n+1} = 0, \\ (\Lambda x^{n+1})_i = 0 & \text{if } i \in \mathcal{A}_n, \\ \lambda_i^{n+1} = \frac{\beta p (\Lambda x^{n+1})_i}{\max(\varepsilon^{2-p}, |(\Lambda x^{n+1})_i|^{2-p})} & \text{if } i \in \mathcal{I}_n, \end{cases} \tag{5.1}$$

where $\mathcal{A}_n = \{i : |B_i y_i^n + \lambda_i^n| \leq \mu_i\}$ are the active indexes and $\mathcal{I}_n = \mathcal{A}_n^c$ are the inactive ones. We write (5.1) in the following form

$$\begin{cases} (A^*A + \Lambda_{\mathcal{I}_n}^* N_{\mathcal{I}_n}^{n+1} \Lambda_{\mathcal{I}_n} + \eta P) x^{n+1} + \Lambda_{\mathcal{A}_n}^* \lambda_{\mathcal{A}_n}^{n+1} = A^* b, \\ \Lambda_{\mathcal{A}_n} x^{n+1} = 0, \end{cases} \tag{5.2}$$

where $N_{\mathcal{I}_n}^{n+1}$ is the diagonal operator with i-entries $(N_{\mathcal{I}_n}^{n+1})_{ii, i \in \mathcal{I}_n} = \frac{\beta p}{\max(\varepsilon^{2-p}, |(\Lambda x^{n+1})_{i \in \mathcal{I}_n}|^{2-p})}$ and $\Lambda_{\mathcal{A}_n}$, $\Lambda_{\mathcal{I}_n}$ are the rows of Λ corresponding to the active and inactive indexes respectively.

In order to solve (5.2) we apply the following iterative procedure solved for $x^{k+1,n+1}, \lambda^{k+1,n+1}$

$$\begin{cases} (A^*A + \Lambda_{\mathcal{I}_n}^* N_{\mathcal{I}_n}^{k,n+1} \Lambda_{\mathcal{I}_n} + \eta P) x^{k+1,n+1} + \Lambda_{\mathcal{A}_n}^* \lambda_{\mathcal{A}_n}^{k+1,n+1} = A^* b, \\ \Lambda_{\mathcal{A}_n} x^{k+1,n+1} = 0, \end{cases} \tag{5.3}$$

where $N_{\mathcal{I}_n}^{k,n+1}$ is the diagonal operator with i-entries $(N_{\mathcal{I}_n}^{k,n+1})_{ii, i \in \mathcal{I}_n} = \frac{\beta p}{\max(\varepsilon^{2-p}, |(\Lambda x^{k,n+1})_{i \in \mathcal{I}_n}|^{2-p})}$. Note that the iterative procedure of (5.3) is convergent by Theorem (2).

Remark 7 Note that the system matrix associated to (5.3) is symmetric.

The algorithm stops when the residue of (5.2) and (4.3) (for the inner and the outer cycle respectively) is $O(10^{-12})$ in the control problem and $O(10^{-8})$ in the microscopy image example.

We remark that in our numerical tests we always took $\eta = 0$. The initialization x^0, λ^0 in the outer cycle is chosen in the following way

$$x^0 = (A^*A + 2\beta \Lambda^* \Lambda)^{-1} A^* b, \;\; \lambda^0 = \Lambda^{-1} A^* (b - A x_0). \tag{5.4}$$

In particular λ^0 is the solution of the first equation in (4.3) for $x = x^0$. As in Sect. 3, for some values of β the previous initialization is not suitable. Following the idea already used for the monotone scheme, we successfully tested an analogous continuation strategy with respect to increasing β-values.

In Algorithm 3 we jump out of at the inner loop in case of presence of singular components. We recall that the singular components are those i such that $|(\Lambda x)_i| < \varepsilon\}$, that is, the components where the ε-regularization is most influential.

Algorithm 3 Active set monotone scheme

1: Initialize $\varepsilon > 0$, x^0, λ^0, $y^0 = \Lambda x^0$. Set $n = 0$.
2: **repeat** {outer loop}
3: Let $\mathcal{A}_n = \{i : |B_i y_i^n + \lambda_i^n| \leq \mu_i\}$, $\mathcal{I}_n = \mathcal{A}_n^c$. Initialize $x^{0,n+1} = x^n$, $\lambda^{0,n+1} = \lambda^n$ and $y^{0,n+1} = \Lambda x^{0,n}$. Set $k = 0$.
4: **repeat** {inner loop}
5: Solve for $x^{k+1,n+1}$, $\lambda_{\mathcal{A}_n}^{k+1,n+1}$

$$\begin{cases} (A^*A + \Lambda_{\mathcal{I}_n}^* N_{\mathcal{I}_n}^{k,n+1} \Lambda_{\mathcal{I}_n} + \eta P) x^{k+1,n+1} + \Lambda_{\mathcal{A}_n}^* \lambda_{\mathcal{A}_n}^{k+1,n+1} = A^*b \\ \Lambda_{\mathcal{A}_n} x^{k+1,n+1} = 0 \end{cases} \tag{5.5}$$

Set $y^{k+1,n+1} = \Lambda x^{k+1,n+1}$, $\lambda_{\mathcal{I}_n}^{k+1,n+1} = \dfrac{\beta p y_{\mathcal{I}_n}^{k+1,n+1}}{\max(\varepsilon^{2-p}, |y_{\mathcal{I}_n}^{k+1,n+1}|^{2-p})}$.

6: If $y_{\mathcal{I}_n}^{k+1,n+1}$ is a singular point, go to 9.
7: Set $k = k + 1$.
8: **until** the stopping criteria for the inner loop is fulfilled.
9: Set $n = n + 1$;
10: **until** the stopping criteria for the outer loop is fulfilled.
11: Reduce ε and go to 3.

In the case Λ coincides with the identity the system (5.1) can be written as

$$\begin{cases} x_i^{n+1} = 0 & \text{if } i \in \mathcal{A}_n, \\ \langle A_i, Ax^{n+1} - b \rangle + \eta P_{ij} x_j^{n+1} + \dfrac{\beta p x_i^{n+1}}{\max(\varepsilon^{2-p}, |x_i^{n+1}|^{2-p})} = 0 & \text{if } i \in \mathcal{I}_n. \end{cases} \tag{5.6}$$

Note that in (5.6) we coupled the first and the third equation in (5.1) and we eliminated the dual variable. The advantage is that now we solve the second equation in (5.6) only for the inactive components $x_{\mathcal{I}_n}$, solving a system of $|\mathcal{I}_n|$ equations, whereas in (5.5) we solve $d + |\mathcal{A}_n|$ equations. Finally we remark that in the case Λ coincides with the identity $\varepsilon > 0$ is fixed. In particular $\varepsilon = \min_i \left(\frac{2\beta(1-p)}{|A_i|_2^2} \right)^{\frac{1}{2-p}}$ accordingly to the lower bound on the inactive components given by Corollary 1.

5.2 Sparsity in a time-dependent control problem

We test the active set monotone algorithm on the time-dependent control problem described in Sect. 3.2, with the same discretization in space and time ($\Delta x = \Delta t = \frac{1}{50}$) and target function b. Also the initialization of x and the ε-range are the same. In Tables 4 we report the results of our tests for $p = .1$ and β incrementally increasing by factor of 10 from 10^{-3} to 1. We report only the values for the second control u_2 since the first control u_1 is always zero. As expected, $|Du_2|_0^c$ increases and $|Du_2|_p^p$ decreases when β is increasing. Note that the number of iterations of the inner and outer cycle are both small.

The algorithm was also tested for the same p as in Sect. 3.2, that is $p = .5$, for the same range of β as in Table 4. Comparing to the results achieved by Algorithm 1,

Table 4 Sparsity in a time-dependent control problem, $p = .1$, mesh size $h = \frac{1}{50}$. Results obtained by Algorithm 3

β	10^{-3}	10^{-2}	10^{-1}	1
No. of outer iter.	1	1	4	1
No. of inner iter.	20	20	30	20
$\lvert Du_2\rvert_0^c$	95	95	98	100
$\lvert Du_2\rvert_p^p$	18	17	14	0
Residue	10^{-15}	10^{-15}	10^{-14}	10^{-16}

we obtained the same values for the ℓ^0-term for corresponding values of β and a considerably smaller residue within a significantly fewer number of inner iterations.

Finally we note that if $\Lambda = I$ the number of inner iterations is even smaller, that is, 6 on the average.

5.3 Compressed sensing approach for microscopy image reconstruction

In this subsection we present an application of the active set monotone scheme to compressed sensing for microscopy image reconstruction. We focus on the STORM (stochastic optical reconstruction microscopy) method, which is based on stochastically switching and high-precision detection of single molecules to achieve an image resolution beyond the diffraction limit. The literature on the STORM has been intensively increasing, see e.g. [4,23,25,46]. The STORM reconstruction process consists in a series of imaging cycles. In each cycle only a fraction of the fluorophores in the field of view are switched on (stochastically), such that each of the active fluorophores is optically resolvable from the rest, allowing the position of these fluorophores to be determined with high accuracy. Despite the advantage of obtaining sub-diffraction-limit spatial resolution, in these single molecule detection-based techniques such as STORM, the time to acquire a super-resolution image is limited by the maximum density of fluorescent emitters that can be accurately localized per imaging frame, see e.g. [31,38,47]. In order to get at the same time better resolution and higher emitter density per imaging frame, compressive sensing methods based on l^1 techniques have been recently applied, see e.g. [2,21,50] and the references therein. In the following, we propose a similar approach based on our l^p with $p < 1$ methods. We mention that l^p with $0 < p \leq 1$ techniques based on a concave-convex regularizing procedure, and hence different from ours, are used in [33].

To be more specific, each single frame reconstruction can be achieved by solving the following constrained-minimization problem:

$$\min_{x\in\mathbb{R}^d} |x|_p^p \quad \text{such that} \quad |Ax - b|_2 \leq \varepsilon, \tag{5.7}$$

where $p \in (0, 1]$, x is the up-sampled, reconstructed image, b is the experimentally observed image, and A is the impulse response (of size $m \times d$, where m and d are the numbers of pixels in b and x, respectively). A is usually called the point spread function (PSF) and describes the response of an imaging system to a point source

or point object. The inequality constraint on the ℓ^2-norm allows some inaccuracy in the image reconstruction to accommodate the statistical corruption of the image by noise [50]. Solving problems as (5.7) is referred to as compressed sensing in the literature of microscopy imaging. Indeed, in the basic compressed sensing problem, an under-determined, sparse signal vector is reconstructed from a noisy measurement in a basis in which the signal is not sparse. In the compressed sensing approach to microscopy image reconstruction, the sparse basis is a high resolution grid, in which fluorophore locations are presented, while the noisy measurement basis is the lower resolution camera pixels, on which fluorescence signal are detected experimentally. In this framework, the optimally reconstructed image is the one that contains the fewest number of fluorophores but reproduces the measured image on the camera to a given accuracy (when convolved with the optical impulse response).

We reformulate problem (5.7) as:

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}|Ax-b|_2^2 + \beta|x|_p^p \tag{5.8}$$

and we solve (5.8) by applying Algorithm 3. Note that we may consider (5.8) arising from (5.7) with β related to the reciprocal of the Lagrange multiplier associated to the inequality constraint $|Ax-b|_2 \leq \varepsilon$. We remark also that in the resolution of (5.8) in Algorithm 3 the parameter ε is kept constant and equal to the lower bound on the inactive components given by Corollary 1.

First we tested the procedure for same resolution images, in particular the conventional and the true images are both 128×128 pixel images. Then the algorithm was tested in the case of a 16×16 pixel conventional image and a 128×128 true image. The values for the impulse response A and the measured data b were chosen according to the literature, in particular A was taken as the Gaussian PSF matrix with variance $\sigma = 8$ and size $3 \times \sigma = 24$, and b was simulated by convolving the impulse response A with a random 0-1 mask over the image adding a white random noise so that the signal to noise ratio is .01.

We carried out several tests with the same data for different values of p, β. We report only our results for $p = .1$ and $\beta = 10^{-6}, \beta = 10^{-9}$ for the same and the different resolution case respectively, since for these values the best reconstructions were achieved. The number of single frame reconstructions carried out to get the full reconstruction was 5, 10 for the same, different resolution case, respectively.

In order to measure the performance of our algorithm, we plot a graphic of the average over six recoveries of the location recovery and the exact recovery (up to a certain tolerance) against the noise. Note that in compressed sensing these quantities are typically used as a measure of the efficacy of the reconstruction method, see for example [17] (where, under certain conditions, a linear decay with respect to the noise is proven) and [7].

The first test is carried out for a sparse 0–1 cross-like image. The STORM reconstructions are presented in Figs. 5 and 6 for the same and different resolution case, respectively. In Fig. 7 the plots of the location and exact recovery are shown. Note that our algorithm can recover quite well the location of the emitters. Also, the location and intensity of the emitters decay linearly with respect to the noise level, in

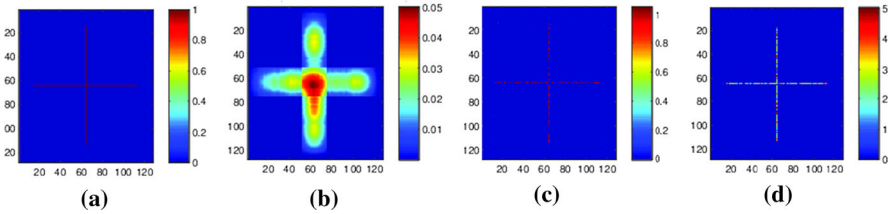

Fig. 5 A STORM reconstruction procedure, same resolution, $p = .1, \beta = 10^{-6}$. Results obtained by Algorithm 3. **a** Real distribution, **b** simulated single frame image, **c** single frame sparse reconstruction, **d** full STORM sparse reconstruction

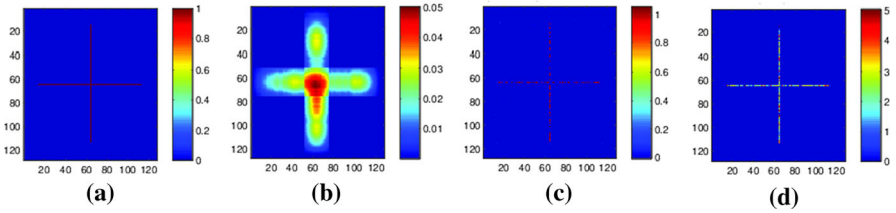

Fig. 6 A STORM reconstruction from a $16x16$ pixel image, different resolution, $p = .1, \beta = 10^{-9}$. Results obtained by Algorithm 3. **a** Real distribution, **b** simulated single frame image, **c** single frame sparse reconstruction, **d** full STORM sparse reconstruction

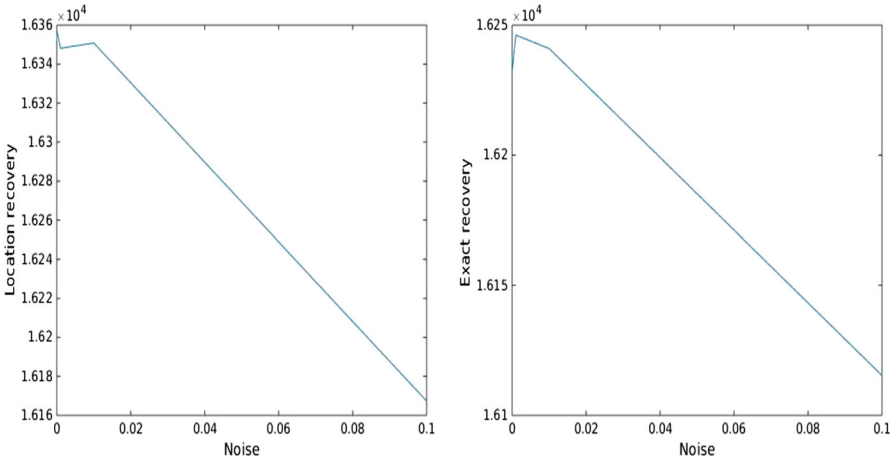

Fig. 7 Left: location recovery. Right: exact recovery. Cross image, different resolution, $p = .1, \beta = 10^{-9}$. Results obtained by Algorithm 3

line with the result of [17]. In particular, for small noise both the recoveries are very near to $d^2 = 16,384$, that is, the exact recovery is 16, 240, 16, 243 and the location is 16, 384, 16, 360 for the same and the different resolution case, respectively. We observe also that the values of the location recovery are higher than the exact recovery for small values of the noise, as expected.

A second test on a non sparse standard phantom image is carried out. In Fig. 8 we show the reconstruction in the case of same resolution images. Note that a high

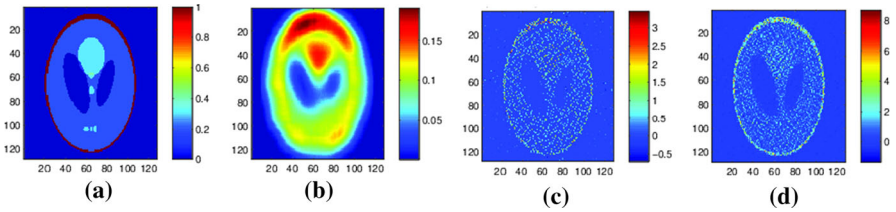

Fig. 8 A STORM reconstruction procedure, same resolution, $p = .1, \beta = 10^{-6}$. Results obtained by Algorithm 3. **a** Real distribution, **b** simulated single frame image, **c** single frame sparse reconstruction, **d** full STORM sparse reconstruction

Table 5 Number of iterations and residue for the cross image (different res.), $p = .1, \beta = 10^{-9}$. Results obtained by Algorithm 3

Frame	1	2	3	4	5	6	7	8	9	10
Iterations outer	100	98	100	100	100	100	100	85	100	100
Iterations inner	147	190	144	184	145	186	146	187	145	165
Residue	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-9}	10^{-8}	10^{-8}	10^{-8}	10^{-8}	10^{-8}

Table 6 Number of iterations for the phantom (same res.), $p = .1, \beta = 10^{-6}$. Results obtained by Algorithm 3

Frame	1	2	3	4	5
Iterations outer	6	11	7	6	6
Iterations inner	9	14	12	7	7
Residue	10^{-8}	10^{-10}	10^{-12}	10^{-8}	10^{-8}

percentage of emitters is correctly localized and the boundaries of the image are well-recovered. Also in this case the location and exact recoveries show a linear decay with respect to the noise.

In Tables 5 and 6 we report the number of iterations needed for each single frame reconstruction. For the cross image in the different resolution case (Table 5), the number of iterations is averagely 100, 164 for the outer cycle and inner cycle, respectively. Note that for the phantom in the same resolution case (Table 6) the number of iterations is lower, that is averagely 7.2, 9.8 for the outer cycle and inner cycle, respectively. The numbers of iterations for the cross image in case of same resolution are comparable to the ones of Table 5. As shown in the third line of each tables, the residue is always less than or equal to 10^{-8}.

We compared our results with the ones obtained by the FISTA in the same situations and same values of the parameters as described above. Figure 9 shows a comparison between the number of surplus and missed emitters recovered (Error+, Error− respectively) by Algorithm 3 and the FISTA in the case of the cross image and different resolution. We remark that the levels of the location and exact recoveries achieved by the FISTA are lower than the ones obtained by Algorithm 3, at least for values of the noise near .01. In particular, by the FISTA the Error+ is always above 410, whereas by Algorithm 3 is zero for small value of the noise. On the other hand,

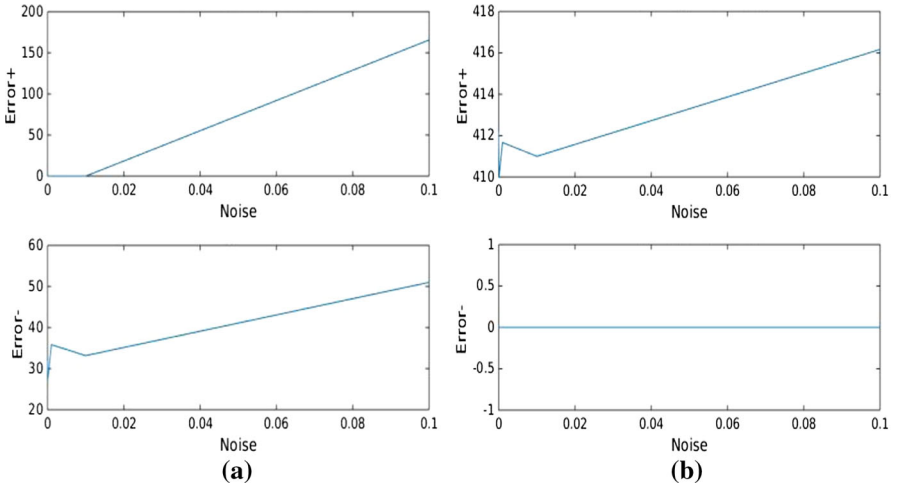

Fig. 9 Graphics of Error+ (surplus of emitters), Error- (missed emitters) against noise. **a** $p = .1$, $\beta = 10^{-6}$ by Algorithm 3, **b** $p = .1$, $\beta = 10^{-4}$ by FISTA

FISTA is faster than our algorithm (as expected, since our algorithm solves a nonlinear equation for each minimization problem.).

Acknowledgements Open access funding provided by University of Graz. We thank the referees for very thoughtful suggestions and remarks, which helped to improve our results.

References

1. Artina, M., Fornasier, M., Solombrino, F.: Linearly constrained nonsmooth and nonconvex minimization. SIAM J. Optim. **23**, 1904–1937 (2013)
2. Babcock, H.P., Moffitt, J.R., Cao, Y., Zhuang, X.: Fast compressed sensing analysis for super-resolution imaging using L1-homotopy. Opt. Express **21**, 28583–28596 (2013)
3. Barenblatt, G.I.: The mathematical theory of equilibrium cracks in brittle fracture. Adv. Appl. Math. Mech. **7**, 55–129 (1962)
4. Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., Hess, H.F.: Imaging intracellular fluorescent proteins at nanometer resolution. Science **313**, 1642–1645 (2006)
5. Black, M.J., Rangarajan, A.: On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. Int. J. Comput. Vis. **19**, 57–91 (1996)
6. Bredies, K., Lorentz, D.A., Reiterer, S.: Minimization of non-smooth, nonconvex functionals by iterative thresholding. J. Optim. Theory Appl. **165**, 78–112 (2015)
7. Candes, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**, 1207–1223 (2006)
8. Candes, E.J., Wakin, M.B., Byod, S.: Enhancing sparsity by reweighted ℓ_1 minimization. J. Fourier Anal. Appl. **14**, 877–905 (2008)
9. Casas, E., Clason, C., Kunisch, K.: Approximation of elliptic control problems in measure spaces with sparse solutions. SIAM J. Control Optim. **50**, 1735–1752 (2012)

10. Chartrand, R.: Exact reconstruction of sparse signals via noconvex minimization. IEEE Signal Process. Lett. **14**, 707–710 (2007)
11. Chartrand, R.: Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data. In: IEEE Interantional Symposium on Biomedical Imaging: From Nano to Macro (2009)
12. Chartrand, R., Staneva, V.: Restricted isometry properties and nonconvex compressing sensing. Inverse Probl. **24**, 035020 (2008)
13. Chartrand, R., Yin, W.: Iteratively reweighted algorithms for compressing sensing. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2008)
14. Chen, X., Zhou, W.: Convergence of the reweighted ℓ_1 minimization algorithm for ℓ_2-ℓ_p minimization. Comput. Optim. Appl. **59**, 47–61 (2014)
15. Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of ℓ^2-ℓ^p minimization. SIAM J. Sci. Comput. **32**, 2832–2852 (2010)
16. Dugdale, D.S.: Yielding of steel sheets containing slits. J. Mech. Phys. Solids **8**, 100–104 (1960)
17. Duval, V., Peyré, G.: Exact support recovery for sparse spikes deconvolution. Found. Comput. Math. **15**, 1315–1355 (2015)
18. Fornasier, M., Ward, R.: Iterative thresholing meets free-discontinuity problems. Found. Comput. Math. **10**, 527–567 (2015)
19. Foucart, S., Lai, M.-J.: Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. Appl. Comput. Harmon. Anal. **26**, 395–407 (2009)
20. Ghilli, D., Kunisch, K.: A monotone scheme for sparsity optimization in ℓ^p with $p \in (0, 1]$. In: IFAC WC Proceedings (2017)
21. Gu, L., Sheng, Y., Chen, Y., Chang, H., Zhang, Y., Lv, P., Ji, W., Xu, T.: High-density 3D single molecular analysis based on compressed sensing. Biophys. J. **106**, 2443–2449 (2014)
22. Herzog, R., Stadler, G., Wachsmuth, G.: Directional sparsity in optimal control of partial differential equations. SIAM J. Control Optim. **50**, 943–963 (2012)
23. Hess, S.T., Girirajan, T.P., Mason, M.D.: Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. Biophys. J. **91**, 4258–4272 (2006)
24. Hintermüller, M.: Wu, Tao: Nonconvex TV^q-models in image restoration: analysis and a trust-region regularization-based superlinearly convergent solver. SIAM J. Imaging Sci. **6**, 1385–1415 (2013)
25. Huang, B., Babcock, H.P., Zhuang, X.: Breaking the diffraction barrier: super-resolution imaging of cells. Cell **143**, 1047–1058 (2010)
26. Huang, J., Mumford, D.: Statistics of natural images and models. In: International Conference on Computer Vision and Pattern Recognition (CVPR), Fort Collins, pp. 541–547 (1999)
27. Ito, K., Kunisch, K.: A variational approach to sparsity optimization based on Lagrange multiplier theory. Inverse Probl. **30**, 015001 (2014)
28. Ito, K., Kunisch, K.: Lagrange multiplier approach to variational problems and applications. In: Advances in Design and Control 15. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2008)
29. Jiao, Y., Jin, B., Lu, X.: A primal dual active set with continuation algorithm for the ℓ^0-regularized optimization problem. Appl. Comput. Harmon. Anal. **39**, 400–426 (2015)
30. Jiao, Y., Jin, B., Lu, X., Ren, W.: A primal dual active set algorithm for a class of nonconvex sparsity optimization (2013) **(preprint)**
31. Jones, S.A., Shim, S.-H., He, J., Zhuang, X.: Fast, three-dimensional super-resolution imaging of live cells. Nat. Methods **8**, 499–505 (2011)
32. Kalise, D., Kunisch, K., Rao, Z.: Infinite horizon sparse optimal control. J. Optim. Theory Appl. **172**, 481–517 (2017)
33. Kim, K., Min, J., Carlini, L., Unser, M., Manley, S., Jeon, D., Ye, J.C.: Fast maximum likelihood high-density low-SNR super-resolution localization microscopy. In: International Conference on Sampling Theory and Applications, Bremen, Federal Republic of Germany, pp. 285–288 (2013)
34. Lai, M.-J., Wang, J.: An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems. SIAM J. Optim. **21**, 82–101 (2011)
35. Lai, M.-J., Xu, Y., Yin, W.: Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. SIAM J. Numer. Anal. **51**, 927–957 (2013)
36. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**, 2434–2460 (2014)
37. Lu, Z.: Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming. Math. Program Ser. A **147**, 277–307 (2014)

38. Nieuwenhuizen, R.P.J., Lidke, K.A., Bates, M., Puig, D.L., Grünwald, D., Stallinga, S., Rieger, B.: Measuring image resolution in optical nanoscopy. Nat. Methods **10**, 557–562 (2013)
39. Nikolova, M.: Minimizers of const-functions involving nonsmooth data-fidelity terms: applications to the processing of outliers. SIAM J. Numer. Anal. **40**, 965–994 (2002)
40. Nikolova, M., Ng, M.K., Tam, C.-P.: Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. IEEE Trans. Image Process. **19**, 3073–3088 (2010)
41. Nikolova, M., Ng, M.K., Zhang, S., Ching, W.-K.: Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. SIAM J. Imaging Sci. **1**, 2–25 (2008)
42. Ochs, P., Dosovitskiy, A., Brox, T., Pock, T.: On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. SIAM J. Imaging Sci. **8**, 331–372 (2015)
43. Del Piero, G.: A variational approach to fracture and other inelastic phenomena. J. Elast. **112**, 3–77 (2013)
44. Ramlau, R., Zarzer, C.: On the minimization of a Tikhonov functional with non-convex sparsity constraints. Electron. Trans. Numer. Anal. **39**, 476–507 (2012)
45. Roth, S., Black, M.J.: Fields of experts. Int. J. Comput. Vis. **82**, 205–229 (2009)
46. Rust, M., Bates, M., Zhuang, X.: Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat. Methods **3**, 793–796 (2006)
47. Shroff, H., Galbraith, C.G., Galbraith, J.A., Betzig, E.: Live-cell photoactivated localization microscopy of nanoscale adhesion dynamics. Nat. Methods **5**, 417–423 (2008)
48. Stadler, G.: Elliptic optimal control problems with L1-control cost and applications for the placement of control devices. Comput. Optim. Appl. **44**, 159–181 (2009)
49. Sun, Q.-: Recovery of sparsest signals via ℓ^q-minimization. Appl. Comput. Hamon. Anal. **32**, 329–341 (2012)
50. Zhu, L., Zhang, W., Elnatan, D., Huang, B.: Faster STORM using compressed sensing. Nat. Methods **9**, 721–723 (2012)
51. Zoubir, A., Koivunen, V., Chakhchoukh, Y., Muma, M.: Robust estimation in signal processing: a tutorial-style treatment of fundamental concepts. IEEE Signal Process. Mag. **29**, 61–80 (2012)
52. Zuo, W., Meng, D., Zhang, L., Feng, X., Zhang, D.: A generalized iterated shrinkage algorithm for non-convex sparse conding. In: IEEE International Conference on Computer Vision, pp. 217–224 (2013)