Inverse Problems

PAPER

Modulus-based iterative methods for constrained $\ell_p - \ell_q$ minimization

To cite this article: A Buccini *et al* 2020 *Inverse Problems* **36** 084001

IOP Publishing Inverse Problems

Inverse Problems **36** (2020) 084001 (26pp) https://doi.org/10.1088/1361-6420/ab9f86

Modulus-based iterative methods for constrained ℓ_p–ℓ_q minimization

A Buccini[1], **M Pasha**[2] **and L Reichel**[2,3]

[1] Dipartimento di Matematica e Informatica, Università di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy
[2] Department of Mathematical Sciences, Kent State University, Kent, OH 44242, United States of America

E-mail: reichel@math.kent.edu

Received 1 December 2019, revised 3 June 2020
Accepted for publication 23 June 2020
Published 20 August 2020

Abstract

The need to solve discrete ill-posed problems arises in many areas of science and engineering. Solutions of these problems, if they exist, are very sensitive to perturbations in the available data. Regularization replaces the original problem by a nearby regularized problem, whose solution is less sensitive to the error in the data. The regularized problem contains a fidelity term and a regularization term. Recently, the use of a p-norm to measure the fidelity term and a q-norm to measure the regularization term has received considerable attention. The balance between these terms is determined by a regularization parameter. In many applications, such as in image restoration, the desired solution is known to live in a convex set, such as the nonnegative orthant. It is natural to require the computed solution of the regularized problem to satisfy the same constraint(s). This paper shows that this procedure induces a regularization method and describes a modulus-based iterative method for computing a constrained approximate solution of a smoothed version of the regularized problem. Convergence of the iterative method is shown, and numerical examples that illustrate the performance of the proposed method are presented.

Keywords: modulus-based method, constrained minimization, ill-posed problem, sparse approximation

1. Introduction

Many applications in science and engineering require the solution of minimization problems of the form

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_p^p, \tag{1}$$

[3] Author to whom any correspondence should be addressed.

where $A \in \mathbb{R}^{m\times n}$ is a large matrix, whose singular values 'cluster' at the origin. Matrices of this kind arise, for instance, from the discretization of Fredholm integral equations of the first kind. The minimization problem (1) is a so-called discrete ill-posed problem; see, e.g., [1–3] for discussions on this kind of problems. The vector $b \in \mathbb{R}^m$ represents measured data that are contaminated by an (unknown) error $e \in \mathbb{R}^m$ that may stem from measurement or discretization inaccuracies. Our solution methods allow $m \geqslant n$ as well as $m < n$.

When $p = 2$, the minimization problem (1) is a linear least-squares problem. We also are interested in computing approximate solutions of (1) when $0 < p < 2$. The choice of p should be informed by the type of error e in b; see below.

Letting $p = 2$ is appropriate when the error e in b can be modeled by white Gaussian noise. However, when the error is non-Gaussian, e.g., when b is contaminated by impulse noise, the use of the Euclidean norm is not effective. When $p \geqslant 1$, the expression

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p}, \quad x = [x_1, x_2, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$$

is a norm. The mapping $x \mapsto \|x\|_p$ is not a norm for $0 < p < 1$, since it does not satisfy the triangle inequality. Nevertheless, minimization of (1) for these values of p also is of interest; see, e.g., [4–6]. For simplicity, we will refer to the mapping $x \mapsto \|x\|_p$ as a norm for all $p > 0$.

Let $b_{\text{true}} \in \mathbb{R}^m$ denote the unknown error-free vector associated with b, i.e.,

$$b = b_{\text{true}} + e.$$

We would like to compute the solution of minimal norm, x_{true}, of the minimization problem (1) with b replaced by b_{true}. We assume that b_{true} is in the range of A, denoted by $\mathcal{R}(A)$. Since the singular values of A 'cluster' at the origin, the matrix A in (1) is numerically rank deficient. The minimization problem (1) therefore might not have a solution or the solution might not be an accurate approximation of x_{true} due to severe propagation of the error e in b into the computed solution. To remedy these difficulties, at least in part, we replace the minimization problem (1) by a penalized minimization problem of the form

$$x_\mu := \arg\min_x \mathcal{J}_\mu(x), \tag{2}$$

where

$$\begin{aligned}
\mathcal{J}_\mu(x) &:= \Phi_{\text{fid}}(x) + \mu\Phi_{\text{reg}}(x),\\
\Phi_{\text{fid}}(x) &:= \frac{1}{p}\|Ax - b\|_p^p = \frac{1}{p}\sum_{i=1}^{m} \phi_p((Ax-b)_i),\\
\Phi_{\text{reg}}(x) &:= \frac{1}{q}\|Lx\|_q^q = \frac{1}{q}\sum_{j=1}^{s} \phi_q((Lx)_j),
\end{aligned}$$

and $0 < p, q \leqslant 2$. This replacement is known as *regularization*. The function $\phi_\gamma : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ is given by

$$\phi_\gamma(t) = |t|^\gamma, \quad \gamma \in \mathbb{R}, \tag{3}$$

and the matrix $L \in \mathbb{R}^{s\times n}$ is referred to as the regularization matrix. Common choices of L include the identity, a finite difference matrix, or a framelet operator; see, e.g., [6–8]. The

regularization parameter $\mu > 0$ balances the influence of the fidelity term Φ_{fid} and the regularization term Φ_{reg} in (2). Let $\mathcal{N}(M)$ denote the null space of the matrix M. It is desirable that L be chosen so that

$$\mathcal{N}(A) \cap \mathcal{N}(L) = \{0\}, \tag{4}$$

because then the minimization problem (2) has a unique solution for any $\mu > 0$ when $p, q > 1$.

Minimization problems of the form (2) arise in a wide variety of research areas, such as in numerical linear algebra [9, 10], image restoration [5, 6, 8], compressed sensing [11–13], pattern recognition [14], and matrix completion [15].

It is in general beneficial to impose the same constraints on the computed solution that the desired solution, x_{true}, is known to satisfy. For example, in image restoration problems, the entries of x_{true} represent pixel values of the image. Pixels are nonnegative and, therefore, one generally obtains a more accurate approximation of x_{true} when solving the constraint minimization problem

$$x_\mu^+ := \arg\min_{x \geqslant 0} \mathcal{J}_\mu(x) \tag{5}$$

than when solving the unconstrained problem (2). In the present paper we first show that (5) induces a regularization method whenever the regularization parameter μ is chosen appropriately with respect to the noise. Then, we describe a solution method for a smoothed version of (5) that is based on the modulus iterative method [16]. As numerically shown in [17] the difference between the results obtained considering the smoothed and the original functional is negligible in terms of quality of the computed reconstructions.

To the best of our knowledge, this is the first time that the regularization properties of ℓ_p–ℓ_q minimization have been investigated. Moreover, the constrained version (5) of the model (2) has not been proposed before, and the majorization–minimization algorithm has never been combined with the modulus method. Particular cases of this regularization technique have been analyzed in [18–20].

The organization of this paper is as follows. Section 2 proves that the described minimization scheme is a regularization method. Section 3 outlines the majorization–minimization generalized Krylov subspace (MM-GKS) method proposed in [5] for the solution of a smoothed version of the unconstrained minimization problem (2). Section 4 reviews modulus-based methods for constrained optimization problems. Modulus-based methods for the solution of a smoothed version of (5) are described in section 5, which discusses two approaches: the first approach uses nested generalized Krylov subspaces and applies the modulus-based method in these subspaces. The second approach is well suited for minimization problems (5) in which A is a block-circulant-circulant-block (BCCB) matrix. Then the fast Fourier transform (FFT) can replace the generalized Krylov subspace method. This replacement reduces the computational cost. Section 6 shows the convergence of the proposed methods. Illustrative numerical examples are presented in section 7, and section 8 contains concluding remarks.

2. Regularization property

In this section we discuss the regularization properties of (5). In particular, we would like to show that, when the norm of the noise e goes to 0, the minimizers of (5) converge to a desirable solution of the noise-free problem. This kind of result is standard in the theory of inverse problems; see, e.g., [1]. The proofs presented here are similar to, and the results can be

derived from, the ones in [21]. We present the proofs for the convenience of the reader. Before showing the regularization properties, we need two auxiliary results.

Lemma 1. *Let $\{x_j\}_{j\in\mathbb{N}}$ be a sequence of elements of $\mathbb{R}^n$ and let $q > 0$. If the $\|x_j\|_q^q$ are uniformly bounded, i.e., if there exists a constant $c > 0$ independent of j such that*

$$\|x_j\|_q^q \leqslant c \quad \forall\ j \in \mathbb{N},$$

then $\|x_j\|_2^2$ is uniformly bounded.

Proof. By definition of the q-norm we have

$$c \geqslant \|x_j\|_q^q = \sum_{i=1}^{n} |(x_j)_i|^q,$$

where $(x_j)_i$ denotes the ith component of x_j. Thus,

$$c \geqslant |(x_j)_i|^q \quad \forall\ 1 \leqslant i \leqslant n,\ j \in \mathbb{N},$$

which yields

$$c^{1/q} \geqslant |(x_j)_i| \quad \forall\ 1 \leqslant i \leqslant n,\ j \in \mathbb{N}.$$

We can now bound the two-norm by

$$\|x_j\|_2^2 = \sum_{i=1}^{n} (x_j)_i^2 \leqslant \sum_{i=1}^{n} c^{2/q} = nc^{2/q}.$$

□

Let

$$\Omega_0 = \{x \in \mathbb{R}^n : x_i \geqslant 0, \quad i = 1, 2, \ \ldots, \ n\}$$

denote the nonnegative orthant, and define the indicator function i_0 for Ω_0,

$$i_0(x) = \begin{cases} 0 & \text{if } x \in \Omega_0, \\ \infty & \text{else.} \end{cases} \tag{6}$$

We can rewrite the minimization problem (5) as

$$\min_{x\in\mathbb{R}^n} \widehat{\mathcal{J}}_\mu(x), \quad \widehat{\mathcal{J}}_\mu(x) = \mathcal{J}_\mu(x) + i_0(x). \tag{7}$$

Let us first show that $\widehat{\mathcal{J}}_\mu$ admits a minimizer.

Lemma 2. *Let condition (4) hold, then the functional $\widehat{\mathcal{J}}_\mu$ defined in (7) admits a global minimizer.*

Proof. It is immediate that the functional $\widehat{\mathcal{J}}_\mu$ is lower semi-continuous, proper, and coercive. Thus, there exists an $x \in \mathbb{R}^n$ such that $\widehat{\mathcal{J}}_\mu(x) < \infty$.

Let

$$\varphi = \inf_{x\in\mathbb{R}^n} \widehat{\mathcal{J}}_\mu(x).$$

There exists a constant M and a sequence $\{x_j\}_j$ such that $\widehat{\mathcal{J}}_\mu(x_j) \to \varphi$ as $j \to \infty$ and $\widehat{\mathcal{J}}_\mu(x_j) \leqslant M$ for all j. In particular, $\|Ax - b\|_p^p \leqslant Mp$ and $\|Lx\|_q^q \leqslant M\frac{q}{\mu}$. With a similar argument as in lemma 1, we have that there are two constants c_1 and c_2 such that

$$\|Ax_j - b\|_2^2 \leqslant c_1 \quad \text{and} \quad \|Lx_j\|_2^2 \leqslant c_2 \quad \forall\ j.$$

Thanks to (4), it is easy to see that there is a constant c such that

$$\|x_j\|_2^2 \leqslant c \quad \forall\ j,$$

i.e., the sequence $\{x_j\}_j$ is uniformly bounded. Hence, it admits a convergent subsequence $\{x_{j_k}\}_{j_k}$. Let $\bar{x}$ be the limit of the subsequence $\{x_{j_k}\}_{j_k}$. We have that

$$\varphi \leqslant \widehat{\mathcal{J}}_\mu(\bar{x}) \leqslant \liminf_{j_k\to\infty} \widehat{\mathcal{J}}_\mu(x_{j_k}) = \lim_{j_k\to\infty} \widehat{\mathcal{J}}_\mu(x_{j_k}) = \varphi,$$

i.e., $\bar{x}$ is a minimizer of $\widehat{\mathcal{J}}_\mu$. □

We are in position to show our main result.

Theorem 3. *Consider the minimization problem* (7) *with* $0 < p, q \leqslant 2$. *Let* $\mathcal{S}$ *denote the set of nonnegative solutions of the noise-free problem associated with* (1)*, i.e.,*

$$\mathcal{S} = \{x \in \mathbb{R}^n :\ Ax = b_{\text{true}} \text{ and } x \in \Omega_0\}.$$

Assume that $\mathcal{S}$ *is non-empty. Let* $\{b_j\}_{j\in\mathbb{N}}$ *be a sequence of vectors in* $\mathbb{R}^m$ *such that* $\|b_j - b_{\text{true}}\|_p \leqslant \delta_j \to 0$ *as* $j \to \infty$, *and let* $\{\mu_j\}_{j\in\mathbb{N}}$ *be a sequence of positive real numbers such that*

$$\mu_j \to 0 \quad \text{and} \quad \frac{\delta_j^p}{\mu_j} \to 0 \quad \text{as} \quad j \to \infty.$$

For all j, *let* x_j *denote a global minimizer of*

$$\widehat{\mathcal{J}}_j(x) = \frac{1}{p}\|Ax - b_j\|_p^p + \frac{\mu_j}{q}\|Lx\|_q^q + i_0(x).$$

There exists a convergent subsequence of $\{x_j\}_{j\in\mathbb{N}}$, *denoted by* $\{x_m\}_{m\in\mathbb{N}}$, *such that*

$$x_m \to x^* \quad \text{as} \quad m \to \infty,$$

where

$$x^* \in \arg\min_{x\in\mathcal{S}} \|Lx\|_p^p.$$

Proof. First, let us observe that the sequence $\{x_j\}_{j\in\mathbb{N}}$ is well defined thanks to lemma 2. Since x_j is a global minimizer of $\widehat{\mathcal{J}}_j$, we have that

$$\widehat{\mathcal{J}}_j(x_j) \leqslant \widehat{\mathcal{J}}_j(x) \quad \forall\ x \in \mathbb{R}^n.$$

In particular, let $x^\dagger \in \arg\min_{x\in\mathcal{S}} \|Lx\|_q^q$. Then

$$\widehat{\mathcal{J}}_j(x_j) \leqslant \widehat{\mathcal{J}}_j(x^\dagger). \tag{8}$$

Observe that obviously $x_j \in \Omega_0$ and that $x^\dagger \in \Omega_0$ by definition. Thus $i_0(x_j) = i_0(x^\dagger) = 0$. This, combined with (8) and the definition of δ_j, implies

$$\frac{1}{p}\|Ax_j - b_j\|_p^p + \frac{\mu_j}{q}\|Lx_j\|_q^q \leqslant \frac{1}{p}\|Ax^\dagger - b_j\|_p^p + \frac{\mu_j}{q}\|Lx^\dagger\|_q^q \leqslant \frac{\delta_j^p}{p} + \frac{\mu_j}{q}\|Lx^\dagger\|_q^q. \tag{9}$$

The inequality above shows that the sequences $\{\|Ax_j - b_j\|_p^p\}_{j\in\mathbb{N}}$ and $\{\|Lx_j\|_q^q\}_{j\in\mathbb{N}}$ are uniformly bounded. Thanks to lemma 1, we also have that the sequences $\{\|Ax_j - b_j\|_2^2\}_{j\in\mathbb{N}}$ and $\{\|Lx_j\|_2^2\}_{j\in\mathbb{N}}$ are uniformly bounded and, since $\mathcal{N}(A) \cap \mathcal{N}(L) = 0$, the sequence $\{x_j\}_{j\in\mathbb{N}}$ admits a convergent subsequence, which we also denote by $\{x_m\}_{m\in\mathbb{N}}$. Let x^* denote the limit of $\{x_m\}_{m\in\mathbb{N}}$. We first show that $Ax^* = b_{\text{true}}$. Consider

$$\begin{aligned}
0 &\leqslant \frac{1}{p}\|Ax^* - b_{\text{true}}\|_p^p \leqslant \liminf_{m\to\infty}\frac{1}{p}\|Ax_m - b_m\|_p^p \\
&\leqslant \liminf_{m\to\infty}\left\{\frac{1}{p}\|Ax_m - b_m\|_p^p + \frac{\mu_m}{q}\|Lx_m\|_q^q\right\} \\
&\leqslant \liminf_{m\to\infty}\left\{\frac{\delta_j^p}{p} + \frac{\mu_m}{q}\|Lx^\dagger\|_q^q\right\} = 0,
\end{aligned}$$

which implies that $Ax^* = b_{\text{true}}$.

We now show that

$$x^* \in \arg\min_{x\in\mathcal{S}} \|Lx\|_q^q.$$

We need to show that $x^* \in \Omega_0$, i.e., that $i_0(x^*) = 0$, and that $\|Lx^*\|_q^q = \|Lx^\dagger\|_q^q$. Consider

$$\begin{aligned}
\frac{1}{q}\|Lx^*\|_q^q + i_0(x^*) &\leqslant \liminf_{m\to\infty}\left\{\frac{1}{q}\|Lx_m\|_q^q + i_0(x_m)\right\} \\
&\leqslant \liminf_{m\to\infty}\left\{\frac{1}{p\mu_j}\|Ax_m - b_m\|_p^p + \frac{1}{q}\|Lx_m\|_q^q + i_0(x_m)\right\} \\
&\leqslant \liminf_{m\to\infty}\left\{\frac{\delta_j^p}{p\mu_j} + \frac{1}{q}\|Lx^\dagger\|_q^q\right\} = \frac{1}{q}\|Lx^\dagger\|_q^q,
\end{aligned}$$

where the last inequality follows from (9) divided by $\mu_j > 0$. Multiplying the left-hand and right-hand sides of the above inequality by $q > 0$, we obtain

$$\|Lx^*\|_q^q + i_0(x^*) \leqslant \|Lx^\dagger\|_q^q,$$

which implies $i_0(x^*) = 0$ and $\|Lx^*\|_q^q = \|Lx^\dagger\|_q^q$. This concludes the proof. □

3. A majorization–minimization solution method

We review one of the majorization–minimization (MM) methods for the solution of (2) described in [5]. In each step of this method one first determines a functional $\mathcal{Q}$ that is a quadratic tangent majorant for $\mathcal{J}_\mu(x)$, and then computes the minimum of this functional.

Definition 4 ([5]). The functional $x \mapsto \mathcal{Q}(x, v) : \mathbb{R}^n \to \mathbb{R}$ is said to be a quadratic tangent majorant for $x \mapsto \mathcal{J}_\mu(x)$ at $x = v \in \mathbb{R}^n$ if

(a) $x \mapsto \mathcal{Q}(x, v)$ is quadratic;
(b) $\mathcal{Q}(v, v) = \mathcal{J}_\mu(v)$;
(c) $\bigtriangledown_x \mathcal{Q}(v, v) = \bigtriangledown_x \mathcal{J}_\mu(v)$;
(d) $\mathcal{Q}(x, v) \geqslant \mathcal{J}_\mu(x) \quad \forall\, x \in \mathbb{R}^n$;

where $\bigtriangledown_x f$ denotes the gradient of $f = f(x)$ with respect to $x \in \mathbb{R}^n$.

The functional $\mathcal{J}_\mu(x)$ admits a quadratic majorant for $1 < p, q \leqslant 2$, but not for $0 < p \leqslant 1$ or $0 < q \leqslant 1$, since $x \mapsto \mathcal{J}_\mu(x)$ is not differentiable for the latter values of p and q, and all x. For this reason, one smooths the function (3) to make it differentiable for $\gamma \in (0, 1]$. A popular smoothed version of (3) is given by

$$\phi_{\gamma,\epsilon}(t) = \left(t^2 + \epsilon^2\right)^{\gamma/2} \quad \text{with} \quad \begin{cases} \epsilon > 0 & \text{for } 0 < \gamma \leqslant 1, \\ \epsilon = 0 & \text{for } \gamma > 1, \end{cases}$$

for some small $\epsilon > 0$. The minimization problem (2) is replaced by the smoothed problem

$$\min_{x \in \mathbb{R}^n} \mathcal{J}_{\mu,\epsilon}(x), \quad \mathcal{J}_{\mu,\epsilon}(x) := \frac{1}{p} \sum_{i=1}^{m} \phi_{p,\epsilon}((Ax - b)_i) + \frac{\mu}{q} \sum_{j=1}^{s} \phi_{q,\epsilon}((Lx)_j). \tag{10}$$

Huang *et al* [5] describe two approaches to construct a quadratic tangent majorant for (10) at an available approximate solution $x = x^{(k)}$. The majorants considered in [5] are referred to as adaptive or fixed quadratic majorants. The latter are cheaper to compute, but may give slower convergence. We develop the analysis only for fixed quadratic majorants, but all theoretical results also hold for adaptive quadratic majorants.

Let $x^{(k)}$ be an available approximate solution of (10), and introduce the vectors

$$v^{(k)} = Ax^{(k)} - b, \quad u^{(k)} = Lx^{(k)}.$$

Define

$$w_{\text{fid}}^{(k)} = v^{(k)} \left(1 - \left(\frac{(v^{(k)})^2 + \epsilon^2}{\epsilon^2}\right)^{p/2-1}\right),$$

$$w_{\text{reg}}^{(k)} = u^{(k)} \left(1 - \left(\frac{(u^{(k)})^2 + \epsilon^2}{\epsilon^2}\right)^{q/2-1}\right),$$

where all operations in the expressions on the right-hand sides, including squaring, are element-wise. It is shown in [5] that the functional

$$\mathcal{Q}(x, x^{(k)}) = \frac{\epsilon^{p-2}}{2} \left(\|Ax - b\|_2^2 - 2\langle w_{\text{fid}}^{(k)}, Ax\rangle\right) + \frac{\mu\epsilon^{q-2}}{2} \left(\|Lx\|_2^2 - 2\langle w_{\text{reg}}^{(k)}, Lx\rangle\right) + c, \tag{11}$$

where c is a suitable constant that is independent of x, is a quadratic tangent majorant for $\mathcal{J}_{\mu,\epsilon}(x)$ at $x^{(k)}$. We determine the next approximation, $x^{(k+1)}$, as the minimizer of $\mathcal{J}_{\mu,\epsilon}(x)$ by minimizing the functional $x \mapsto \mathcal{Q}(x, x^{(k)})$. It follows from (11) that

$$x^{(k+1)} = \arg\min_{x \in \mathbb{R}^n} \left[\|Ax - b\|_2^2 - 2\langle w_{\text{fid}}^{(k)}, Ax\rangle + \eta \left(\|Lx\|_2^2 - 2\langle w_{\text{reg}}^{(k)}, Lx\rangle\right)\right], \tag{12}$$

where $\eta = \mu \frac{\epsilon^{q-2}}{\epsilon^{p-2}}$. Details of the derivation of this expression are provided in [5].

Since the functional $x \mapsto \mathcal{Q}(x, x^{(k)})$ is quadratic, the minimizer $x^{(k+1)}$ can be computed by determining the zero of the gradient of the expression in the right-hand side of (12), i.e., by solving the linear system of equations

$$(A^{\mathrm{T}}A + \eta L^{\mathrm{T}}L)x = A^{\mathrm{T}}(b + w_{\mathrm{fid}}^{(k)}) + \eta L^{\mathrm{T}} w_{\mathrm{reg}}^{(k)}, \tag{13}$$

where the superscript T denotes transposition. The matrix on the left-hand side of (13) is non-singular for $\mu > 0$ when (4) holds. This condition typically is satisfied for image restoration problems, since in this application the matrix A represents a blurring operator, which is a low-pass filter, while the regularization matrix L usually is the identity matrix or a difference operator, which is a high-pass filter. For future reference, we formulate equation (13) as the equivalent least-squares problem

$$\min_{x\in\mathbb{R}^n} \left\| \begin{bmatrix} A \\ \eta^{1/2} L \end{bmatrix} x - \begin{bmatrix} b + w_{\mathrm{fid}}^{(k)} \\ \eta^{1/2} w_{\mathrm{reg}}^{(k)} \end{bmatrix} \right\|_2^2. \tag{14}$$

An algorithm for the MM method of this section is described in [[5], section 5]. This algorithm determines an approximate solution in a low-dimensional solution subspace. The dimension of this subspace is increased by one in each iteration. We will in section 5 present an extension of this algorithm to the constrained minimization problems (5).

4. Modulus-based iterative methods

In [22, 23] a constrained least-squares problem is reduced to a linear complementarity problem, which can be solved by a modulus-based iterative method. We will apply such a method to the solution of a nonnegatively constrained least-squares problem associated with (12). For the convenience of the reader, we describe the following results that are discussed by Bai [16] and Cottle *et al* [24].

Theorem 5. *Let M be a symmetric positive definite matrix. Then the nonnegatively constrained quadratic programming problem,*

$$\min_{z\geqslant 0} \left(\frac{1}{2} z^{\mathrm{T}} M z + c^{\mathrm{T}} z \right),$$

which we denote by NNQP(M, c), is equivalent to the linear complementarity problem,

$$z \geqslant 0, \quad Mz + c \geqslant 0, \quad z^{\mathrm{T}}(Mz + c) = 0,$$

which is denoted by LCP(M, c).

Corollary 6. *Let $M \in \mathbb{R}^{n\times n}$ be symmetric and positive definite and let $c \in \mathbb{R}^n$. Then the problems NNQP(M, c) and LCP(M, c) have the same unique solution.*

Corollary 7. *The nonnegative least squares (NNLS) problem*

$$\min_{z\geqslant 0} \|Gz - g\|_2$$

is equivalent to LCP$(G^{\mathrm{T}}G, -G^{\mathrm{T}}g)$, $z \geqslant 0$, $r = G^{\mathrm{T}}Gz - G^{\mathrm{T}}g \geqslant 0$, and $z^{\mathrm{T}}r = 0$. It has a unique solution when the matrix G is of full column rank.

Theorem 8. *Let $D \in \mathbb{R}^{n\times n}$ be a positive definite diagonal matrix, and define for any vector $y = [y_1, y_2, \ldots, y_n]^{\mathrm{T}} \in \mathbb{R}^n$ the vector $|y| = [|y_1|, |y_2|, \ldots, |y_n|]^{\mathrm{T}} \in \mathbb{R}^n$.*

Algorithm 1. (Modulus-based iterative method). Let $D \in \mathbb{R}^{n\times n}$ be a positive definite diagonal matrix and let $y^{(0)}$ be an initial approximate solution of (15).

for $k = 0, 1, \ldots$ **.do**
$y^{(k+1)} = (D + G^{\mathrm{T}}G)^{-1}((D - G^{\mathrm{T}}G)|y^{(k)}| + G^{\mathrm{T}}g)$;
Exit when $\|y^{(k+1)} - y^{(k)}\|_2$ is small enough;
end
$z = y^{(k+1)} + |y^{(k+1)}|$;

(*a*) *If* (z, r) *is a solution of LCP*$(G^{\mathrm{T}}G, -G^{\mathrm{T}}g)$, *then* $y = (z - D^{-1}r)/2$ *satisfies*

$$(D + G^{\mathrm{T}}G)y = (D - G^{\mathrm{T}}G)|y| + G^{\mathrm{T}}g. \tag{15}$$

(*b*) *If* y *satisfies* (15), *then* $z = |y| + y$ *and* $r = D(|y| - y)$ *is a solution of LCP*$(G^{\mathrm{T}}G, -G^{\mathrm{T}}g)$.

Proof. The theorem follows from results in [16]. □

We are interested in the situation when $D = \alpha I$ with $\alpha > 0$ in algorithm 1. Convergence can be shown when the matrix $G^{\mathrm{T}}G$ is nonsingular; see, e.g., [22, 23]. The matrix $(D + G^{\mathrm{T}}G)^{-1}$ is not explicitly formed when executing the algorithm; this is commented on further below. The iterations with the algorithm are repeated until two consecutive iterates are close enough or the maximal number of iterations is achieved. We will apply this algorithm in the following section.

5. Constrained ℓ_p–ℓ_q minimization methods

We describe the application of the modulus-based iterative method of the previous section to the ℓ_p–ℓ_q minimization problem with nonnegativity constraint (5). The method can be modified to handle other inequality constraints.

Consider the minimization problem,

$$\min_{x\geqslant 0} \frac{1}{p}\|Ax - b\|_p^p + \frac{\mu}{q}\|Lx\|_q^q. \tag{16}$$

To impose the nonnegativity constraint on the solution, we replace the functional (10) by

$$\min_{x\in\mathbb{R}^n} \widehat{\mathcal{J}}_{\mu,\epsilon}(x), \quad \widehat{\mathcal{J}}_{\mu,\epsilon}(x) = \mathcal{J}_{\mu,\epsilon}(x) + i_0(x), \tag{17}$$

where the indicator function i_0 is defined in (6). Since the functional $\widehat{\mathcal{J}}_{\mu,\epsilon}$ is not differentiable on the boundary of Ω_0, instead of constructing a quadratic tangent majorant, we define and construct a constrained quadratic tangent majorant. We will require that the majorant is quadratic in Ω_0, and that it takes on the value ∞ in $\mathbb{R}^n\backslash\Omega_0$.

Definition 9. The functional $x \mapsto \widehat{\mathcal{Q}}(x, v) : \mathbb{R}^n \to \mathbb{R}$ is said to be a constrained quadratic tangent majorant for $x \mapsto \widehat{\mathcal{J}}_{\mu,\epsilon}(x) = \mathcal{J}_{\mu,\epsilon}(x) + i_0(x)$ at $x = v \in \mathbb{R}^n$ if $\widehat{\mathcal{Q}}(x, v)$ can be expressed as

$$\widehat{\mathcal{Q}}(x, v) = \mathcal{Q}(x, v) + i_0(x),$$

where $\mathcal{Q}(x, v)$ is a quadratic tangent majorant of $\mathcal{J}_{\mu,\epsilon}(x)$ in the sense of definition 4.

For an available approximate solution $x^{(k)}$ of (5) with $\mu > 0$ and $\epsilon > 0$, the expression

$$\widehat{\mathcal{Q}}(x, x^{(k)}) = \mathcal{Q}(x, x^{(k)}) + i_0(x), \tag{18}$$

with $\mathcal{Q}$ defined in (11), is a constrained quadratic tangent majorant of the functional $x \mapsto \widehat{\mathcal{J}}_{\mu,\epsilon}(x)$ at $x = x^{(k)}$.

Proposition 10. *Let $0 < p, q \leqslant 2$ and assume that condition (4) holds. Then, for any $x^{(k)} \in \mathbb{R}^n$, the functional $\widehat{\mathcal{Q}}$ defined by (18) is a constrained quadratic tangent majorant for $\widehat{\mathcal{J}}_{\mu,\epsilon}(x)$.*

Proof. This result follows trivially from the definitions of $\widehat{\mathcal{Q}}$ and $\mathcal{Q}$. □

5.1. Minimization method for general matrices

When the matrix A is large, the computations required by algorithm 1 may be prohibitive. We will show how the computational cost can be reduced by determining an approximate solution in a generalized Krylov subspace (GKS). We remark that GKS methods have previously been applied in [5] to solve the unconstrained minimization problem (2).

The GKS method first determines an initial reduction of A to a small bidiagonal matrix by applying $1 \leqslant \ell \ll \min\{m, n\}$ steps of Golub–Kahan bidiagonalization to A with initial vector b. This gives a decomposition

$$AV_0 = U_0 B_0, \tag{19}$$

where the matrix $V_0 \in \mathbb{R}^{n\times\ell}$ has orthonormal columns that span the Krylov subspace $K_\ell(A^{\mathrm{T}}A, A^{\mathrm{T}}b) = \mathrm{span}\{A^{\mathrm{T}}b, (A^{\mathrm{T}}A)A^{\mathrm{T}}b, \ldots, (A^{\mathrm{T}}A)^{\ell-1}A^{\mathrm{T}}b\}$, the matrix $U_0 \in \mathbb{R}^{m\times(\ell+1)}$ has orthonormal columns, the first one of which is $b/\|b\|_2$, and the matrix $B_0 \in \mathbb{R}^{(\ell+1)\times\ell}$ is lower bidiagonal. It is inexpensive to compute the QR factorization $AV_0 = Q_A R_A$, where $Q_A \in \mathbb{R}^{m\times\ell}$ has orthonormal columns and $R_A \in \mathbb{R}^{\ell\times\ell}$ is upper triangular. We also compute the QR factorization $LV_0 = Q_L R_L$, where $Q_L \in \mathbb{R}^{s\times\ell}$ has orthonormal columns and $R_L \in \mathbb{R}^{\ell\times\ell}$ is upper triangular (recall that $L \in \mathbb{R}^{s\times n}$). Here we assume that $1 \leqslant \ell \leqslant s$ is small enough so that the decomposition (19) exists. This is the generic situation.

To begin with, we determine an initial approximate solution in $\mathcal{R}(V_0)$ of the least-squares problem (14). Thus, we solve the problem

$$\min_{y\in\mathbb{R}^\ell} \left\| \begin{bmatrix} AV_0 \\ \eta^{1/2} L V_0 \end{bmatrix} y - \begin{bmatrix} (b + w_{\mathrm{fid}}^{(k)}) \\ \eta^{1/2} w_{\mathrm{reg}}^{(k)} \end{bmatrix} \right\|_2^2,$$

which simplifies to

$$\min_{y\in\mathbb{R}^\ell} \left\| \begin{bmatrix} R_A \\ \eta^{1/2} R_L \end{bmatrix} y - \begin{bmatrix} Q_A^{\mathrm{T}}(b + w_{\mathrm{fid}}^{(k)}) \\ \eta^{1/2} Q_L^{\mathrm{T}} w_{\mathrm{reg}}^{(k)} \end{bmatrix} \right\|_2^2. \tag{20}$$

This gives the approximate solution

$$x^{(0)} = V_0 y^{(0)}$$

of (14), where $y^{(0)} \in \mathbb{R}^\ell$ denotes the solution of (20).

To determine an approximate solution of the constrained minimization problem in $\mathcal{R}(V_0)$, we replace (20) by

$$x^{(0)} = \arg\min_{x\geqslant 0} \left\| \begin{bmatrix} R_A \\ \eta^{1/2} R_L \end{bmatrix} V_0^{\mathrm{T}} x - \begin{bmatrix} Q_A^{\mathrm{T}}(b + w_{\mathrm{fid}}^{(k)}) \\ \eta^{1/2} Q_L^{\mathrm{T}} w_{\mathrm{reg}}^{(k)} \end{bmatrix} \right\|_2^2. \tag{21}$$

This problem is solved by the modulus-based iterative method described by algorithm 1. We apply this algorithm with the matrices and vector

$$G = \begin{bmatrix} R_A \\ \eta^{1/2} R_L \end{bmatrix} V_0^{\mathrm{T}}, \quad D = \alpha I, \quad g = \begin{bmatrix} Q_A^{\mathrm{T}}(b + w_{\mathrm{fid}}^{(k)}) \\ \eta^{1/2} Q_L^{\mathrm{T}} w_{\mathrm{reg}}^{(k)} \end{bmatrix},$$

and initial approximate solution $z^{(0)} = \max\{x^{(0)}, 0\}$, where the operation 'max' is componentwise. The parameter $\alpha > 0$ is user-defined. Its choice is discussed in [23]. The iterations with algorithm 1 can be expressed as

$$z_{j+1}^{(0)} = V_0(\alpha I + R_A^{\mathrm{T}} R_A + \eta R_L^{\mathrm{T}} R_L)^{-1} \cdot \left((\alpha I - R_A^{\mathrm{T}} R_A - \eta R_L^{\mathrm{T}} R_L) V_0^{\mathrm{T}} |z_j^{(0)}| + G^{\mathrm{T}} g \right) \tag{22}$$

for $j = 0, 1, 2, \ldots$. We remark that the matrix $(\alpha I + R_A^{\mathrm{T}} R_A + \eta R_L^{\mathrm{T}} R_L)^{-1}$ is not explicitly formed; instead, we compute the Cholesky factorization of the matrix $\alpha I + R_A^{\mathrm{T}} R_A + \eta R_L^{\mathrm{T}} R_L$.

Theorem 11. *Assume that the matrix $R_A^{\mathrm{T}} R_A + \eta R_L^{\mathrm{T}} R_L$ is of full rank. Then the sequence $\{z_j^{(0)}\}_j$ generated by iteration (22) converges to the solution of (21).*

Proof. Convergence for the situation when $V_0 = I$ is shown in, e.g., [22, 23]. It is based on the observation that the largest eigenvalue of the matrix

$$M = (\alpha I + R_A^{\mathrm{T}} R_A + \eta R_L^{\mathrm{T}} R_L)^{-1} \cdot (\alpha I - R_A^{\mathrm{T}} R_A - \eta R_L^{\mathrm{T}} R_L)$$

is strictly smaller than one. Let the columns of $\widetilde{V}_0 \in \mathbb{R}^{n \times (n-\ell)}$ be such that the matrix $W = [V_0, \widetilde{V}_0] \in \mathbb{R}^{n \times n}$ is orthogonal, and define

$$\widetilde{M} = W \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} W^{\mathrm{T}}.$$

Then the iterations (22) can be expressed as

$$z_{j+1}^{(0)} = \widetilde{M} |z_j^{(0)}| + V_0 G^{\mathrm{T}} g, \quad j = 0, 1, 2, \ldots . \tag{23}$$

The largest eigenvalue of $\widetilde{M}$ is strictly smaller than one. Therefore, the convergence proof in [22, 23] carries over to the iterations (23) and, hence, to the iterations (22). □

Let the iterations (22) terminate with the iterate $z_{j+1}^{(0)}$. An approximate solution of (21) then is furnished by

$$x_+^{(0)} = z_{j+1}^{(0)} + \left| z_{j+1}^{(0)} \right|. \tag{24}$$

Substituting (24) into (13) gives the residual vector

$$r^{(0)} = A^{\mathrm{T}} \left(A x_+^{(0)} - \left(b + w_{\mathrm{fid}}^{(0)} \right) \right) + \eta L^{\mathrm{T}} \left(L x_+^{(0)} - w_{\mathrm{reg}}^{(0)} \right).$$

We expand the solution subspace by including the scaled residual vector $v_{\mathrm{new}} = r^{(0)} / \|r^{(0)}\|_2$ in the solution subspace. Note that, at least in exact arithmetic, the vector v_{new} is orthogonal to the columns of V_0. We define the matrix $V_1 = [V_0, v_{\mathrm{new}}] \in \mathbb{R}^{n \times (\ell+1)}$, whose columns form an orthonormal basis for the expanded solution subspace. If v_{new} is not numerically orthogonal to the columns of V_0, then we reorthogonalize.

We store the matrices

$$AV_1 = [AV_0, Av_{\mathrm{new}}], \quad LV_1 = [LV_0, Lv_{\mathrm{new}}]$$

Algorithm 2. (NN-(ℓ_p–ℓ_q)$_{\text{GKS}}$). Let $A \in \mathbb{R}^{m\times n}$ and $L \in \mathbb{R}^{s\times n}$ be such that (4) holds. Let $0 < p \leqslant 2$, and $0 < q \leqslant 2$. Fix $\epsilon > 0$, $\ell > 0$, and $\mu > 0$. Let $b \in \mathbb{R}^m$ denote the noise-corrupted data vector and let $x_+^{(0)} \in \Omega_0$ be an initial guess for the solution of (1).

$\eta = \mu \frac{\epsilon^{q-2}}{\epsilon^{p-2}}$
Generate the initial subspace basis: $V_0 \in \mathbb{R}^{n\times \ell}$ such that $V_0^{\mathrm{T}} V_0 = I$
Compute AV_0 and LV_0
Compute the QR factorizations $AV_0 = Q_A R_A$ and $LV_0 = Q_L R_L$
for $k = 0, 1, 2, \ldots$ **.do**
  $v^{(k)} = Ax_+^{(k)} - b$
  $u^{(k)} = Lx_+^{(k)}$
  $w_{\text{fid}}^{(k)} = v^{(k)} \left(1 - \left(\frac{(v^{(k)})^2+\epsilon^2}{\epsilon^2}\right)^{p/2-1}\right)$
  $w_{\text{reg}}^{(k)} = u^{(k)} \left(1 - \left(\frac{(u^{(k)})^2+\epsilon^2}{\epsilon^2}\right)^{q/2-1}\right)$
  $G = \begin{bmatrix} R_A \\ \eta^{1/2} R_L \end{bmatrix} V_0^{\mathrm{T}}$
  Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue of $G^{\mathrm{T}}G$
  $\alpha = \sqrt{\lambda_{\min}\lambda_{\max}}$
  $g = \begin{bmatrix} Q_A^{\mathrm{T}}(b + w_{\text{fid}}^{(k)}) \\ \eta^{1/2} Q_L^{\mathrm{T}} w_{\text{reg}}^{(k)} \end{bmatrix}$
  $x^{(k)} = (R_A^{\mathrm{T}}R_A + \eta R_L^{\mathrm{T}}R_L)^{-1}(R_A^{\mathrm{T}}Q_A^{\mathrm{T}}(b + w_{\text{fid}}^{(k)}) + \eta R_L^{\mathrm{T}}Q_L^{\mathrm{T}} w_{\text{reg}}^{(k)})$
  $z_0^{(k)} = \max\{x^{(k)}, 0\}$
  for $j = 0, 1, \ldots$ **.do**
    $z_{j+1}^{(k)} = V_0(\alpha I + R_A^{\mathrm{T}}R_A + \eta R_L^{\mathrm{T}}R_L)^{-1}\left((\alpha I - R_A^{\mathrm{T}}R_A - \eta R_L^{\mathrm{T}}R_L)V_0^{\mathrm{T}}|z_j^{(k)}| + G^{\mathrm{T}}g\right)$
    Exit loop when $\|z_{j+1}^{(k)} - z_j^{(k)}\|_2$ is small enough
  end
  $x_+^{(k)} = z_{j+1}^{(k)} + |z_{j+1}^{(k)}|$
  Compute the residual $r^{(k)} = A^{\mathrm{T}}(Ax_+^{(k)} - (b + w_{\text{fid}}^{(k)})) + \eta L^{\mathrm{T}}(Lx_+^{(k)} - w_{\text{reg}}^{(k)})$
  Reorthogonalize, if needed, $r^{(k)} = r^{(k)} - V_k V_k^{\mathrm{T}} r^{(k)}$
  Enlarge the solution subspace with $v_{\text{new}} = \frac{r^{(k)}}{\|r^{(k)}\|_2}$
  $V_{k+1} = [V_k, v_{\text{new}}]$
  Update the QR factorizations $AV_{k+1} = Q_A R_A$ and $LV_{k+1} = Q_L R_L$
end

and compute their QR factorizations by updating the QR factorizations of AV_0 and LV_0 according to

$$AV_1 = [AV_0, Av_{\text{new}}] = [Q_A, \tilde{q}_A]\begin{bmatrix} R_A & r_A \\ 0^{\mathrm{T}} & \tau_A \end{bmatrix}, \tag{25}$$

$$LV_1 = [LV_0, Lv_{\text{new}}] = [Q_L, \tilde{q}_L]\begin{bmatrix} R_L & r_L \\ 0^{\mathrm{T}} & \tau_L \end{bmatrix}, \tag{26}$$

where

$$\begin{aligned} r_A &= Q_A^{\mathrm{T}}(Av_{\text{new}}), \quad q_A = Av_{\text{new}} - Q_A r_A, \quad \tau_A = \|q_A\|_2, \quad \tilde{q}_A = q_A/\tau_A, \\ r_L &= Q_L^{\mathrm{T}}(Lv_{\text{new}}), \quad q_L = Lv_{\text{new}} - Q_L r_L, \quad \tau_L = \|q_L\|_2, \quad \tilde{q}_L = q_L/\tau_L; \end{aligned}$$

see [25] for details on updating the QR factorization of a matrix. We now apply the modulus-based iterations (22) with R_A and R_L replaced by the upper triangular matrices in the QR factorizations (25) and (26), respectively, and use the initial iterate $x_+^{(0)}$. The modulus-based

Algorithm 3. (NN-(ℓ_p–ℓ_q)$_{\text{FFT}}$). Let $A \in \mathbb{R}^{n\times n}$ be a BCCB matrix and $W \in \mathbb{R}^{s\times n}$ be an analysis operator such that $W^{\mathrm{T}}W = I$. Let $0 < p \leqslant 2$, and $0 < q \leqslant 2$. Fix $\epsilon > 0$ and $\mu > 0$. Let $b \in \mathbb{R}^n$ denote the noise-corrupted data vector and let $x_+^{(0)} \in \Omega_0$ be an initial guess for the solution of (1). Let F denote the Fourier matrix such that $A = F^*\Sigma F$.

$\eta = \mu\frac{\epsilon^{q-2}}{\epsilon^{p-2}}$
Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue of $A^{\mathrm{T}}A$
$\alpha = \sqrt{(\lambda_{\min}+\eta)(\lambda_{\max}+\eta)}$
for $k = 0, 1, 2, \ldots$ **do**
  $v^{(k)} = Ax_+^{(k)} - b$
  $u^{(k)} = Wx_+^{(k)}$
  $w_{\text{fid}}^{(k)} = v^{(k)}\left(1 - \left(\frac{(v^{(k)})^2+\epsilon^2}{\epsilon^2}\right)^{p/2-1}\right)$
  $w_{\text{reg}}^{(k)} = u^{(k)}\left(1 - \left(\frac{(u^{(k)})^2+\epsilon^2}{\epsilon^2}\right)^{q/2-1}\right)$
  $x^{(k)} = F^*(\Sigma^*\Sigma + \eta I)^{-1}(\Sigma^*F(b + w_{\text{fid}}^{(k)}) + \eta FW^{\mathrm{T}}w_{\text{reg}}^{(k)})$
  $z_0^{(k)} = \max\{x^{(k)}, 0\}$
  for $j = 0, 1, \ldots$ **do**
    $z_{j+1}^{(k)} = F^*(\alpha I + \Sigma^*\Sigma + \eta I)^{-1}\left(((\alpha-\eta)I - \Sigma^*\Sigma)F|z_j^{(k)}| + \Sigma^*F(b + w_{\text{fid}}^{(k)}) + \eta FW^{\mathrm{T}}w_{\text{reg}}^{(k)}\right)$
    Exit loop when $\|z_{j+1}^{(k)} - z_j^{(k)}\|_2$ is small enough
  end
  $x_+^{(k+1)} = z_{j+1}^{(k)} + |z_{j+1}^{(k)}|$
end

iterations give a new approximate solution $x_+^{(1)}$ of (16), a new associated residual vector

$$r^{(1)} = A^{\mathrm{T}}\left(Ax_+^{(1)} - \left(b + w_{\text{fid}}^{(1)}\right)\right) + \eta L^{\mathrm{T}}\left(Lx_+^{(1)} - w_{\text{reg}}^{(1)}\right),$$

and a new solution subspace defined by the range of the matrix

$$V_2 = \left[V_1, r^{(1)}/\|r^{(1)}\|_2\right].$$

The computations proceed in this manner until an approximate solution of (16) with desired accuracy has been determined. Details of the computations are described by algorithm 2.

We observe that algorithm 2 requires the computation of the smallest and largest eigenvalues of $G^{\mathrm{T}}G$. Thanks to the projection into the generalized Krylov subspace, the matrix $G^{\mathrm{T}}G$ is of fairly small dimension. Therefore, these eigenvalues can be estimated very cheaply. This can be done by computing the largest and the smallest singular values of G, for instance, by using the method described in [26]. Iterations in the j-loop are terminated when two consecutive iterates are close enough. The stopping criterion is described in section 7; see (32).

5.2. Minimization method for BCCB matrices

When deblurring images with large black areas close to the edges of the image, the blurring matrix $A \in \mathbb{R}^{n\times n}$ often can be chosen to be a BCCB (block circulant with circulant blocks) matrix without affecting the quality of the restoration in a negative way. Many astronomical images allow the use of a BCCB blurring matrix; see, e.g., [27] for a discussion. The advantage of using a BCCB blurring matrix A is that it can be diagonalized by the unitary two-dimensional Fourier matrix $F \in \mathbb{C}^{n\times n}$. Thus,

$$A = F^*\Sigma F, \tag{27}$$

where the matrix Σ is diagonal, possibly with complex diagonal entries. The superscript * denotes transposition and complex conjugation. The factorization (27) can be computed in $\mathcal{O}(n\ \log_2(n))$ arithmetic floating point operations (flops) and allows at each step of the algorithm to transform the ℓ_2–ℓ_2 minimization problem (14), whose solution provides an approximation of the solution of the original ℓ_p–ℓ_q minimization problem, to a diagonal system. This makes the application of the modulus-based iterations (algorithm 1) inexpensive when $D = \alpha I$ for special regularization matrices L. In the computed examples reported in section 7, we let $L = W$ be an analysis operator defined by the transformation to a framelet domain; see section 7 for details. Here it suffices to note that $W^{\mathrm{T}}W = I$. The matrix $G^{\mathrm{T}}G$ in the modulus-based method then can be expressed as

$$G^{\mathrm{T}}G = A^{\mathrm{T}}A + \eta L^{\mathrm{T}}L = A^{\mathrm{T}}A + \eta W^{\mathrm{T}}W = F^*(\Sigma^*\Sigma + \eta I)F.$$

The evaluation of matrix-vector products with the matrix in the right-hand side requires only $\mathcal{O}(n\ \log_2(n))$ flops when using the FFT. Details of the computations are described by algorithm 3. We remark that the matrix F is not explicitly formed; only matrix-vector products with F and F^* are evaluated.

The stopping criterion for the j-loop is described by (32).

6. Convergence

This section shows convergence of the iterates generated by the modulus-based constrained ℓ_p–ℓ_q minimization method. We focus on the method described in subsection 5.1 and comment at the end of this section on the convergence of the method discussed in subsection 5.2. The proofs below extend results in [5] on the convergence of the unconstrained ℓ_p–ℓ_q minimization method to allow constraints. Several of the proofs are analogous to those in [5]. For the convenience of the reader, we provide enough details to make the present paper self-contained.

Assume that the condition of theorem 11 holds. Then the nonnegative approximate solutions $x_+^{(0)}, x_+^{(1)}, x_+^{(2)}, \ \ldots$ of (17) defined in subsection 5.1 exist. We note that $x_+^{(0)}$ is obtained by an element of a subspace of $\mathbb{R}^n$ of dimension ℓ, and, more generally, $x_+^{(j)}$ lives in some subspace of $\mathbb{R}^n$ for $j = 1, 2, 3, \ldots$. For $j \geqslant n - \ell$, the approximate solutions $x_+^{(j)}$ live in $\mathbb{R}^n$. Thus, for large values of j all iterates are in the same space. This simplifies the convergence analysis. Of course, the rate of convergence of the iterates $x_+^{(j)}$ for small j to the desired solution x_{true} is affected by the subspaces, in which the $x_+^{(j)}$ for j small live. In the following we will assume that enough steps of the algorithm have been performed so that this does not constitute an issue. We may require $n - \ell$ steps of the algorithm to be performed for the following results to hold. However, in practical application, this is never the case and convergence is reached within a reasonable number of iterations.

Proposition 12. *Let* $0 < p, q \leqslant 2$ *and assume that condition* (4) *holds. Let* $\{x_+^{(k)}\}_{k=1}^{\infty}$ *denote the sequence of approximate solutions generated by algorithm 2. For any initial approximate solution* $x_+^{(0)} \in \Omega_0$ *and all* $k \geqslant 1$ *we have*

$$\widehat{\mathcal{Q}}(x_+^{(k+1)}, x_+^{(k)}) \leqslant \widehat{\mathcal{Q}}(x_+^{(k)}, x_+^{(k)}).$$

Proof. An analogous result for the unconstrained minimization problem (10) with $\widehat{\mathcal{Q}}$ replaced by the majorant $\mathcal{Q}$, defined by (11), is shown in [[6], lemma 5.2]. The proof of this result carries over to the functional $\widehat{\mathcal{Q}}$. □

Note that the result above holds for the constrained minimizers $x_+^{(k)}$. The exact computations of these points may require that the inner iterations in algorithms 2 and 3 be carried out an arbitrarily large number of times. However, in practice only a fairly small number of iterations are needed to ensure convergence to accurate approximations of the constrained minimizers.

Theorem 13. *Let condition (4) hold. Then, for any initial approximate solution $x_+^{(0)} \in \Omega_0$, the sequence $\{\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k)})\}_{k=0}^{\infty}$ is monotonically nonincreasing and convergent, where $\widehat{\mathcal{J}}_{\mu,\epsilon}$ is defined by (17).*

Proof. The sequence $\{\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k)})\}_{k=0}^{\infty}$ is bounded from below by zero and is monotonically nonincreasing,

$$\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k+1)}) \leqslant \widehat{\mathcal{Q}}\left(x_+^{(k+1)}, x_+^{(k)}\right) \leqslant \widehat{\mathcal{Q}}\left(x_+^{(k)}, x_+^{(k)}\right) = \widehat{\mathcal{J}}_{\mu,\epsilon}\left(x_+^{(k)}\right).$$

The first inequality and the equality follow from the fact that $\widehat{\mathcal{Q}}(x, x_+^{(k)})$ is a constrained quadratic tangent majorant of $\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k)})$, i.e., they are a consequence of proposition 10. The second inequality follows from proposition 12. Since the sequence $\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k)})$, $k = 0, 1, 2, \ldots$, is monotonically nonincreasing and bounded from below, it is convergent. □

In the remainder of this section, we investigate the behavior of the sequence of iterates $\{x_+^{(k)}\}_{k\geqslant 0}$.

Proposition 14. *Let the initial approximate solution $x_+^{(0)} \in \mathbb{R}^n$ of (17) belong to Ω_0, and let subsequent approximate solutions $x_+^{(k)}$, $k = 1, 2, 3, \ldots$, be determined as described in section 5. Then the unconstrained majorization error functional*

$$x \mapsto \varepsilon(x, x_+^{(k)}) := \mathcal{Q}(x, x_+^{(k)}) - \mathcal{J}_{\mu,\epsilon}(x)$$

has the following properties:

(*a*) $\varepsilon(x, x_+^{(k)}) \in C^1(\mathbb{R}^n)$;
(*b*) $\varepsilon(x, x_+^{(k)}) \geqslant 0 \quad \forall\, x \in \mathbb{R}^n$;
(*c*) $\varepsilon(x_+^{(k)}, x_+^{(k)}) = 0$;
(*d*) $0 = \nabla_x \varepsilon(x_+^{(k)}, x_+^{(k)})$;
(*e*) $\nabla_x \varepsilon(x_+^{(k+1)}, x_+^{(k)}) = -\nabla_x \widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k+1)})$.

Proof. The proof of this result is identical to the one of [[5], proposition 3]. □

Definition 15. A convex (not necessarily differentiable) function $f(x)$ is said to be δ-strongly convex if there is a constant $\delta > 0$, such that the function $f(x) - \frac{\delta}{2}\|x\|_2^2$ is convex. The constant δ is referred to as the modulus of strong convexity of f.

Lemma 16. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strongly convex function with modulus of strong convexity $\delta > 0$. Let $x^* \in \mathbb{R}$ be a minimizer of $f(x)$. Then*

$$\frac{\delta}{2}\|x - x^*\|_2^2 \leqslant f(x) - f(x^*) \quad \forall\, x \in \mathbb{R}^n. \tag{28}$$

Proof. A proof can be found in, e.g., [28]. □

Theorem 17. *Let condition (4) hold and let $\{x_+^{(k)}\}$ denote the sequence of the iterates generated by either algorithms 2 or 3. Then the following statements hold*:

(*a*) $\lim_{k\to\infty} \|x_+^{(k+1)} - x_+^{(k)}\|_2 = 0$;

(*b*) *There exists a convergent subsequence* $\{x_+^{(j_k)}\}$ *that converges to a point* $x^* \in \Omega_0$;
(*c*) *Let* $I = \{i : (x^*) > 0\}$. *Then* $(\nabla \mathcal{J}_{\mu,\epsilon})_i = 0$ *for* $i \in I$.

Proof. Consider the quadratic majorant function $x \mapsto \mathcal{Q}(x, x_+^{(k)})$ at step k of the iterative method. Since, thanks to (4), $x \mapsto \mathcal{Q}(x, x_+^{(k)})$ is δ-strongly convex, we can apply lemma 16. In particular, inequality (28) with the function $\mathcal{Q}(\cdot, x_+^{(k)})$ in place of $f(\cdot)$ and $x_+^{(k+1)}$ in place of x^* yields

$$\frac{\delta}{2}\left\|x - x_+^{(k+1)}\right\|_2^2 \leqslant \mathcal{Q}\left(x, x_+^{(k)}\right) - \mathcal{Q}\left(x_+^{(k+1)}, x_+^{(k)}\right) \quad \forall\, x \in \mathbb{R}^n. \tag{29}$$

The above inequality holds for all $k \geqslant 0$. Substituting x by the iterate $x_+^{(k)}$ in (29), and observing that $x_+^{(k+1)}, x_+^{(k)} \in \Omega_0$, we obtain

$$\begin{aligned}
\frac{\delta}{2}\|x_+^{(k)} - x_+^{(k+1)}\|_2^2 &\leqslant \mathcal{Q}\left(x_+^{(k)}, x_+^{(k)}\right) - \mathcal{Q}\left(x_+^{(k+1)}, x_+^{(k)}\right) \\
&= \mathcal{J}_{\mu,\epsilon}\left(x_+^{(k)}, x_+^{(k)}\right) + \varepsilon\left(x_+^{(k)}, x_+^{(k)}\right) - \mathcal{J}_{\mu,\epsilon}\left(x_+^{(k+1)}\right) - \varepsilon\left(x_+^{(k+1)}, x_+^{(k)}\right) \\
&\leqslant \mathcal{J}_{\mu,\epsilon}\left(x_+^{(k)}\right) - \mathcal{J}_{\mu,\epsilon}\left(x_+^{(k+1)}\right) \\
&= \widehat{\mathcal{J}}_{\mu,\epsilon}\left(x_+^{(k)}\right) - \widehat{\mathcal{J}}_{\mu,\epsilon}\left(x_+^{(k+1)}\right) \quad \forall\, k,
\end{aligned}$$

where the inequality follows from proposition 14. Summing the above inequalities over k gives

$$\begin{aligned}
\sum_{k=0}^{\infty} \|x_+^{(k+1)} - x_+^{(k)}\|_2^2 &\leqslant \frac{2}{\delta}\sum_{k=0}^{\infty}\left(\widehat{\mathcal{J}}_{\mu,\epsilon}\left(x_+^{(k)}\right) - \widehat{\mathcal{J}}_{\mu,\epsilon}\left(x_+^{(k+1)}\right)\right) \\
&= \frac{2}{\delta}\left(\left(\widehat{\mathcal{J}}_{\mu,\epsilon}\left(x_+^{(0)}\right) - \cancel{\widehat{\mathcal{J}}_{\mu,\epsilon}\left(x^{(1)}\right)}\right) + \cancel{\widehat{\mathcal{J}}_{\mu,\epsilon}\left(x^{(1)}\right)} + \cdots\right) \\
&= \frac{2}{\delta}\left(\widehat{\mathcal{J}}_{\mu,\epsilon}\left(x_+^{(0)}\right) - \widehat{\mathcal{J}}_{\mu,\epsilon}^*\right),
\end{aligned} \tag{30}$$

where $\widehat{\mathcal{J}}_{\mu,\epsilon}^*$ denotes the limit point of the sequence $\{\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k)})\}_{k\geqslant 0}$. According to theorem 13, the $\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(k)})$ are nonnegative and a decreasing function of k. It follows that the limit point exists and $\widehat{\mathcal{J}}_{\mu,\epsilon}(x_+^{(0)}) - \widehat{\mathcal{J}}_{\mu,\epsilon}^*$ is nonnegative. We conclude that the series on the left-hand side of inequality (30) is convergent. Hence, statement (a) holds.

We turn to statement (b). With a similar argument as the one in the proof of lemma 2, we obtain that the sequence $\{x_+^{(k)}\}_k$ is uniformly bounded and, thus, admits a convergent subsequence $\{x_+^{(j_k)}\}_{j_k}$. Let x^* be the limit point of $\{x_+^{(j_k)}\}_{j_k}$. Since $x_+^{(j_k)} \in \Omega_0$ for all j_k and Ω_0 is closed we have that $x^* \in \Omega_0$.

We now show statement (c). Let $i \in I$, then there exists J such that $(x_+^{(j_k)})_i > 0$ for all $j_k > J$. Then, for all $j_k > J$, it holds that $\frac{\partial \mathcal{Q}}{\partial x_i}(x_+^{(j_k+1)}, x_+^{(j_k)}) = 0$, where $\frac{\partial \mathcal{Q}}{\partial x_i}(x_+^{(j_k+1)}, x_+^{(j_k)})$ denotes the partial derivative of $\mathcal{Q}$ with respect to the ith component of x. Then, the definition of ε yields $\frac{\partial \varepsilon}{\partial x_i}(x_+^{(j_k+1)}, x_+^{(j_k)}) = -\frac{\partial \mathcal{J}_{\mu,\epsilon}}{\partial x_i}(x_+^{(j_k)})$. We obtain that

$$\begin{aligned}
\left|\frac{\partial \mathcal{J}_{\mu,\epsilon}}{\partial x_i}\left(x_+^{(j_k)}\right)\right| &= \left|\frac{\partial \varepsilon}{\partial x_i}\left(x_+^{(j_k+1)}, x_+^{(k)}\right)\right| \\
&= \left|\frac{\partial \varepsilon}{\partial x_i}\left(x_+^{(j_k+1)}, x_+^{(j_k)}\right) - \frac{\partial \varepsilon}{\partial x_i}\left(x_+^{(j_k)}, x_+^{(j_k)}\right)\right|
\end{aligned}$$

$$\leqslant \|\nabla\varepsilon\left(x_+^{(j_k+1)}, x_+^{(j_k)}\right) - \nabla\varepsilon\left(x_+^{(j_k+1)}, x_+^{(j_k+1)}\right)\|$$
$$\leqslant L\|x_+^{(j_k+1)} - x_+^{(j_k)}\|,$$

where the last inequality follows from the Lipschitz continuity of the gradient of the functions $\mathcal{Q}$ and $\mathcal{J}_{\mu,\epsilon}$. Thanks to point (a) we have that $\|x_+^{(j_k+1)} - x_+^{(j_k)}\| \to 0$ as $j_k \to \infty$. Thus $\left|\frac{\partial \mathcal{J}_{\mu,\epsilon}}{\partial x_i}(x_+^{(j_k)})\right| \to 0$ as $j_k \to \infty$, i.e., $\frac{\partial \mathcal{J}_{\mu,\epsilon}}{\partial x_i}(x^*) = 0$. □

7. Numerical examples

This section presents a few computed examples that illustrate the performance of the numerical methods described in the previous sections. We consider some imaging problems. Therefore, we use a two-level framelet analysis operator as regularization operator L, since it is well-known that images have sparse representation in the framelet domain. We recall that framelets are extensions of wavelets. Following [7, 29], we define them as follows:

Definition 18. Let $W \in \mathbb{R}^{r\times n}$ with $1 \leqslant n \leqslant r$. The set of the rows of W is a framelet system for $\mathbb{R}^n$ if $\forall\, x \in \mathbb{R}^n$ it holds

$$\|x\|_2^2 = \sum_{j=1}^{r} (w_j^{\mathrm{T}} x)^2, \tag{31}$$

where $w_j \in \mathbb{R}^n$ denotes the jth row of the matrix W (written as a column vector), i.e., $W = [w_1, w_2, \ldots, w_r]^{\mathrm{T}}$. The matrix W is referred to as an analysis operator and W^{T} as a synthesis operator.

Equation (31) is equivalent to the perfect reconstruction formula

$$x = W^{\mathrm{T}} y, \quad y = Wx.$$

Thus, the matrix W defines a tight frame if and only if $W^{\mathrm{T}}W = I$. We remark that in general $WW^{\mathrm{T}} \neq I$, unless $r = n$ and the framelets are orthonormal. Observe that $\mathcal{N}(W) = \{0\}$. Therefore, (4) is trivially satisfied.

We use the same tight frames as in [7, 17, 29–31]; they are determined by linear B-splines. Specifically, for problems in one space-dimension, they are formed by a low-pass filter $W_0 \in \mathbb{R}^{n\times n}$ and two high-pass filters $W_1 \in \mathbb{R}^{n\times n}$ and $W_2 \in \mathbb{R}^{n\times n}$. The corresponding masks are given by

$$u^{(0)} = \frac{1}{4}[1, 2, 1], \quad u^{(1)} = \frac{\sqrt{2}}{4}[1, 0, -1], \quad u^{(2)} = \frac{1}{4}[-1, 2, -1].$$

The analysis operator W is determined by these masks and by imposing reflexive boundary conditions, which ensure that $W^{\mathrm{T}}W = I$. Define the matrices

$$W_0 = \frac{1}{4}\begin{pmatrix} 3 & 1 & 0 & \ldots & 0 \\ 1 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ 0 & \ldots & 0 & 1 & 3 \end{pmatrix}, \quad W_1 = \frac{\sqrt{2}}{4}\begin{pmatrix} -1 & 1 & 0 & \ldots & 0 \\ -1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 0 & 1 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix},$$

and

$$W_2 = \frac{1}{4}\begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}.$$

Then the operator W is defined as

$$W = \begin{pmatrix} W_0 \\ W_1 \\ W_2 \end{pmatrix}.$$

We are concerned with image restoration problems in two space-dimensions. Therefore, we construct the two-dimensional framelet analysis operator by means of the tensor products

$$W_{ij} = W_i \otimes W_j, \quad i, j = 0, 1, 2.$$

The matrix W_{00} is a low-pass filter; all the other matrices W_{ij} contain at least one high-pass filter. The analysis operator is given by

$$W = \begin{bmatrix} W_{00} \\ W_{01} \\ \vdots \\ W_{22} \end{bmatrix}.$$

We consider two types of noise, white Gaussian noise and impulse noise. The first is obtained when the entries of the vector e in the data vector b are realizations of a Gaussian random variable with 0 mean. In this case we refer to the ratio

$$\sigma = \frac{\|e\|_2}{\|Ax_{\text{true}}\|_2}$$

as the noise level. Impulse noise is obtained when the entries of the vector b are constructed as follows:

$$b_i = \begin{cases} (b_{\text{true}})_i & \text{with probability } 1-\sigma, \\ u_i & \text{with probability } \sigma, \end{cases}$$

where $0 \leqslant \sigma < 1$ and u_i is a number chosen randomly in the dynamical range of b_{true}. In this case we will refer to σ as the noise level.

The outer iterations of the algorithms considered are terminated as soon as the relative change of the computed approximate solution $x_+^{(k)}$ drops below a user-specified threshold, i.e., we terminate the iterations as soon as

$$\frac{\left\| x_+^{(k+1)} - x_+^{(k)} \right\|_2}{\left\| x_+^{(k)} \right\|_2} < \text{tol}_{\text{outer}},$$

(a) (b) (c)

Figure 1. Hubble test problem: (a) true image (230×230 pixels), (b) PSF (26×26 pixels), (c) blurred image with 2% of white Gaussian noise and 20% of impulse noise (230×230 pixels).

Table 1. Comparison of the RREs and SSIMs obtained with NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$ or NN-$(\ell_p$–$\ell_q)_{\text{FFT}}$ and with (unconstrained) ℓ_p–ℓ_q for different noise levels in the considered examples. For the Hubble example, the noise that corrupts the data is the sum of 2% of white Gaussian and $\sigma_1 = 20\%$, $\sigma_2 = 30\%$, and $\sigma_3 = 40\%$ of impulse noise. For the tomography example, the data is corrupted by white Gaussian noise of levels $\sigma_1 = 1\%$, $\sigma_2 = 5\%$, and $\sigma_3 = 10\%$.

Example	Quality measure	Method	Noise level		
			σ_1	σ_2	σ_3
Hubble	RRE	NN-$(\ell_p$–$\ell_q)_{\text{FFT}}$	0.136 35	0.150 83	0.267 87
		ℓ_p–ℓ_q	0.138 39	0.152 24	0.271 94
	SSIM	NN-$(\ell_p$–$\ell_q)_{\text{FFT}}$	0.846 85	0.809 82	0.417 53
		ℓ_p–ℓ_q	0.829 46	0.823 00	0.355 83
Tomography	RRE	NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$	0.194 91	0.261 88	0.321 57
		ℓ_p–ℓ_q	0.211 61	0.296 28	0.390 64
	SSIM	NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$	0.779 30	0.578 54	0.432 03
		ℓ_p–ℓ_q	0.557 56	0.327 40	0.205 30

or if the number of (outer) iterations reaches 200. The inner iterations in the modulus method are stopped as soon as the relative change of the computed approximate solution $x_+^{(k)}$ drops below a user-specified threshold

$$\frac{\left\| z_{j+1}^{(k)} - z_j^{(k)} \right\|_2}{\left\| z_j^{(k)} \right\|_2} < \text{tol}_{\text{inner}}, \tag{32}$$

or if the number of inner iterations reaches 100. In our experiments we set $\text{tol}_{\text{outer}} = \text{tol}_{\text{inner}} = 10^{-4}$.

In all the experiments, we set the dimension of the initial space to $\ell = 1$ and choose the initial approximate solution $x_+^{(0)} = \max\{A^{\text{T}}b, 0\}$. Consequently, $V_0 = A^{\text{T}}b/\|A^{\text{T}}b\|_2$.

To assess the quality of the reconstructed solution, we compute the relative reconstruction error (RRE) defined by

$$\text{RRE}(x) = \frac{\|x - x_{\text{true}}\|_2}{\|x_{\text{true}}\|_2}. \tag{33}$$

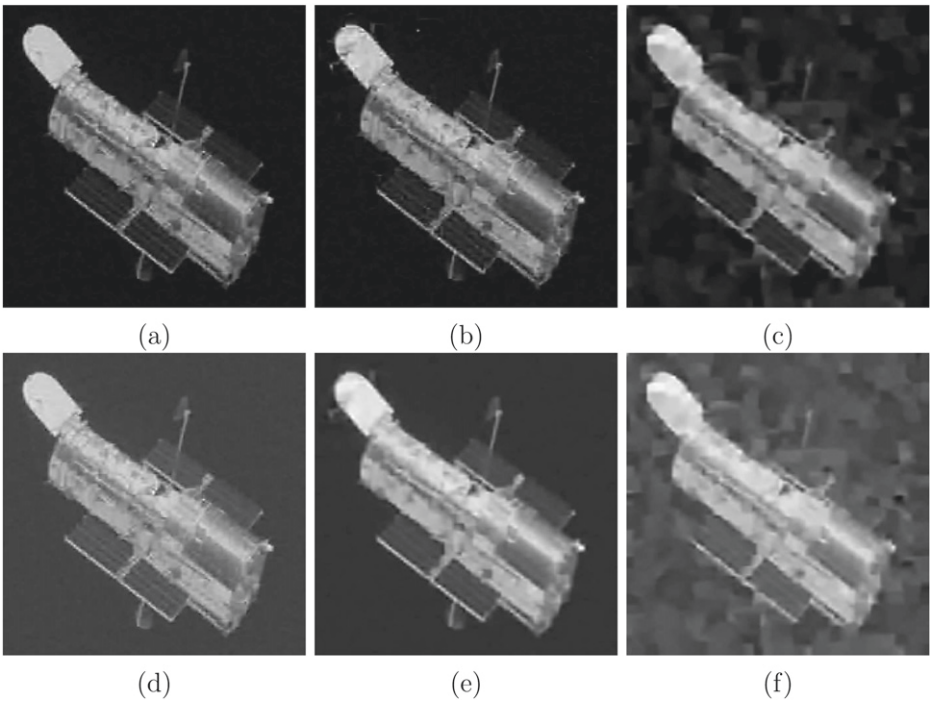


Figure 2. Hubble test problem reconstructions: panels (a)–(c) report the reconstructions obtained with NN-$(\ell_p$–$\ell_q)_{\text{FFT}}$ with 20%, 30%, and 40% of impulse noise, respectively; panels (d)–(f) report the reconstructions obtained with ℓ_p–ℓ_q with 20%, 30%, and 40% of impulse noise, respectively.

The parameter q is set to 0.1 in all the experiments, while p will depend on the noise. The regularization parameter μ is tuned by hand to minimize the RRE. A discussion on how to determine a good value for μ is outside the scope of this paper; see [17, 32] for discussions. The former reference uses the discrepancy principle and the latter cross validation to determine μ.

To measure the quality of the computed approximate solutions, we in addition to (33) use the structural similarity index (SSIM). The definition of the SSIM is involved and we refer to [33] for details. Here we just recall that the SSIM measures how well the overall structure of the image is recovered; the higher the index, the better the reconstruction. The highest value achievable is 1.

All computations were carried out in MATLAB R2018b with about 15 significant decimal digits running on a laptop computer with core CPU Intel® Core™ i7-8750H@2.20 GHz with 16 GB of RAM.

Hubble. In our first example we consider a synthetic astronomical image deblurring problem. We construct this example by blurring the image of the hubble telescope in figure 1(a) with the non-symmetric PSF in figure 1(b), and cut the boundary of the image to simulate boundary effects in real data. We then add 2% of white Gaussian noise and three different levels of impulse noise, namely 20%, 30%, and 40%. Figure 1(c) displays the blurred and noisy image with 20% impulse noise. Thanks to the nature of the image, we may impose periodic boundary conditions. This makes $A \in \mathbb{R}^{230^2 \times 230^2}$ a BCCB matrix, which allows us to use

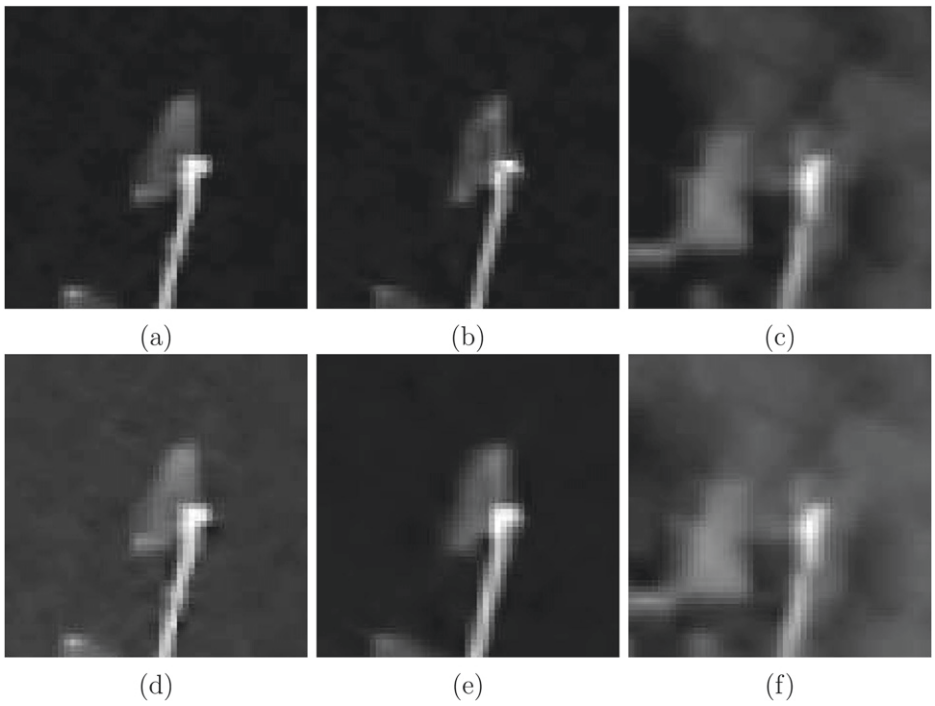


Figure 3. Hubble test problem reconstructions: panels (a)–(c) report a blow-up of the reconstructions obtained with NN-$(\ell_p-\ell_q)_{\text{FFT}}$ with 20%, 30%, and 40% of impulse noise, respectively; panels (d)–(f) report a blow-up of the reconstructions obtained with $\ell_p-\ell_q$ with 20%, 30%, and 40% of impulse noise, respectively.

algorithm 3 for the reconstruction; see, e.g., [27] for details on image deblurring. Since we added impulse noise, we set $p < 1$. Specifically, we let $p = 0.8$.

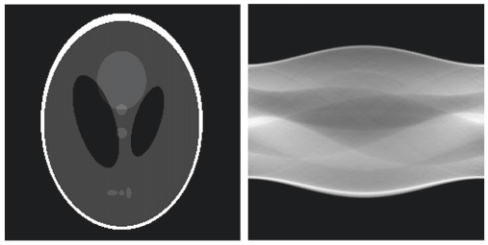
We report the errors and SSIMs obtained with algorithm 3 and the unconstrained method, implemented with the FFT as well, in table 1. We observe that the RRE obtained with the constrained method is always smaller than the RRE obtained with the unconstrained version. Moreover, the difference becomes larger as the noise increases. This is confirmed by both the SSIM (except for the second noise level) and by visual inspection of the reconstructions in figure 2. We observe that the presence of the nonnegativity constraint allows a more uniform reconstruction of the black background of the image. Moreover, the presence of the constraint allows us to select a smaller regularization parameter, thus, obtaining more detailed reconstructions. This is confirmed by visual inspection of the blow-ups of the reconstructions in figure 3.

Finally, in table 2, we report the CPU times in seconds for both the constrained and unconstrained methods. We can observe that, as expected, the computational cost of the constrained method is higher than of the unconstrained one. However, the total cost is not very high and it is possible to obtain the reconstructions in a reasonable amount of time.

Tomography. In our second example, we consider a synthetic tomography problem. In tomography, the data are the Radon transform of the attenuation coefficients of some scanned object; see, e.g., [34] for details on computerized tomography. We consider parallel beam tomography, where J parallel x-ray beams are shined through an object at different angles θ_k

Table 2. Comparison of the CPU times in seconds required for NN-$(\ell_p-\ell_q)_{\text{GKS}}$ or NN-$(\ell_p-\ell_q)_{\text{FFT}}$ and with (unconstrained) $\ell_p-\ell_q$ for different noise levels in the considered examples. For the Hubble example, the noise that corrupts the data is the sum of 2% of white Gaussian and $\sigma_1 = 20\%$, $\sigma_2 = 30\%$, and $\sigma_3 = 40\%$ of impulse noise. For the Tomography example, the data is corrupted by white Gaussian noise of levels $\sigma_1 = 1\%$, $\sigma_2 = 5\%$, and $\sigma_3 = 10\%$.

Example	Method	CPU time		
		σ_1	σ_2	σ_3
Hubble	NN-$(\ell_p-\ell_q)_{\text{FFT}}$	29.998	29.710	26.628
	$\ell_p-\ell_q$	13.850	13.990	14.978
Tomography	NN-$(\ell_p-\ell_q)_{\text{GKS}}$	69.423	47.724	98.215
	$\ell_p-\ell_q$	15.692	6.6424	5.5741



(a) (b)

Figure 4. Tomography test problem: (a) true image (256 × 256 pixels), (b) noise-free sinogram (362 × 90 pixels).

with $k = 1, 2, \ldots, K$. The datum $b_{j,k}$, the so-called sinogram, is the line integral of the attenuation coefficient of the object scanned along the jth beam at angle θ_k. We generate the synthetic data using the Matlab program package IR tools [35]. In particular, we use the command `PRtomo`. We set the dimension of the image to 256 × 256, and consider 90 angles equispaced between 0 and 179 degrees, and 362 beams. This leads to an underdetermined system where $A \in \mathbb{R}^{32\,580\times 65\,536}$. We report in figure 4(a) the exact attenuation coefficient, and in figure 4(b) the noise-free sinogram. We add different levels of white Gaussian noise, namely, 1%, 5%, and 10%. Since the noise is Gaussian, we set $p = 2$. The matrix of the system is not a BCCB matrix. Therefore, we use algorithm 2 for the solution of the constrained problem.

We report the results obtained with both the constrained and unconstrained approach in table 1. We can observe that the difference in the computed solutions determined by the constrained and unconstrained methods is more significant in this example than in the previous one. This can be motivated by the larger black area present in the image. Visual inspection of the reconstructions in figure 5 confirms the large difference between the reconstructions obtained with the unconstrained and the constrained approaches. In particular, we can observe that the reconstructions obtained with the unconstrained method appear affected by unwanted oscillations in the black areas. On the other hand, the constrained method is able to provide constant black areas around the phantom and does not reconstruct the noise, thus avoiding the unwanted oscillations present in the other reconstructions. Table 2 reports the CPU times required for the computation of the reconstructions. Like in the previous example, the timings

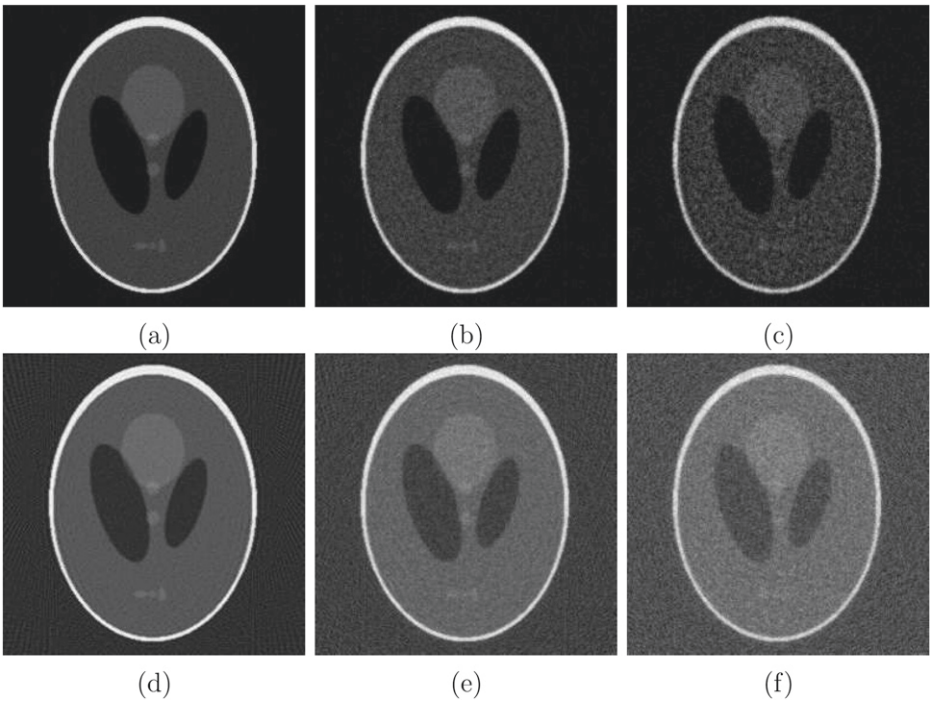


Figure 5. Tomography test problem reconstructions: panels (a)–(c) report the reconstructions obtained with NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$ with 1%, 5%, and 10% of white Gaussian noise, respectively; panels (d)–(f) display reconstructions obtained with (unconstrained) ℓ_p–ℓ_q with 1%, 5%, and 10% of white Gaussian noise, respectively.

for the constrained method are higher than the ones for the unconstrained one. However, they are not too high to make the method unfeasible.

Walnut. For our final example, we consider a real tomography problem. We use the data obtained by tomography of a walnut [36]. In particular, we consider the data in the `Data164.mat` file. The attenuation coefficients are stored in a 164×164 image and the sinogram is obtained by shining 164 fan-beams at 120 angles. This procedure generates an underdetermined problem where $A \in \mathbb{R}^{19\,680 \times 26\,896}$. Figure 6(b) shows the sinogram. In [36] a high-quality reconstruction obtained by a higher-dimensional data set of the true attenuation coefficients is provided; see figure 6(a). However, due to different scaling and size it is very difficult to use this image as ground truth to evaluate the RRE and SSIM. Thus, we compute the reconstructions obtained with the constrained and unconstrained method with five different regularization parameters. The advantages of the constrained method are already evident by visual inspection of the reconstruction and, therefore, we rely on this for the comparison of the two approaches. We do not know which kind of noise contaminated the data, however, it is safe to assume that the noise is not too far from Gaussian. We therefore let $p = 2$. Finally, since the matrix A is not a BCCB matrix, we use the NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$ method for the solution of the constrained problem.

The computed reconstructions are shown in figure 7. We can observe that, similarly to the synthetic example, the unconstrained model tends to amplify the noise and introduces unwanted oscillations in the reconstructed solution. On the other hand, these oscillations are not present in the reconstructions obtained with the constrained model. Moreover, we can

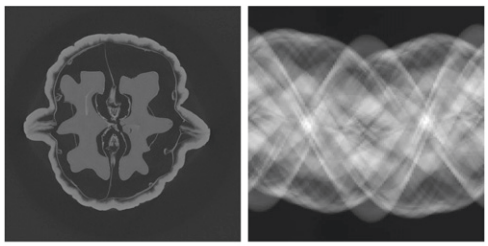


Figure 6. Walnut test problem: (a) high-quality reconstruction (2296 $\times$ 2296 pixels), (b) under-sampled sinogram (164 $\times$ 120 pixels).

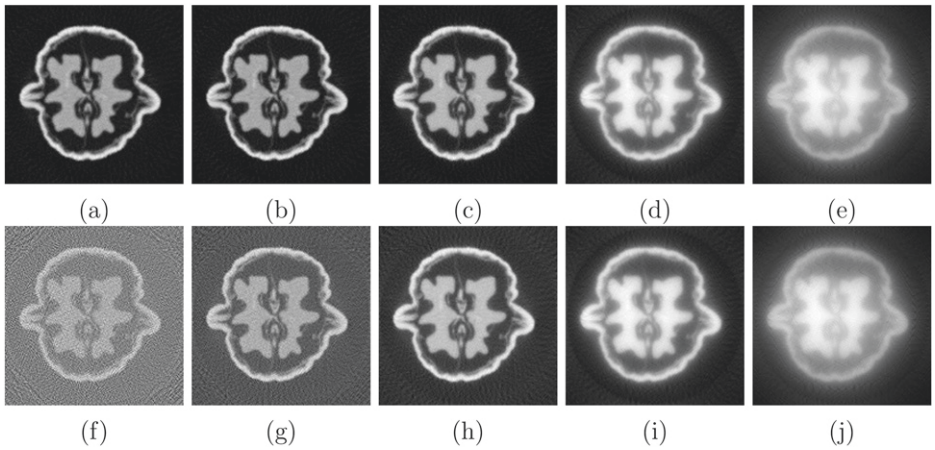


Figure 7. Walnut test problem reconstructions: panels (a)–(e) report the reconstructions obtained by the NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$ method with different regularization parameters μ, panels (f)–(j) show reconstructions determined by the (unconstrained) ℓ_p–ℓ_q method with different regularization parameters μ. The reconstructions in panels (a) and (f) are obtained with $\mu = 10^{-1}$, (b) and (g) with $\mu = 1$, (c) and (h) with $\mu = 10$, (d) and (i) with $\mu = 100$, and (e) and (j) with $\mu = 1000$.

Table 3. Comparison of the CPU times in seconds required for NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$ and with (unconstrained) ℓ_p–ℓ_q for different values of μ.

Method	Values of μ				
	10^{-1}	1	10	100	1000
NN-$(\ell_p$–$\ell_q)_{\text{GKS}}$	15.067	35.288	20.842	1.324	0.9686
ℓ_p–ℓ_q	20.933	6.0861	2.2145	0.9123	0.9109

observe that the reconstructions obtained with the constrained method are much more stable with respect to the choice of the parameter μ; the method is able to provide satisfactory reconstructions for a large interval of μ-values. Finally, in table 3 we show the CPU times required for the computation of all the reconstructions in figure 7. We can observe that the constrained method, while being more expensive than the unconstrained one, is able to maintain a reasonable computational cost.

8. Conclusions

In this paper we proposed new approaches for solving discrete ill-posed problems with non-negativity constraint. We started from the ℓ_p–ℓ_q regularization method described in [5] and combined it with the modulus-based algorithm [16, 23] to impose nonnegativity. The use of the non-convex models obtained when either p or q are smaller than 1 allowed us to determine high-quality reconstructions and to consider noise models different from the Gaussian one. We differentiated the cases in which the system matrix A is general and when it has a circulant structure. In the first case, we apply a generalized Krylov subspace method to lower the computational effort by projecting the problem into an appropriate subspace of fairly small dimension. In the second case, we exploited the fact that circulant and BCCB matrices can be efficiently diagonalized by the Fourier transform, thus, obtaining a diagonal problem. We provided a proof of convergence of approximate solutions computed with the new algorithms described. Several numerical examples, both on synthetic and real data, illustrated the performances of the proposed methods in terms of the quality of the reconstructed solutions.

Acknowledgments

The authors would like to thank a referee for carefully reading the manuscript and for comments that lead to improvements of the presentation. AB is a member of the GNCS-INdAM group that partially supported this work with the Young Researchers Project (Progetto Giovani Ricercatori) 'Variational methods for the approximation of sparse data'. Moreover, AB research is partially supported by the Regione Autonoma della Sardegna research project 'Algorithms and Models for Imaging Science [AMIS]' (RASSR57257, intervento finanziato con risorse FSC 2014–2020—Patto per lo Sviluppo della Regione Sardegna). The work of LR is supported in part by NSF grants DMS-1720259 and DMS-1729509.

ORCID iDs

A Buccini https://orcid.org/0000-0002-6456-4150
M Pasha https://orcid.org/0000-0003-4249-2421
L Reichel https://orcid.org/0000-0003-1729-6816

References

[1] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)
[2] Hansen P C 1994 Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems *Numer. Algorithms* **6** 1–35
[3] Hansen P C 1998 *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion* (Philadelphia, PA: SIAM)
[4] Estatico C, Gratton S, Lenti F and Titley-Peloquin D 2017 A conjugate gradient like method for p-norm minimization in functional spaces *Numer. Math.* **137** 895–922
[5] Huang G, Lanza A, Morigi S, Reichel L and Sgallari F 2017 Majorization-minimization generalized Krylov subspace methods for ℓ_p–ℓ_q optimization applied to image restoration *BIT Numer. Math.* **57** 351–78
[6] Lanza A, Morigi S, Reichel L and Sgallari F 2015 A generalized Krylov subspace method for ℓ_p-ℓ_q minimization *SIAM J. Sci. Comput.* **37** S30–50
[7] Cai J-F, Osher S and Shen Z 2009 Linearized Bregman iterations for frame-based image deblurring *SIAM J. Imaging Sci.* **2** 226–52

[8] Lanza A, Morigi S and Sgallari F 2016 Constrained TV_p-ℓ_2 model for image restoration *J. Sci. Comput.* **68** 64–91

[9] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.* **2** 183–202

[10] Xie Z and Hu J 2013 Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems *2013 6th Int. Congress on Image and Signal Processing* (*CISP*) vol 3 (IEEE) pp 1660–4

[11] Candès E J, Romberg J and Tao T 2006 Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information *IEEE Trans. Inf. Theory* **52** 489–509

[12] Donoho D L 2006 Compressed sensing *IEEE Trans. Inf. Theory* **52** 1289–306

[13] Liu Z, Wei Z and Sun W 2014 An iteratively approximated gradient projection algorithm for sparse signal reconstruction *Appl. Math. Comput.* **228** 454–62

[14] Tou J T and Gonzalez R C 1974 *Pattern Recognition Principles* (Boston, MA: Addison-Wesley)

[15] Lai M-J, Xu Y and Yin W 2013 Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization *SIAM J. Numer. Anal.* **51** 927–57

[16] Bai Z-Z 2010 Modulus-based matrix splitting iteration methods for linear complementarity problems *Numer. Linear Algebr. Appl.* **17** 917–33

[17] Buccini A and Reichel L 2019 An ℓ^2-ℓ^q regularization method for large discrete ill-posed problems *J. Sci. Comput.* **78** 1526–49

[18] Bredies K and Lorenz D A 2009 Regularization with non-convex separable constraints *Inverse Problems* **25** 085011

[19] Daubechies I, Defrise M and De Mol C 2004 An iterative thresholding algorithm for linear inverse problems with a sparsity constraint *Commun. Pure Appl. Math.* **57** 1413–57

[20] Grasmair M 2010 Non-convex sparse regularisation *J. Math. Anal. Appl.* **365** 19–28

[21] Hofmann B, Kaltenbacher B, Poeschl C and Scherzer O 2007 A convergence rates result for tikhonov regularization in banach spaces with non-smooth operators *Inverse Problems* **23** 987–1010

[22] Zheng N, Hayami K and Yin J-F 2016 Modulus-type inner outer iteration methods for nonnegative constrained least squares problems *SIAM J. Matrix Anal. Appl.* **37** 1250–78

[23] Bai Z-Z, Buccini A, Hayami K, Reichel L, Yin J-F and Zheng N 2017 Modulus-based iterative methods for constrained Tikhonov regularization *J. Comput. Appl. Math.* **319** 1–13

[24] Cottle R, Pang J-S and Venkateswaran V 1989 Sufficient matrices and the linear complementarity problem *Linear Algebr. Appl.* **114** 231–49

[25] Daniel J W, Gragg W B, Kaufman L and Stewart G W 1976 Reorthogonalization and stable algorithms for updating the Gram-Schmidt QR factorization *Math. Comput.* **30** 772–95

[26] Baglama J and Reichel L 2005 Augmented implicitly restarted Lanczos bidiagonalization methods *SIAM J. Sci. Comput.* **27** 19–42

[27] Hansen P C, Nagy J G and O'Leary D P 2006 *Deblurring Images: Matrices, Spectra, and Filtering* (Philadelphia, PA: SIAM)

[28] Mairal J 2015 Incremental majorization-minimization optimization with application to large-scale machine learning *SIAM J. Optim.* **25** 829–55

[29] Cai J-F, Osher S and Shen Z 2009 Split Bregman methods and frame based image restoration *Multiscale Model. Simul.* **8** 337–69

[30] Huang J, Donatelli M and Chan R H 2013 Nonstationary iterated thresholding algorithms for images deblurring *Inverse Problems Imaging* **7** 717–36

[31] Cai Y, Donatelli M, Bianchi D and Huang T-Z 2016 Regularization preconditioners for frame-based image deblurring with reduced boundary artifacts *SIAM J. Sci. Comput.* **38** B164–89

[32] Buccini A and Reichel L 2020 An ℓ^p-ℓ^q minimization method with cross-validation for the restoration of impulse noise contaminated images *J. Comput. Appl. Math.* **375** 112824

[33] Wang Z, Bovik A C, Sheikh H R and Simoncelli E P 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12

[34] Buzug T M 2011 Computed tomography *Springer Handbook of Medical Technology* eds R Kramme, K-P Hoffmann and R S Pozos (Berlin: Springer) pp 311–42

[35] Gazzola S, Hansen P C and Nagy J G 2019 IR tools: a MATLAB package of iterative regularization methods and large-scale test problems *Numer. Algorithms* **81** 773–811

[36] Hämäläinen K, Harhanen L, Kallonen A, Kujanpää A, Niemi E and Siltanen S 2015 Tomographic x-ray data of a walnut (arXiv:1502.04064)