A nested primal–dual FISTA-like scheme for composite convex optimization problems

S. Bonettini[1] · **M. Prato**[1] · **S. Rebegoldi**[2]

Abstract

We propose a nested primal–dual algorithm with extrapolation on the primal variable suited for minimizing the sum of two convex functions, one of which is continuously differentiable. The proposed algorithm can be interpreted as an inexact inertial forward–backward algorithm equipped with a prefixed number of inner primal–dual iterations for the proximal evaluation and a "warm–start" strategy for starting the inner loop, and generalizes several nested primal–dual algorithms already available in the literature. By appropriately choosing the inertial parameters, we prove the convergence of the iterates to a saddle point of the problem, and provide an $O(1/n)$ convergence rate on the primal–dual gap evaluated at the corresponding ergodic sequences. Numerical experiments on some image restoration problems show that the combination of the "warm–start" strategy with an appropriate choice of the inertial parameters is strictly required in order to guarantee the convergence to the real minimum point of the objective function.

Keywords Primal–dual algorithms · Forward–backward algorithms · Inertial techniques · Convex optimization · Image deblurring

1 Introduction

In this paper, we address the following composite convex optimization problem

---

✉ S. Rebegoldi
simone.rebegoldi@unifi.it

S. Bonettini
silvia.bonettini@unimore.it

M. Prato
marco.prato@unimore.it

1 Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, Via Campi 213, 41125 Modena, Italy

2 Dipartimento di Ingegneria Industriale, Università di Firenze, Viale Morgagni 40, 50143 Firenze, Italy

$$\min_{u \in \mathbb{R}^d} F(u) \equiv f(u) + g(Au), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, $A \in \mathbb{R}^{d' \times d}$ and $g : \mathbb{R}^{d'} \to \mathbb{R} \cup \{\infty\}$ is convex. Problems of the form (1) typically arise from variational models in image processing, computer vision and machine learning, such as the classical Total Variation (TV) [38], Total Generalized Variation (TGV) [11] or Directional TGV models [22, 25, 26] for image denoising, TV-ℓ_1 models for optical flow estimation [15], regularized logistic regression in multinomial classification problems [10] and several others.

One standard approach to tackle problem (1) is the forward–backward (FB) method [5, 8, 19, 20], which combines iteratively a gradient step on the differentiable part f with a proximal step on the convex (possibly non differentiable) term $g \circ A$, thus generating an iterative sequence $\{u_n\}_{n \in \mathbb{N}}$ of the form

$$u_{n+1} = \text{prox}_{\alpha g \circ A}(u_n - \alpha \nabla f(u_n)), \quad n = 0, 1, \dots \tag{2}$$

where $\alpha > 0$ is an appropriate steplength along the descent direction $-\nabla f(u_n)$ and the proximal operator $\text{prox}_{\alpha g \circ A}$ is defined by

$$\text{prox}_{\alpha g \circ A}(a) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \; \alpha g(Au) + \frac{1}{2}\|u - a\|^2. \tag{3}$$

Convergence of the iterates (2) to a solution of problem (1) can be obtained by assuming that ∇f is Lipschitz continuous and choosing $\alpha \in (0, 2/L)$, where L is the Lipschitz constant of the gradient [20]. Accelerated versions of the standard FB scheme can be obtained by introducing an inertial step inside (2), thus obtaining

$$\begin{cases} \bar{u}_n = u_n + \gamma_n(u_n - u_{n-1}) \\ u_{n+1} = \text{prox}_{\alpha g \circ A}(\bar{u}_n - \alpha \nabla f(\bar{u}_n)) \end{cases}, \quad n = 0, 1, \dots \tag{4}$$

where $\bar{u}_n$ is usually called the inertial point and $\gamma_n \in [0, 1]$ is the inertial parameter. Such an approach was first considered in a pioneering work by Nesterov [32] for differentiable problems (i.e assuming $g \equiv 0$), and then later generalized in [5] to the more general problem (1) under the name of FISTA - "Fast Iterative Soft-Thresholding Algorithm". FISTA shares an optimal $O(1/n^2)$ convergence rate [2, 5], which is a lower bound for first-order methods [33], and convergence of the iterates still holds under appropriate assumptions on the inertial parameter [14].

The original versions of methods (2) and (4) require that the proximal operator of $g \circ A$ is computed exactly. However, such an assumption is hardly realistic, as the proximal operator might not admit a closed-form formula or its computation might be too costly in several practical situations, for instance in TV regularized imaging problems or low-rank minimization (see [3, 30, 35, 39] and references therein). In these cases, assuming that prox_g and the matrix A are explicitly available, the proximal operator can be approximated at each outer iteration n by means of an inner iterative algorithm, which is applied to the dual problem associated to (3), see [9, 13, 17, 39]. The choice of an appropriate number of inner iterations for the evaluation of the proximal operator is thus a crucial issue, as it affects the numerical performance

of FB algorithms as well as their theoretical convergence. In this regard, we can identify two main strategies in the literature for addressing the inexact computation of the FB iterate. On the one hand, several works have devised inexact versions of the FB algorithm by requiring that the FB iterate is computed with sufficient accuracy up to a prefixed (or adaptive) tolerance [8, 9, 20, 39, 41]. In these works, the inner iterations are usually run until some *stopping condition* is met, in order to ensure that the inexact FB iterate is sufficiently close to the exact proximal point. Since the tolerances are required to be vanishing, the number of inner iterations will grow unbounded as the outer iterations increase, resulting in an increasing computational cost per proximal evaluation and hence an overall slowdown of the algorithm. On the other hand, one could fix the number of inner iterations in advance, thus losing control on the accuracy of the proximal evaluation, while employing an appropriate *starting condition* for the inner numerical routine. In particular, the authors in [17] consider a nested primal–dual algorithm that can be interpreted as an inexact version of the standard FB scheme (2), where the proximal operator is approximated by means of a prefixed number of primal–dual iterates, and each inner primal–dual loop is "warm-started" with the outcome of the previous one. As proved in [17, Theorem 3.1], such a "warm-start" strategy is sufficient for proving the convergence of the iterates to a solution of (1) when the accuracy in the proximal evaluation is prefixed.

In this paper, we propose an accelerated variant of the nested primal–dual algorithm devised in [17] that performs an inertial step on the primal variable. The proposed algorithm could be interpreted as an inexact version of the FISTA-like method (4) equipped with a prefixed number of inner primal–dual iterations for the proximal evaluation and a "warm–start" strategy for the initialization of the inner loop. Under an appropriate choice of the inertial parameters, we prove the convergence of the iterates towards a solution of (1) and an ergodic $O(1/n)$ convergence rate result on the primal–dual gap. We note that other nested primal–dual algorithms with inertial steps on the primal variable have been studied in the previous literature, however they are either shown to diverge in numerical counterexamples [4, 18] or its convergence relies on additional over relaxation steps [28] in the same fashion as in the popular "Chambolle-Pock" primal–dual algorithm [15]. We claim that our proposed method is the first inexact inertial FB method whose convergence relies solely on its "warm–start" procedure, without requiring any increasing accuracy in the proximal evaluation. Furthermore, our numerical experiments on a Poisson image deblurring test problem show that the theoretical assumption made on the inertial parameters is strictly necessary for proving convergence, as the proposed nested inertial primal–dual algorithm equipped with the same inertial parameter rule of FISTA (which does not comply with the theoretical assumption) is observed to diverge from the solution of the minimization problem.

The paper is organized as follows. In Sect. 2 we recall some useful notions of convex analysis, the problem setting, and a practical procedure for approximating the proximal operator by means of a primal–dual loop. In Sect. 3 we present the algorithm and the related convergence analysis. Section 4 is devoted to numerical experiments on some image restoration test problems. Finally, the conclusions and future perspectives are reported in Sect. 5.

2 Preliminaries

In the remainder of the paper, the symbol $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^d$. If $v \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d' \times d}$, $\|v\|$ is the usual Euclidean vectorial norm, and $\|A\|$ represents the largest singular value of A. If $v = (v_1, \ldots, v_d) \in \mathbb{R}^{dn}$, where $v_i \in \mathbb{R}^n$, $i = 1, \ldots, d$, then $\|v\|_{2,1} = \sum_{i=1}^{d} \|v_i\|$. If $\Omega \subseteq \mathbb{R}^d$, then $\text{relint}(\Omega) = \{x \in \Omega : \exists \epsilon > 0 \text{ s.t. } B(x, \epsilon) \cap \text{aff}(\Omega) \subseteq \Omega\}$ denotes the relative interior of the set Ω, where $B(x, \epsilon)$ is the ball of center x and radius ϵ, and $\text{aff}(\Omega)$ is the affine hull of Ω. Finally, the symbol $\mathbb{R}_{\geq 0}$ stands for the set of nonnegative real numbers.

2.1 Basic notions of convex analysis

We start by recalling some well-known definitions and properties of convex analysis that will be employed throughout the paper.

Definition 1 [37, p. 214] Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper, convex and lower semicontinuous function. The subdifferential of φ at point $u \in \mathbb{R}^d$ is defined as the following set

$$\partial \varphi(u) = \{w \in \mathbb{R}^d : \varphi(v) \geq \varphi(u) + \langle w, v - u \rangle \ \forall \ v \in \mathbb{R}^d\}.$$

From Definition 1, it easily follows that $\hat{u}$ is a minimum point of φ if and only $0 \in \partial \varphi(\hat{u})$.

Lemma 1 [37, *Theorem* 23.8, *Theorem* 23.9] *Let* $\varphi_1, \varphi_2 : \mathbb{R}^{d'} \to \mathbb{R} \cup \{\infty\}$ *be proper, convex and lower semicontinuous functions and* $A \in \mathbb{R}^{d' \times d}$.

(i) *If* $\varphi(u) = \varphi_1(u) + \varphi_2(u)$ *and there exists* $u_0 \in \mathbb{R}^{d'}$ *such that* $u_0 \in \text{relint}(\text{dom}(\varphi_1)) \cap \text{relint}(\text{dom}(\varphi_2))$, *where* relint *denotes the relative interior of a set, then*

$$\partial(\varphi_1 + \varphi_2)(u) = \partial\varphi_1(u) + \partial\varphi_2(u), \quad \forall \ u \in \text{dom}(\varphi).$$

(ii) *If* $\varphi(u) = \varphi_1(Au)$ *and there exists* $u_0 \in \mathbb{R}^d$ *such that* $Au_0 \in \text{relint}(\text{dom}(\varphi_1))$, *then*

$$\partial\varphi(u) = A^T \partial\varphi_1(Au) = \{A^T w : w \in \partial\varphi_1(Au)\}, \quad \forall \ u \in \text{dom}(\varphi).$$

Definition 2 [37, p. 104] Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper, convex and lower semicontinuous function. The Fenchel conjugate of φ is the function $\varphi^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ defined as

$$\varphi^*(w) = \sup_{u \in \mathbb{R}^d} \langle w, u \rangle - \varphi(u).$$

Lemma 2 [37, *Theorem* 12.2] *Let* $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *be a proper, convex and lower semicontinuous function. Then* φ^* *is itself convex and lower semicontinuous and* $(\varphi^*)^* = \varphi$, *which implies the following variational representation for* φ:

$$\varphi(u) = \sup_{w \in \mathbb{R}^d} \langle u, w \rangle - \varphi^*(w).$$

Definition 3 [31, p. 278] Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper, convex and lower semicontinuous function. The proximal operator of φ is defined as

$$\text{prox}_\varphi(a) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \, \frac{1}{2} \|u - a\|^2 + \varphi(u).$$

The following two lemmas involving the proximal operator will be extensively employed in the convergence analysis of Sect. 3.

Lemma 3 *Let* $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *be proper, convex and lower semicontinuous. For all* $\alpha, \beta > 0$, *the following statements are equivalent*:

(i)$u = \text{prox}_{\alpha\varphi}(u + \alpha w)$;
(ii)$w \in \partial\varphi(u)$;
(iii)$\varphi(u) + \varphi^*(w) = \langle w, u \rangle$;
(iv)$u \in \partial\varphi^*(w)$;
(v)$w = \text{prox}_{\beta\varphi^*}(\beta u + w)$.

Proof See [31, Sections 11–12]. □

Lemma 4 *Let* $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *be proper, convex and lower semicontinuous, and* $x, e \in \mathbb{R}^d$. *Then the equality* $y = \text{prox}_\varphi(x + e)$ *is equivalent to the following inequality*:

$$\|y - z\|^2 \leq \|x - z\|^2 - \|x - y\|^2 + 2\langle y - z, e \rangle + 2\varphi(z) - 2\varphi(y), \quad \forall \, z \in \mathbb{R}^d.$$

Proof See [17, Lemma 3.3]. □

Finally, we recall the well-known descent lemma holding for functions having a Lipschitz continuous gradient.

Lemma 5 *Suppose that* $\varphi : \mathbb{R}^d \to \mathbb{R}$ *is a continuously differentiable function with* $L-$*Lipschitz continuous gradient, i.e.*

$$\|\nabla\varphi(u) - \nabla\varphi(v)\| \leq L\|u - v\|, \quad \forall \, u, v \in \mathbb{R}^d.$$

Then the following inequality holds

$$\varphi(u) \leq \varphi(v) + \langle \nabla\varphi(v), u - v \rangle + \frac{L}{2}\|u - v\|^2, \quad \forall \, u, v \in \mathbb{R}^d.$$

Proof See [7, Proposition A.24]. □

2.2 Problem setting

In the remainder, we consider the optimization problem

$$\min_{u\in\mathbb{R}^d} F(u) \equiv f(u) + g(Au),$$ (5)

under the following blanket assumptions on the involved functions.

Assumption 1

(i) $f : \mathbb{R}^d \to \mathbb{R}$ is convex and continuously differentiable and ∇f is L–Lipschitz continuous;

(ii) $g : \mathbb{R}^{d'} \to \mathbb{R} \cup \{\infty\}$ is proper, lower semicontinuous and convex;

(iii) $A \in \mathbb{R}^{d'\times d}$, and there exists $u_0 \in \mathbb{R}^d$ such that $Au_0 \in \text{relint}(\text{dom}(g))$;

(iv) problem (5) admits at least one solution $\hat{u} \in \mathbb{R}^d$.

Note that the hypothesis on A is needed only to ensure that the subdifferential rule in Lemma 1 (ii) holds, so that we can write $\partial(g\circ A)(u) = A^T\partial g(Au)$. Such rule allows us to interpret the minimum points of (5) as solutions of appropriate variational equations, as stated below.

Lemma 6 *A point $\hat{u} \in \mathbb{R}^d$ is a solution of problem (5) if and only if the following conditions hold*

$$\nabla f(\hat{u}) + A^T\hat{v} = 0, \quad \text{with } \hat{v} = \text{prox}_{\beta\alpha^{-1}g^*}(\hat{v} + \beta\alpha^{-1}A\hat{u}).$$ (6)

for any $\alpha, \beta > 0$.

Proof Observe that $\hat{u}$ is a solution of problem (5) if and only if $0 \in \partial F(\hat{u})$. According to Lemma 1 this is equivalent to writing

$$\nabla f(\hat{u}) + A^T\hat{v} = 0, \quad \text{with } \hat{v} \in \partial g(A\hat{u}).$$

An application of Lemma 3 shows that the inclusion $\hat{v} \in \partial g(A\hat{u})$ is equivalent to the equality $\hat{v} = \text{prox}_{\beta\alpha^{-1}g^*}(\hat{v} + \beta\alpha^{-1}A\hat{u})$, which gives the thesis. □

Using Lemma 2, the convex problem (5) can be equivalently reformulated as the following convex-concave saddle-point problem [37]

$$\min_{u\in\mathbb{R}^d} \max_{v\in\mathbb{R}^{d'}} \mathcal{L}(u,v) \equiv f(u) + \langle Au, v\rangle - g^*(v)$$ (7)

where $\mathcal{L}(u,v)$ denotes the primal–dual function. Under Assumptions 1, we can switch the minimum and maximum operator [37, Corollary 31.2.1] and use again Lemma 2, thus obtaining

$$\max_{v \in \mathbb{R}^{d'}} \; -f^*(-A^T v) - g^*(v). \tag{8}$$

Problems (5)–(7)–(8) are usually referred to as the primal, primal–dual, and dual problem, respectively.

2.3 Inexact proximal evaluation via primal–dual iterates

A function of the form $g \circ A$ does not necessarily admit a closed-form formula for its proximal operator, as it is the case of several sparsity-inducing priors appearing in image and signal processing applications. However, when a closed-form formula is not accessible, the proximal operator can be still approximately evaluated by means of a primal–dual scheme, as we show in the following.

Theorem 1 *Let f, g, and A be defined as in problem* (5) *under Assumptions* 1. *Given* $z \in \mathbb{R}^d$, $\alpha > 0$, $0 < \beta < 2/\|A\|^2$ *and* $v^0 \in \mathbb{R}^{d'}$, *consider the dual sequence* $\{v^k\}_{k \in \mathbb{N}} \subseteq \mathbb{R}^{d'}$ *defined as follows*

$$v^{k+1} = \text{prox}_{\beta \alpha^{-1} g^*}(v^k + \beta \alpha^{-1} A(z - \alpha A^T v^k)), \quad \forall \, k \in \mathbb{N}. \tag{9}$$

Then there exists $\hat{v} \in \mathbb{R}^{d'}$ *such that*

(i) $\lim_{k \to \infty} v^k = \hat{v}$;
(ii) $\text{prox}_{\alpha g \circ A}(z) = z - \alpha A^T \hat{v}$.

Proof Theorem 1 is a special case of [17, Lemma 2.3], where the authors prove the same result for objective functions of the form $F(u) = f(u) + g(Au) + h(u)$ being h a lower semicontinuous convex function. The proof is based on the differential inclusion characterizing the proximal point $\hat{z} = \text{prox}_{\alpha g \circ A}(z)$, that is

$$\hat{z} - z + \alpha A^T \hat{v} = 0, \quad \text{with } \hat{v} \in \partial g(A\hat{z}). \tag{10}$$

Using Lemma 3, the above variational equation can be equivalently written as

$$\hat{z} = z - \alpha A^T \hat{v}, \quad \text{with } \hat{v} = \text{prox}_{\beta \alpha^{-1} g^*}(\hat{v} + \beta \alpha^{-1} A\hat{z}) \tag{11}$$

where $\alpha, \beta > 0$ and g^* is the Fenchel conjugate of g. By replacing the first equation of (11) into the second one, we obtain

$$\hat{v} = \text{prox}_{\beta \alpha^{-1} g^*}(\hat{v} + \beta \alpha^{-1} A(z - \alpha A^T \hat{v})), \tag{12}$$

that is, $\hat{v}$ is a fixed point of the operator $T(v) = \text{prox}_{\beta \alpha^{-1} g^*}(v + \beta \alpha^{-1} A(z - \alpha A^T v))$. Consequently, the dual sequence $\{v^k\}_{k \in \mathbb{N}}$ defined in (9) can be interpreted as the fixed-point iteration applied to the operator T. From this remark, the thesis easily follows (see [17, Lemma 2.3] for the complete proof). $\qquad \square$

Remark 1 By combining items (i)-(ii) of Theorem 1, we conclude that

$$\text{prox}_{\alpha g \circ A}(z) = \lim_{k \to \infty} z - \alpha A^T \text{prox}_{\beta \alpha^{-1} g^*}(v^k + \beta \alpha^{-1} A(z - \alpha A^T v^k)).$$

Hence, the proximal operator of $g \circ A$ can be approximated by a finite number of steps of a primal–dual procedure, provided that the operators prox_{g^*} and A are easily computable.

3 Algorithm and convergence analysis

In this section, we propose an accelerated version of the nested primal–dual algorithm considered in [17], which features an extrapolation step in the same fashion as in the popular FISTA and other Nesterov-type forward–backward algorithms [2, 5, 9]. The resulting scheme can be itself considered as an inexact inertial forward–backward algorithm in which the backward step is approximately computed by means of a prefixed number of primal–dual steps.

We report the proposed scheme in Algorithm 1. Given the initial guess $u_0 \in \mathbb{R}^d$, the steplengths $0 < \alpha < 1/L$, $0 < \beta < 1/\|A\|^2$, and a prefixed positive integer $k_{\max}$, we first compute the extrapolated point $\bar{u}_n$ (STEP 1) and initialize the primal–dual loop with the outcome of the previous inner loop, i.e. $v_n^0 = v_{n-1}^{k_{\max}}$ (STEP 2). This is the so-called "warm-start" strategy borrowed from [17]. Subsequently we perform $k_{\max}$ inner primal–dual steps (STEP 3), compute the additional primal iterate $u_n^{k_{\max}}$ (STEP 4), and average the primal iterates $\{u_n^k\}_{k=1}^{k_{\max}}$ over the number of inner iterations (STEP 5). Note that STEPS 3–4 represent an iterative procedure for computing an approximation of the proximal–gradient point $\hat{u}_{n+1} = \text{prox}_{\alpha g \circ A}(\bar{u}_n - \alpha \nabla f(\bar{u}_n))$ (see Theorem 1 and Remark 1). In that spirit, we can consider Algorithm 1 as an inexact inertial forward–backward scheme where the inexact computation of the proximal operator is made explicit in the description of the algorithm itself.

Algorithm 1 Nested primal–dual FISTA-like algorithm

Choose $u_{-1} \in \mathbb{R}^d$, $u_0 = u_{-1}$, $v_{-1}^{k_{\max}} \in \mathbb{R}^{d'}$, $0 < \alpha < \frac{1}{L}$, $0 < \beta < \frac{1}{\|A\|^2}$, $k_{\max} \in \mathbb{N}$, $\{\gamma_n\}_{n \in \mathbb{N}} \subseteq \mathbb{R}_{\geq 0}$.

FOR $n = 0, 1, 2, \ldots$

  STEP 1. Compute the extrapolated point

$$\bar{u}_n = u_n + \gamma_n(u_n - u_{n-1}). \tag{13}$$

  STEP 2. Set $v_n^0 = v_{n-1}^{k_{\max}}$.
  STEP 3. Compute $k_{\max}$ primal-dual iterates:
        FOR $k = 0, 1, \ldots, k_{\max} - 1$

$$u_n^k = \bar{u}_n - \alpha \nabla f(\bar{u}_n) - \alpha A^T v_n^k \tag{14}$$

$$v_n^{k+1} = \operatorname{prox}_{\beta \alpha^{-1} g^*}(v_n^k + \beta \alpha^{-1} A u_n^k). \tag{15}$$

  STEP 4. Compute $u_n^{k_{\max}} = \bar{u}_n - \alpha \nabla f(\bar{u}_n) - \alpha A^T v_n^{k_{\max}}$.
  STEP 5. Compute the next iterate as

$$u_{n+1} = \frac{1}{k_{\max}} \sum_{k=1}^{k_{\max}} u_n^k. \tag{16}$$

END

Algorithm 1 includes some nested primal–dual schemes previously proposed in the literature as special cases:

- when $\gamma_n \equiv 0$, $k_{\max} = 1$, and $f(u) = \|Hu - y\|^2/2$, Algorithm 1 becomes the Generalized Iterative Soft-Thresholding Algorithm (GISTA), which was first devised in [29] for regularized least squares problems;
- when $\gamma_n \equiv 0$ and $k_{\max} = 1$, we recover the Proximal Alternating Predictor-Corrector (PAPC) algorithm proposed in [23], which can be seen as an extension of GISTA to a general data-fidelity term $f(u)$;
- when $\gamma_n \equiv 0$, Algorithm 1 reduces to the nested (non inertial) primal–dual scheme proposed in [17], which generalizes GISTA and PAPC by introducing multiple inner–primal dual iterations.

In the case $\gamma_n \neq 0$, Algorithm 1 can be interpreted as an inexact version of FISTA, which differs from other inexact FISTA-like schemes existing in the literature (see e.g. [9, 43]) due to the warm-start strategy, the prefixed number of primal–dual iterates and the averaging of the inner primal sequence at the end of each outer iteration. Note that similar nested inertial primal–dual algorithms were proposed in [4,

18] for solving least squares problems with total variation regularization. However, both these extensions do not adopt a "warm-start" strategy for the inner loop: indeed the authors in [4] set $v_n^0 = 0$, whereas in [18] the initial guess is computed by extrapolating the two previous dual variables. Interestingly, in the numerical experiences of [4, 18], the function values of the methods were observed to diverge or converge to a higher limit value than their non-extrapolated counterpart. These numerical "failures" might be due to the lack of the warm-start strategy, which appears to be crucial also for the convergence of inertial schemes, as we will demonstrate in the next section.

We also note that Algorithm 1 differs significantly from the classical Chambolle-Pock primal–dual algorithm for minimizing sums of convex functions of the form $F(u) = f(u) + g(Au)$ [15], and its generalizations to the case where f is differentiable [16, 21, 35, 44]. Indeed, while Algorithm 1 performs the extrapolation step on the primal variable $\bar{u}_n$ *before* the inner primal–dual loop begins, the algorithms proposed in [15, 16, 21, 35, 44] include extrapolation as an intermediate step between the primal and the dual iteration (or viceversa). For instance, the Chambolle-Pock algorithm described in [15, Algorithm 1] can be written as

$$\begin{cases} u_{n+1} = \text{prox}_{\tau f}(u_n - \tau A^T v_n) \\ \bar{u}_{n+1} = 2u_{n+1} - u_n \\ v_{n+1} = \text{prox}_{\sigma g^*}(v_n + \sigma A\bar{u}_{n+1}), \end{cases} \tag{17}$$

where $\bar{u}_{n+1}$ represents the extrapolated iterate, which is computed after the primal iterate u_{n+1} and then employed inside the dual iterate v_{n+1}. All the algorithms in [16, 21, 35, 44] are built upon similar extrapolation strategies. Furthermore, we remark that the works [21, 35, 44] also take into account the possible inexact computation of the involved proximal operators, either by introducing additive errors [21, 44] or considering ϵ_n–approximations of the proximal points [35], as well as errors into the computation of the gradient of f. By contrast, we put ourselves in the easier scenario where the proximal operators and gradient are all computable in closed form.

3.1 Convergence analysis

The aim of this section is to prove the convergence of the iterates generated by Algorithm 1, as well as an $O(1/n)$ convergence rate on the primal–dual gap evaluated at the average (ergodic) sequences associated to $\{u_n\}_{n\in\mathbb{N}}$ and $\{v_n^0\}_{n\in\mathbb{N}}$. Similar results in the (non strongly) convex setting have been proved for several other primal–dual methods, see e.g. [15–17, 23, 35].

We start by proving some technical descent inequalities holding for the primal–dual function (7), where an approximate forward–backward point of the form $u = \bar{u} - \alpha\nabla f(\bar{u}) - \alpha A^T v$ and its associated dual variable v appear. The result is similar to the one given in [23, Lemma 3.1] for the iterates of the PAPC algorithm, which in turn is a generalization of a technical result employed for GISTA [29, Lemma 3]. Such inequalities also recall the key inequality of FISTA [14, Lemma 1],

which however involves the primal function values instead of the primal–dual ones. For our purposes, we need to introduce the following function

$$\|v\|_A^2 := \|v\|^2 - \beta\|A^T v\|^2, \quad v \in \mathbb{R}^{d'},$$

which is a norm whenever $0 < \beta < 1/\|A\|^2$, as is the case for Algorithm 1.

Lemma 7 *Let f, g, A be defined as in problem* (5) *under Assumptions* 1. *Given $\bar{u} \in \mathbb{R}^d$, $\bar{v} \in \mathbb{R}^{d'}$, $0 < \alpha < \frac{1}{L}$, $0 < \beta < \frac{1}{\|A\|^2}$, define the points $\tilde{u} \in \mathbb{R}^d$, $v \in \mathbb{R}^{d'}$, $u \in \mathbb{R}^d$ as follows*

$$\begin{cases} \tilde{u} = \bar{u} - \alpha\nabla f(\bar{u}) - \alpha A^T \bar{v} \\ v = \text{prox}_{\beta\alpha^{-1}g^*}(\bar{v} + \beta\alpha^{-1}A\tilde{u}) \\ u = \bar{u} - \alpha\nabla f(\bar{u}) - \alpha A^T v \end{cases} \tag{18}$$

The following facts hold.

(i) *For all $z \in \mathbb{R}^d$, we have*

$$\mathcal{L}(u, v) + \frac{1}{2\alpha}\|z - u\|^2 \le \mathcal{L}(z, v) + \frac{1}{2\alpha}\|z - \bar{u}\|^2 - \frac{1}{2}\left(\frac{1}{\alpha} - L\right)\|u - \bar{u}\|^2. \tag{19}$$

(ii) *For all $z' \in \mathbb{R}^{d'}$, we have*

$$\mathcal{L}(u, z') + \frac{\alpha}{2\beta}\|z' - v\|_A^2 \le \mathcal{L}(u, v) + \frac{\alpha}{2\beta}\|z' - \bar{v}\|_A^2 - \frac{\alpha}{2\beta}\|v - \bar{v}\|_A^2. \tag{20}$$

(iii) *For all $z \in \mathbb{R}^d$, $z' \in \mathbb{R}^{d'}$, we have*

$$\mathcal{L}(u, z') + \frac{1}{2\alpha\beta}\left(\beta\|z - u\|^2 + \alpha^2\|z' - v\|_A^2\right)$$
$$\le \mathcal{L}(z, v) + \frac{1}{2\alpha\beta}\left(\beta\|z - \bar{u}\|^2 - \beta(1 - \alpha L)\|u - \bar{u}\|^2\right) \tag{21}$$
$$+ \frac{1}{2\alpha\beta}\left(\alpha^2\|z' - \bar{v}\|_A^2 - \alpha^2\|v - \bar{v}\|_A^2\right).$$

Proof (i) We first apply the following three-point equality

$$\|a - c\|^2 = \|a - b\|^2 + \|b - c\|^2 + 2\langle a - b, b - c\rangle$$

with $a = \bar{u}$, $b = u$, $c = z$ and then divide the result by the factor 2α, thus obtaining

$$\frac{1}{2\alpha}\|u - z\|^2 = \frac{1}{2\alpha}\|\bar{u} - z\|^2 - \frac{1}{2\alpha}\|\bar{u} - u\|^2 - \langle u - z, \nabla f(\bar{u}) + A^T v\rangle, \ \forall z \in \mathbb{R}^d. \tag{22}$$

Starting from $\mathcal{L}(z, v)$, we can write the following chain of inequalities.

$$
\begin{aligned}
\mathcal{L}(z, v) &= f(z) + \langle Az, v \rangle - g^*(v) \\
&\geq f(\bar{u}) + \langle z - \bar{u}, \nabla f(\bar{u}) \rangle + \langle Az, v \rangle - g^*(v) \\
&= f(\bar{u}) + \langle u - \bar{u}, \nabla f(\bar{u}) \rangle - \langle u - z, \nabla f(\bar{u}) \rangle + \langle Az, v \rangle - g^*(v) \\
&\geq f(u) - \frac{L}{2} \|u - \bar{u}\|^2 + \frac{1}{2\alpha} \|u - z\|^2 - \frac{1}{2\alpha} \|\bar{u} - z\|^2 \\
&\quad + \frac{1}{2\alpha} \|\bar{u} - u\|^2 + \langle u - z, A^T v \rangle + \langle Az, v \rangle - g^*(v) \\
&= \underbrace{f(u) + \langle u, A^T v \rangle - g^*(v)}_{=\mathcal{L}(u,v)} + \frac{1}{2\alpha} \|u - z\|^2 - \frac{1}{2\alpha} \|\bar{u} - z\|^2 + \frac{1}{2} \left(\frac{1}{\alpha} - L \right) \|u - \bar{u}\|^2
\end{aligned}
\tag{23}
$$

where the first inequality follows from the convexity of f, and the second one from Eq. (22) combined with the descent lemma on f. The result follows by reordering the terms in (23).

(ii) Let us consider the inequality obtained using Lemma 4 with $x = \bar{v}$, $e = \beta \alpha^{-1} A \tilde{u}$, $y = v$ and then multiply it by $(\alpha \beta^{-1})/2$, namely

$$
\frac{\alpha}{2\beta} \|v - z'\|^2 \leq \frac{\alpha}{2\beta} \|\bar{v} - z'\|^2 - \frac{\alpha}{2\beta} \|\bar{v} - v\|^2 + \langle v - z', A\tilde{u} \rangle + g^*(z') - g^*(v), \ \forall \, z' \in \mathbb{R}^{d'}.
\tag{24}
$$

Starting from $\mathcal{L}(u, v)$, we can write the following chain of inequalities.

$$
\mathcal{L}(u, v) = f(u) + \langle Au, v \rangle - g^*(v)
\tag{25}
$$

$$
\begin{aligned}
&\geq \underbrace{f(u) + \langle Au, z' \rangle - g^*(z')}_{=\mathcal{L}(u,z')} + \langle z' - v, A(\tilde{u} - u) \rangle \\
&\quad + \frac{\alpha}{2\beta} \|\bar{v} - v\|^2 - \frac{\alpha}{2\beta} \|\bar{v} - z'\|^2 + \frac{\alpha}{2\beta} \|v - z'\|^2 \\
&= \mathcal{L}(u, z') + \langle z' - v, A(\tilde{u} - u) \rangle + \frac{\alpha}{2\beta} \|\bar{v} - v\|^2 - \frac{\alpha}{2\beta} \|\bar{v} - z'\|^2 + \frac{\alpha}{2\beta} \|v - z'\|^2
\end{aligned}
\tag{26}
$$

From (18), we also have

$$
\tilde{u} - u = \alpha A^T (v - \bar{v}).
\tag{27}
$$

Plugging (27) inside (25) yields

$$
\mathcal{L}(u, v) \geq \mathcal{L}(u, z') + \alpha \langle A^T(z' - v), A^T(v - \bar{v}) \rangle + \frac{\alpha}{2\beta} \|\bar{v} - v\|^2 - \frac{\alpha}{2\beta} \|\bar{v} - z'\|^2 + \frac{\alpha}{2\beta} \|v - z'\|^2.
\tag{28}
$$

Now observe that the scalar product in the previous inequality can be rewritten as

$$
\langle A^T(z' - v), A^T(v - \bar{v}) \rangle = -\frac{1}{2} \|A^T(z' - v)\|^2 - \frac{1}{2} \|A^T(v - \bar{v})\|^2 + \frac{1}{2} \|A^T(z' - \bar{v})\|^2.
$$

Finally, inserting the previous fact inside (28) and recalling the definition of the norm $\| \cdot \|_A$ leads to

$$\mathcal{L}(u, v) \geq \mathcal{L}(u, z') + \frac{\alpha}{2\beta} \|v - \bar{v}\|_A^2 - \frac{\alpha}{2\beta} \|\bar{v} - z'\|_A + \frac{\alpha}{2\beta} \|z' - v\|_A^2.$$

By reordering the terms in the above inequality, we get the thesis.

(iii) The inequality follows by combining points (i)-(ii). □

Remark 2 Note that Eq. (25) is a specialized version of the result in [35, Lemma 3], obtained by setting $\check{y} = v, \check{x} = u, \bar{y} = \bar{v}, \bar{x} = \bar{u}, \tau = \alpha, \sigma = \beta \alpha^{-1}, K = A, \epsilon = 0, \delta = 0,$ $e = 0$, and most importantly, $\tilde{y} = \check{y}$ and $g \equiv 0$. In particular, the condition $\tilde{y} = \check{y}$ makes the inner product $\langle K(x - \check{x}), \tilde{y} - \check{y} \rangle$ disappear in [35, Lemma 3]; the assumption $g \equiv 0$ allows us to rewrite (25) as (28) through relation (27) to get the result. Hence, we can say that our Lemma 7(iii) is a specialized version of [35, Lemma 3], which is appropriately rewritten by employing our specific assumptions $\tilde{y} = \check{y}$ and $g \equiv 0$.

The next two lemmas are also needed to prove convergence of Algorithm 1. The former has been employed in the convergence analysis of several FISTA variants with inexact proximal evaluations (see e.g. [9, 41]), whereas the second one is a classical tool for proving convergence of first-order algorithms.

Lemma 8 [41, *Lemma* 1] *Let* $\{p_n\}_{n \in \mathbb{N}}, \{q_n\}_{n \in \mathbb{N}}, \{\lambda_n\}_{n \in \mathbb{N}}$ *be sequences of real non-negative numbers, with* $\{q_n\}_{n \in \mathbb{N}}$ *being a monotone nondecreasing sequence, satisfying the following recursive property*

$$p_n^2 \leq q_n + \sum_{k=1}^n \lambda_k p_k, \quad \forall\, n \geq 1. \tag{29}$$

Then the following inequality holds:

$$p_n \leq \frac{1}{2} \sum_{k=1}^n \lambda_k + \left(q_n + \left(\frac{1}{2} \sum_{k=1}^n \lambda_k \right)^2 \right)^{\frac{1}{2}}, \quad \forall\, n \geq 1. \tag{30}$$

Lemma 9 [34] *Let* $\{p_n\}_{n \in \mathbb{N}}, \{\zeta_n\}_{n \in \mathbb{N}}$ *and* $\{\xi_n\}_{n \in \mathbb{N}}$ *be sequences of real nonnegative numbers such that* $p_{n+1} \leq (1 + \zeta_n)p_n + \xi_n$ *and* $\sum_{n=0}^{\infty} \zeta_n < \infty, \sum_{n=0}^{\infty} \xi_n < \infty.$ *Then,* $\{p_n\}_{n \in \mathbb{N}}$ *converges.*

In the following theorem, we state and prove the convergence of the primal-dual sequence $\{(u_n, v_n^0)\}_{n \in \mathbb{N}}$ generated by Algorithm 1 to a solution of the saddle-point problem (7), provided that the inertial parameters satisfy an appropriate technical assumption. Our result extends the convergence theorem in [17, Theorem 3.1], which holds only for the case $\gamma_n \equiv 0$. As in [17], the line of proof adopted here can be summarized according to the following points: (i) Prove that the sequence $\{(u_n, v_n^0)\}_{n \in \mathbb{N}}$ is convergent; (ii) Show that each limit point of $\{(u_n, v_n^0)\}_{n \in \mathbb{N}}$ is a solution to the saddle-point problem (7); (iii) Conclude that

the primal–dual sequence is converging to a saddle-point. However, some major modifications are required here in order to take into account the presence of the inertial step on the primal variable.

Theorem 2 *Suppose that f, g, A are defined as in problem* (5) *under Assumptions* 1. *Let* $\{(u_n, v_n^0)\}_{n \in \mathbb{N}}$ *be the primal-dual sequence generated by Algorithm* 1. *Suppose that the parameters sequence* $\{\gamma_n\}_{n \in \mathbb{N}}$ *satisfies the following condition*:

$$\sum_{n=0}^{\infty} \gamma_n \|u_n - u_{n-1}\| < \infty. \tag{31}$$

Then the following statements hold true:

  (i)   *the sequence* $\{(u_n, v_n^0)\}_{n \in \mathbb{N}}$ *is bounded*;
 (ii)   *given* $(\hat{u}, \hat{v})$ *solution of* (7), *the sequence* $\{\beta k_{\max} \|\hat{u} - u_n\|^2 + \alpha^2 \|\hat{v} - v_n^0\|_A^2\}_{n \in \mathbb{N}}$ *converges*;
(iii)   *the sequence* $\{(u_n, v_n^0)\}_{n \in \mathbb{N}}$ *converges to a solution of* (7).

Proof (i) Let $(\hat{u}, \hat{v}) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ be a solution of problem (7), i.e.

$$\mathcal{L}(\hat{u}, v) \leq \mathcal{L}(\hat{u}, \hat{v}) \leq \mathcal{L}(u, \hat{v}), \quad \forall u \in \mathbb{R}^d, v \in \mathbb{R}^{d'}.$$

By directly applying Eq. (21) with $z = \hat{u}$, $z' = \hat{v}$, $u = u_n^{k+1}$, $\bar{u} = \bar{u}_n$, $v = v_n^{k+1}$, $\bar{v} = v_n^k$, for $k = 0, \ldots, k_{\max} - 1$, observing that $\mathcal{L}(u_n^{k+1}, \hat{v}) \geq \mathcal{L}(\hat{u}, v_n^{k+1})$ and discarding the terms proportional to $\|u_n^{k+1} - \bar{u}_n\|^2$ and $\|v_n^{k+1} - v_n^k\|_A^2$, we obtain the following inequality

$$\frac{1}{2\alpha\beta} \left(\beta \|\hat{u} - u_n^{k+1}\|^2 + \alpha^2 \|\hat{v} - v_n^{k+1}\|_A^2 \right) \leq \frac{1}{2\alpha\beta} \left(\beta \|\hat{u} - \bar{u}_n\|^2 + \alpha^2 \|\hat{v} - v_n^k\|_A^2 \right),$$

$$k = 0, \ldots, k_{\max} - 1.$$

Recalling that $\bar{u}_n = u_n + \gamma_n(u_n - u_{n-1})$ and using the Cauchy-Schwarz inequality, we can write the following inequalities

$$\frac{1}{2\alpha\beta} \left(\beta \|\hat{u} - u_n^{k+1}\|^2 + \alpha^2 \|\hat{v} - v_n^{k+1}\|_A^2 \right) \leq \frac{1}{2\alpha\beta} \left(\beta \|\hat{u} - u_n\|^2 + \alpha^2 \|\hat{v} - v_n^k\|_A^2 \right)$$

$$+ \frac{1}{2\alpha} \left(\gamma_n^2 \|u_n - u_{n-1}\|^2 - 2\gamma_n \langle \hat{u} - u_n, u_n - u_{n-1} \rangle \right)$$

$$\leq \frac{1}{2\alpha\beta} \left(\beta \|\hat{u} - u_n\|^2 + \alpha^2 \|\hat{v} - v_n^k\|_A^2 \right)$$

$$+ \frac{1}{2\alpha} \left(\gamma_n^2 \|u_n - u_{n-1}\|^2 + 2\gamma_n \|\hat{u} - u_n\| \|u_n - u_{n-1}\| \right)$$

$$k = 0, \ldots, k_{\max} - 1.$$

Summing the previous relation for $k = 0, \ldots, k_{\max} - 1$ allows to write

$$\frac{1}{2\alpha\beta}\left(\beta\sum_{k=0}^{k_{\max}-1}\|\hat{u}-u_n^{k+1}\|^2 + \alpha^2\|\hat{v}-v_n^{k_{\max}}\|_A^2\right) \leq \frac{1}{2\alpha\beta}\left(\beta k_{\max}\|\hat{u}-u_n\|^2 + \alpha^2\|\hat{v}-v_n^0\|_A^2\right)$$
$$+ \frac{k_{\max}}{2\alpha}\left(\gamma_n^2\|u_n-u_{n-1}\|^2 + 2\gamma_n\|\hat{u}-u_n\|\|u_n-u_{n-1}\|\right).$$
(32)

By exploiting the update rule (16) together with the convexity of the function $\phi(u) = \|\hat{u}-u\|^2$ and recalling the warm-start strategy $v_n^{k_{\max}} = v_{n+1}^0$, it easily follows that

$$\beta\sum_{k=0}^{k_{\max}-1}\|\hat{u}-u_n^{k+1}\|^2 + \alpha^2\|\hat{v}-v_n^{k_{\max}}\|_A^2 \geq \beta k_{\max}\|\hat{u}-u_{n+1}\|^2 + \alpha^2\|\hat{v}-v_{n+1}^0\|_A^2.$$

Plugging the previous inequality inside (32) leads to

$$\frac{1}{2\alpha\beta}(\beta k_{\max}\|\hat{u}-u_{n+1}\|^2 + \alpha^2\|\hat{v}-v_{n+1}^0\|_A^2) \leq \frac{1}{2\alpha\beta}\left(\beta k_{\max}\|\hat{u}-u_n\|^2 + \alpha^2\|\hat{v}-v_n^0\|_A^2\right)$$
$$+ \frac{k_{\max}}{2\alpha}\left(\gamma_n^2\|u_n-u_{n-1}\|^2 + 2\gamma_n\|\hat{u}-u_n\|\|u_n-u_{n-1}\|\right).$$
(33)

If we apply recursively inequality (33), we get

$$\frac{1}{2\alpha\beta}(\beta k_{\max}\|\hat{u}-u_{n+1}\|^2 + \alpha^2\|\hat{v}-v_{n+1}^0\|_A^2) \leq \frac{1}{2\alpha\beta}\left(\beta k_{\max}\|\hat{u}-u_1\|^2 + \alpha^2\|\hat{v}-v_1^0\|_A^2\right)$$
$$+ \frac{k_{\max}}{2\alpha}\left(\sum_{k=1}^{n}\gamma_k^2\|u_k-u_{k-1}\|^2 + \sum_{k=1}^{n}2\gamma_k\|\hat{u}-u_k\|\|u_k-u_{k-1}\|\right).$$
(34)

By discarding the term proportional to $\|\hat{v}-v_{n+1}^0\|_A^2$ in the left-hand side, multiplying both sides by $2\alpha/k_{\max}$ and adding the term $\gamma_{n+1}^2\|u_{n+1}-u_n\|^2 + 2\gamma_{n+1}\|\hat{u}-u_{n+1}\|\|u_{n+1}-u_n\|$ to the right-hand side, we come to

$$\|\hat{u}-u_{n+1}\|^2 \leq \|\hat{u}-u_1\|^2 + \frac{\alpha^2}{\beta k_{\max}}\|\hat{v}-v_1^0\|_A^2 + \sum_{k=1}^{n+1}\gamma_k^2\|u_k-u_{k-1}\|^2 + \sum_{k=1}^{n+1}2\gamma_k\|\hat{u}-u_k\|\|u_k-u_{k-1}\|.$$

Finally, we can use Lemma 8 with $p_n = \|\hat{u}-u_n\|$, $q_n = \|\hat{u}-u_1\|^2 + \frac{\alpha^2}{\beta k_{\max}}\|\hat{v}-v_1^0\|_A^2 + \sum_{k=1}^{n}\gamma_k^2\|u_k-u_{k-1}\|^2$ and $\lambda_n = 2\gamma_n\|u_n-u_{n-1}\|$, thus obtaining

$$\|\hat{u}-u_{n+1}\| \leq \frac{1}{2}\sum_{k=1}^{n+1}2\gamma_k\|u_k-u_{k-1}\|$$
$$+ \left(\|\hat{u}-u_1\|^2 + \frac{\alpha^2}{\beta k_{\max}}\|\hat{v}-v_1^0\|_A^2 + \sum_{k=1}^{n+1}\gamma_k^2\|u_k-u_{k-1}\|^2 + \left(\frac{1}{2}\sum_{k=1}^{n+1}2\gamma_k\|u_k-u_{k-1}\|\right)^2\right)^{\frac{1}{2}}.$$

From the previous inequality combined with condition (31), it follows that the sequence $\{u_n\}_{n\in\mathbb{N}}$ is bounded. By discarding the term proportional to $\|\hat{u}-u_{n+1}\|^2$

in Eq. (34), using again (31) and the boundedness of $\{u_n\}_{n\in\mathbb{N}}$, we also deduce the boundedness of $\{v_n^0\}_{n\in\mathbb{N}}$.

(ii) Point (i) guarantees the existence of a constant $M > 0$ such that $\|\hat{u} - u_n\| \leq M$ for all $n \in \mathbb{N}$. Then, from Eq. (33), we get

$$\frac{1}{2\alpha\beta}\left(\beta k_{\max}\|\hat{u} - u_{n+1}\|^2 + \alpha^2\|\hat{v} - v_{n+1}^0\|_A^2\right) \leq \frac{1}{2\alpha\beta}\left(\beta k_{\max}\|\hat{u} - u_n\|^2 + \alpha^2\|\hat{v} - v_n^0\|_A^2\right)$$
$$+ \frac{k_{\max}}{2\alpha}\left(\gamma_n^2\|u_n - u_{n-1}\|^2 + 2M\gamma_n\|u_n - u_{n-1}\|\right).$$

The previous inequality and condition (31) allows to apply Lemma 9 with $p_n = \beta k_{\max}\|\hat{u} - u_n\|^2 + \alpha^2\|\hat{v} - v_n^0\|_A^2$, $\zeta_n \equiv 0$, $\xi_n = \beta k_{\max}\left(\gamma_n^2\|u_n - u_{n-1}\|^2 + 2M\gamma_n\|u_n - u_{n-1}\|\right)$, thus we get the thesis.

(iii) Since $\{(u_n, v_n^0)\}_{n\in\mathbb{N}}$ is bounded, there exists a point $(u^\dagger, v^\dagger)$ and $I \subseteq \mathbb{N}$ such that $\lim_{n\in I}(u_n, v_n^0) = (u^\dagger, v^\dagger)$. Furthermore, based on condition (31), we can write

$$\lim_{n\to\infty}\|\bar{u}_n - u_n\| = \lim_{n\to\infty}\gamma_n\|u_n - u_{n-1}\| = 0.$$

Therefore $\lim_{n\in I}\|\bar{u}_n - u_n\| = 0$, which implies $\lim_{n\in I}\bar{u}_n = u^\dagger$. By writing down again Eq. (21) with $z = \hat{u}$, $z' = \hat{v}$, $u = u_n^{k+1}$, $\bar{u} = \bar{u}_n$, $v = v_n^{k+1}$, $\bar{v} = v_n^k$, where $(\hat{u}, \hat{v})$ is a solution of (7), and using $\mathcal{L}(u_n^{k+1}, \hat{v}) \geq \mathcal{L}(\hat{u}, v_n^{k+1})$, we obtain for $k = 0, \ldots, k_{\max} - 1$

$$\beta(1 - \alpha L)\|u_n^{k+1} - \bar{u}_n\|^2 + \alpha^2\|v_n^{k+1} - v_n^k\|_A^2 \leq \beta\|\hat{u} - \bar{u}_n\|^2 - \beta\|\hat{u} - u_n^{k+1}\|^2$$
$$+ \alpha^2\|\hat{v} - v_n^k\|_A^2 - \alpha^2\|\hat{v} - v_n^{k+1}\|_A^2.$$

By employing the definition of $\bar{u}_n$, the Cauchy-Schwarz inequality and the boundedness of $\{u_n\}_{n\in\mathbb{N}}$ inside the previous relation, we come to

$$\beta(1 - \alpha L)\|u_n^{k+1} - \bar{u}_n\|^2 + \alpha^2\|v_n^{k+1} - v_n^k\|_A^2 \leq \beta\|\hat{u} - u_n\|^2 - \beta\|\hat{u} - u_n^{k+1}\|^2$$
$$+ \alpha^2\|\hat{v} - v_n^k\|_A^2 - \alpha^2\|\hat{v} - v_n^{k+1}\|_A^2$$
$$+ \beta\left(\gamma_n^2\|u_n - u_{n-1}\|^2 + 2M\gamma_n\|u_n - u_{n-1}\|\right).$$

Summing the inequality for $k = 0, \ldots, k_{\max} - 1$ leads to

$$\sum_{k=0}^{k_{\max}-1}\beta(1 - \alpha L)\|u_n^{k+1} - \bar{u}_n\|^2 + \alpha^2\|v_n^{k+1} - v_n^k\|_A^2 \leq \beta k_{\max}\|\hat{u} - u_n\|^2$$
$$- \beta\sum_{k=0}^{k_{\max}-1}\|\hat{u} - u_n^{k+1}\|^2 + \alpha^2\|\hat{v} - v_n^0\|_A^2 - \alpha^2\|\hat{v} - v_n^{k_{\max}}\|_A^2$$
$$+ \beta k_{\max}\left(\gamma_n^2\|u_n - u_{n-1}\|^2 + 2M\gamma_n\|u_n - u_{n-1}\|\right).$$

If we combine the above inequality with the update rule (16), the convexity of the function $\phi(u) = \|\hat{u} - u\|^2$ and the warm-start strategy $v_n^{k_{\max}} = v_{n+1}^0$, we get

$$\sum_{k=0}^{k_{\max}-1} \beta(1-\alpha L)\|u_n^{k+1} - \bar{u}_n\|^2 + \alpha^2\|v_n^{k+1} - v_n^k\|_A^2 \leq \beta k_{\max}\|\hat{u} - u_n\|^2$$
$$- \beta k_{\max}\|\hat{u} - u_{n+1}\|^2 + \alpha^2\|\hat{v} - v_n^0\|_A^2 - \alpha^2\|\hat{v} - v_{n+1}^0\|_A^2$$
$$+ \beta k_{\max}\left(\gamma_n^2\|u_n - u_{n-1}\|^2 + 2M\gamma_n\|u_n - u_{n-1}\|\right).$$

We now sum the previous relation over $n = 1, \dots, N$, thus obtaining

$$\sum_{n=1}^{N} \sum_{k=0}^{k_{\max}-1} \beta(1-\alpha L)\|u_n^{k+1} - \bar{u}_n\|^2 + \alpha^2\|v_n^{k+1} - v_n^k\|_A^2 \leq \beta k_{\max}\|\hat{u} - u_1\|^2 + \alpha^2\|\hat{v} - v_1^0\|_A^2$$
$$+ \beta k_{\max} \sum_{n=1}^{N} \left(\gamma_n^2\|u_n - u_{n-1}\|^2 + 2M\gamma_n\|u_n - u_{n-1}\|\right).$$

Taking the limit for $N \to \infty$ and recalling condition (31), it follows that

$$\lim_{n\to\infty} \sum_{k=0}^{k_{\max}-1} \beta(1-\alpha L)\|u_n^{k+1} - \bar{u}_n\|^2 + \alpha^2\|v_n^{k+1} - v_n^k\|_A^2 = 0.$$

Consequently, we have $\lim_{n\to\infty} \beta(1-\alpha L)\|u_n^{k+1} - \bar{u}_n\|^2 + \alpha^2\|v_n^{k+1} - v_n^k\|_A^2 = 0$ for $k = 0, \dots, k_{\max} - 1$. For $k = 0$, this implies that $\lim_{n\in I}(u_n^1, v_n^1) = (u^\dagger, v^\dagger)$. By combining this fact with the continuity of $\text{prox}_{\beta\alpha^{-1}g^*}$ inside the steps (14)-(15), we conclude that $(u^\dagger, v^\dagger)$ satisfies the fixed-point Eq. (6), which is equivalent to say that $(u^\dagger, v^\dagger)$ is a solution of (7) (see Lemma 6). Since $(u^\dagger, v^\dagger)$ is a saddle point, it follows from point (ii) that the sequence $\{\beta k_{\max}\|u^\dagger - u_n\|^2 + \alpha^2\|v^\dagger - v_n^0\|_A^2\}_{n\in\mathbb{N}}$ converges and, by definition of limit point, it admits a subsequence converging to zero. Then the sequence $\{(u_n, v_n^0)\}_{n\in\mathbb{N}}$ must converge to $(u^\dagger, v^\dagger)$. □

Remark 3 Condition (31) is quite similar to the one adopted by Lorenz and Pock in [28, Eq. 23] for their primal–dual forward–backward algorithm, which applies extrapolation on both the primal and dual variable and then performs a primal–dual iteration in the same fashion as in the popular "Chambolle–Pock" algorithm [15, 16]. Indeed, the inertial parameter α_n adopted in [28] needs to satisfy the following condition

$$\sum_{n=1}^{\infty} \alpha_n\|(x_n, y_n) - (x_{n-1}, y_{n-1})\|^2 < \infty, \tag{35}$$

where $\{x_n\}_{n\in\mathbb{N}}, \{y_n\}_{n\in\mathbb{N}}$ denote the primal and dual iterates of the algorithm, respectively. Differently from (31), condition (35) depends also on the dual variable and squares the norm of the gap between two successive iterates.

As also noted in [28], conditions (31) and (35) can be easily enforced "online", since they depend only on the past iterates. One possible strategy to compute γ_n according to (31) consists in modifying the usual FISTA inertial parameter as follows

$$\gamma_n = \begin{cases} 0, & n = 0 \\ \min\left\{\gamma_n^{FISTA}, \dfrac{C\rho_n}{\|u_n - u_{n-1}\|}\right\}, & n = 1, 2, \dots \end{cases} \tag{36}$$

where $C > 0$ is a positive constant, $\{\rho_n\}_{n \in \mathbb{N}}$ is a prefixed summable sequence and γ_n^{FISTA} is computed according to the usual FISTA rule, namely [5]

$$t_0 = 1, \quad \begin{cases} t_{n+1} & = \dfrac{1+\sqrt{1+4t_n^2}}{2} \\ \gamma_n^{FISTA} & = \dfrac{t_n - 1}{t_{n+1}} \end{cases} \quad n = 0, 1, \dots \tag{37}$$

In the numerical section, we will see that the modified FISTA-like rule (36) has the practical effect of shrinking the inertial parameter after a certain number of iterations, thus avoiding the possible divergence of the algorithm from the true solution of (5).

We now prove an ergodic $O(1/N)$ convergence rate for Algorithm 1 in terms of the quantity $\mathcal{L}(U_{N+1}, z') - \mathcal{L}(z, \overline{V}_{N+1})$, where $(z, z') \in \mathbb{R}^d \times \mathbb{R}^{d'}$ and U_{N+1} and $\overline{V}_{N+1}$ are the ergodic sequences. On the one hand, our result generalizes the convergence rate obtained in [23, Theorem 3.1] under the assumptions $\gamma_n \equiv 0$ and $k_{max} = 1$, which in turn extends a previous result due to [29, Theorem 1] holding for Algorithm 1 in the case $f(u) = \frac{1}{2}\|Hu - y\|^2$, $\gamma_n \equiv 0$, $k_{max} = 1$. On the other hand, the result is novel also in the case $\gamma_n \equiv 0$ and $k_{max} \geq 1$, for which we recover the nested primal–dual algorithm presented in [17].

Theorem 3 *Suppose that f, g, A are defined as in problem* (5) *under Assumptions* 1. *Let $\{(u_n, v_n^0)\}_{n \in \mathbb{N}}$ be the primal-dual sequence generated by Algorithm 1 with inertial parameters $\{\gamma_n\}_{n \in \mathbb{N}}$ satisfying* (31). *For all $n \geq 1$, define the sequence $\bar{v}_n = \dfrac{1}{k_{max}} \sum_{k=0}^{k_{max}-1} v_{n-1}^{k+1}$. Then, given $(\hat{u}, \hat{v})$ solution of* (7), $U_N = \frac{1}{N} \sum_{n=1}^{N} u_n$ *and $\overline{V}_N = \frac{1}{N} \sum_{n=1}^{N} \bar{v}_n$ the ergodic primal-dual sequences, and for any $(z, z') \in \mathbb{R}^d \times \mathbb{R}^{d'}$, we have*

$$\mathcal{L}(U_{N+1}, z') - \mathcal{L}(z, \overline{V}_{N+1}) \leq \frac{\frac{1}{2\alpha}\|z - u_0\|^2 + \frac{\alpha}{2\beta k_{max}}\|z' - v_0^0\|_A^2 + \frac{\gamma(\gamma+2M)}{2\alpha}}{N + 1}, \tag{38}$$

where $M > 0$ is such that $\|u_n - z\| \leq M$ for all $n \geq 0$, and $\gamma = \sum_{n=0}^{\infty} \gamma_n \|u_n - u_{n-1}\| < \infty$.

Proof Considering Eq. (21) with $u = u_n^{k+1}$, $\bar{u} = \bar{u}_n$, $v = v_n^{k+1}$, $\bar{v} = v_n^k$, and discarding the terms proportional to $\|u_n^{k+1} - \bar{u}_n\|^2$ and $\|v_n^{k+1} - v_n^k\|_A^2$, we can write

$$\mathcal{L}(u_n^{k+1}, z') - \mathcal{L}(z, v_n^{k+1}) \leq \frac{1}{2\alpha} \left(\|z - \bar{u}_n\|^2 - \|z - u_n^{k+1}\|^2 \right)$$
$$+ \frac{\alpha}{2\beta} \left(\|z' - v_n^k\|_A^2 - \|z' - v_n^{k+1}\|_A^2 \right)$$
$$\leq \frac{1}{2\alpha} \left(\|z - u_n\|^2 - \|z - u_n^{k+1}\|^2 \right)$$
$$+ \frac{\alpha}{2\beta} \left(\|z' - v_n^k\|_A^2 - \|z' - v_n^{k+1}\|_A^2 \right)$$
$$+ \frac{1}{2\alpha} \left(\gamma_n^2 \|u_n - u_{n-1}\|^2 - 2\gamma_n \langle z - u_n, u_n - u_{n-1} \rangle \right)$$
$$\leq \frac{1}{2\alpha} \left(\|z - u_n\|^2 - \|z - u_n^{k+1}\|^2 \right)$$
$$+ \frac{\alpha}{2\beta} \left(\|z' - v_n^k\|_A^2 - \|z' - v_n^{k+1}\|_A^2 \right)$$
$$+ \frac{1}{2\alpha} \left(\gamma_n^2 \|u_n - u_{n-1}\|^2 + 2M\gamma_n \|u_n - u_{n-1}\| \right),$$

where the last relation follows from the Cauchy-Schwarz inequality and the boundedness of the sequence $\{u_n\}_{n \in \mathbb{N}}$, being $M > 0$ such that $\|z - u_n\| \leq M$ for all $n \in \mathbb{N}$.

Summing the previous inequality for $k = 0, \ldots, k_{\max} - 1$ and exploiting the update rule (16), the convexity of the function $\phi(u) = \|z - u\|^2$ and the warm-start strategy $v_n^{k_{\max}} = v_{n+1}^0$ in the right-hand side, we obtain

$$\sum_{k=0}^{k_{\max}-1} (\mathcal{L}(u_n^{k+1}, z') - \mathcal{L}(z, v_n^{k+1})) \leq \frac{k_{\max}}{2\alpha} \left(\|z - u_n\|^2 - \|z - u_{n+1}\|^2 \right)$$
$$+ \frac{\alpha}{2\beta} \left(\|z' - v_n^0\|_A^2 - \|z' - v_{n+1}^0\|_A^2 \right) \tag{39}$$
$$+ \frac{k_{\max}}{2\alpha} \left(\gamma_n^2 \|u_n - u_{n-1}\|^2 + 2M\gamma_n \|u_n - u_{n-1}\| \right).$$

For the left-hand side of the previous inequality, it is possible to write the following lower bound

$$\sum_{k=0}^{k_{\max}-1} (\mathcal{L}(u_n^{k+1}, z') - \mathcal{L}(z, v_n^{k+1})) = k_{\max} \sum_{k=0}^{k_{\max}-1} \frac{1}{k_{\max}} (\mathcal{L}(u_n^{k+1}, z') - \mathcal{L}(z, v_n^{k+1}))$$
$$\geq k_{\max} (\mathcal{L}(u_{n+1}, z') - \mathcal{L}(z, \bar{v}_{n+1})),$$

where the second inequality follows from the convexity of the function $\mathcal{L}(\cdot, z') - \mathcal{L}(z, \cdot)$, the update rule (16), and the definition of the sequence $\{\bar{v}_n\}_{n \in \mathbb{N}}$. Combining the last lower bound with (39) results in

$$\mathcal{L}(u_{n+1}, z') - \mathcal{L}(z, \bar{v}_{n+1}) \leq \frac{1}{2\alpha}\left(\|z - u_n\|^2 - \|z - u_{n+1}\|^2\right)$$
$$+ \frac{\alpha}{2\beta k_{\max}}\left(\|z' - v_n^0\|_A^2 - \|z' - v_{n+1}^0\|_A^2\right)$$
$$+ \frac{1}{2\alpha}\left(\gamma_n^2\|u_n - u_{n-1}\|^2 + 2M\gamma_n\|u_n - u_{n-1}\|\right).$$

Summing the previous inequality for $n = 0, \dots, N$, multiplying and dividing the left-hand side by $N + 1$, exploiting the convexity of the function $\mathcal{L}(\cdot, z') - \mathcal{L}(z, \cdot)$ and condition (31) yields

$$(N + 1)\Big(\mathcal{L}(U_{N+1}, z') - \mathcal{L}(z, \overline{V}_{N+1})\Big) \leq \frac{1}{2\alpha}\|z - u_0\|^2 + \frac{\alpha}{2\beta k_{\max}}\|z' - v_0^0\|_A^2 + \frac{\gamma(\gamma + 2M)}{2\alpha},$$

from which the thesis follows. $\qquad\square$

Remark 4 Theorem 3 is analogous to the one obtained in [35, Theorem 3] in the (non strongly) convex case for an inexact version of the Chambolle-Pock primal–dual method (17), where ϵ_n−approximations of the proximal operators prox_f and prox_{g*} are employed. In particular, assuming that $\epsilon_n = O(1/n^\alpha)$ with $\alpha > 0$, the authors in [35] show that the convergence rate on the quantity $\mathcal{L}(U_{N+1}, z') - \mathcal{L}(z, \overline{V}_{N+1})$ is either $O(1/N)$ if $\alpha > 1$ (similarly to our Theorem 3), $O(\ln(N)/N)$ if $\alpha = 1$, or $O(1/N^\alpha)$ if $\alpha \in (0, 1)$. Furthermore, they also derive improved convergence rates for a couple of accelerated versions of the same algorithm, under the assumption that one of the terms in the objective function is strongly convex [35, Corollary 2-3]. Although we were not able to show accelerated rates in the strongly convex scenario, a linear rate for Algorithm 1 with $\gamma_n \equiv 0$ has already been given in [17, Theorem 3.2], hence the extension of this result to the case $\gamma_n \neq 0$ is a possible topic for future research.

Remark 5 Note that the numerator of (38) gets smaller as $k_{\max}$ increases, suggesting that a higher number of inner primal–dual iterations corresponds to a better convergence rate factor. Such a result is coherent with the one in [17, Theorem 3.2], where a convergence rate including a $1/k_{\max}$ factor is provided for Algorithm 1 with $\gamma_n \equiv 0$, under the assumption that f is strongly convex. However, one has to consider that a bigger $k_{\max}$ also corresponds to a higher computational cost per iteration. Hence, the choice of an appropriate $k_{\max}$ is crucial for the practical performance of the algorithm. We refer the reader to Sect. 4 for some preliminary numerical investigation on this issue.

Theorem 3 allows us to derive an $O(1/N)$ convergence rate on the primal function values, as shown in the following corollary.

Corollary 1 *Suppose that the same assumptions of Theorem 3 hold and, additionally, the function g has full domain. Then, we have*

$$F(U_{N+1}) - F(\hat{u}) \leq \frac{\frac{1}{2\alpha}\|\hat{u} - u_0\|^2 + \frac{\alpha M'}{2\beta k_{\max}} + \frac{\gamma(\gamma + 2M)}{2\alpha}}{N + 1}, \tag{40}$$

where $M' > 0$ is such that $\|z'_N - v^0_0\|^2_A \leq M'$ for all $z'_N \in \partial g(AU_N)$, $N \geq 0$.

Proof Following e.g. [35, Remark 3], we have that, if g has full domain, then g^* is superlinear, i.e. $g^*(v)/|v| \to \infty$ as $|v| \to \infty$. As a result, the supremum of $\mathcal{L}(U_{N+1}, z')$ over z' is attained at some point $z'_{N+1} \in \partial g(AU_{N+1})$, namely

$$\sup_{z' \in \mathbb{R}^{d'}} \mathcal{L}(U_{N+1}, z') = \mathcal{L}(U_{N+1}, z'_{N+1}). \tag{41}$$

Then, given a solution $\hat{u}$ to problem (5), we can write the following chain of inequalities

$$\begin{aligned}
F(U_{N+1}) - F(\hat{u}) &= (f(U_{N+1}) + g(AU_{N+1})) - (f(\hat{u}) + g(A\hat{u})) \\
&= \sup_{z' \in \mathbb{R}^{d'}} \mathcal{L}(U_{N+1}, z') - \sup_{z' \in \mathbb{R}^{d'}} \mathcal{L}(\hat{u}, z') \\
&\leq \mathcal{L}(U_{N+1}, z'_{N+1}) - \mathcal{L}(\hat{u}, \overline{V}_{N+1}) \\
&\leq \frac{\frac{1}{2\alpha}\|\hat{u} - u_0\|^2 + \frac{\alpha}{2\beta k_{\max}}\|z'_{N+1} - v^0_0\|^2_A + \frac{\gamma(\gamma + 2M)}{2\alpha}}{N + 1},
\end{aligned} \tag{42}$$

where the second equality is due to Lemma 2, the third inequality follows from (41) and the definition of supremum, and the fourth inequality is an application of the rate (38). Since U_N is a bounded sequence (due to Theorem 2(i)), $z'_{N+1} \in \partial g(AU_{N+1})$, and g is locally Lipschitz, it follows that z'_{N+1} is uniformly bounded over N, which implies that the sequence $\{\|z'_{N+1} - v^0_0\|^2_A\}_{N \in \mathbb{N}}$ is bounded by a constant $M' > 0$. Hence, the thesis follows. □

Analogously, one could prove an $O(1/N)$ convergence rate for the primal-dual gap $\mathcal{G}(U_{N+1}, \overline{V}_{N+1}) = \sup_{(z,z') \in \mathbb{R}^d \times \mathbb{R}^{d'}} \mathcal{L}(U_{N+1}, z') - \mathcal{L}(z, \overline{V}_{N+1})$, see e.g. [15, Remark 3].

4 Numerical experiments

We are now going to show the numerical performance of Algorithm 1 on some image deblurring test problems where the observed image has been corrupted by Poisson noise. All the experiments are carried out by running the routines in MAT-LAB R2019a on a laptop equipped with a 2.60 GHz Intel Core i7-4510U processor and 8 GB of RAM.

4.1 Weighted ℓ_2-TV image restoration under Poisson noise

Variational models encoding signal-dependent Poisson noise have become ubiquitous in microscopy and astronomy imaging applications [6]. Let $z \in \mathbb{R}^d$ with

$z_i > 0$ $i = 1, \ldots, d$ be the observed image, which is assumed to be a realization of a Poisson-distributed $d-$dimensional random vector, namely

$$z = \mathcal{P}(Hu),$$

where $u \in \mathbb{R}^d$ is the ground truth, $H \in \mathbb{R}^{d \times d}$ is the blurring matrix associated to a given Point Spread Function (PSF), and $\mathcal{P}(w)$ denotes a realization of a Poisson-distributed random vector with parameter $w \in \mathbb{R}^d$, $w_i > 0$ for $i = 1, \ldots, d$. According to the Maximum a Posteriori approach, the problem of recovering the unknown image u from the acquired image z consists in solving the following penalized optimization problem

$$\min_{u \in \mathbb{R}^d} KL(Hu;z) + \lambda R(u),$$

where

$$KL(Hu;z) := \sum_{i=1}^{d} z_i \log \left(\frac{z_i}{(Hu)_i} \right) + (Hu)_i - z_i \qquad (43)$$

denotes the generalized Kullback-Leibler (KL) divergence, playing the role of the data-fidelity term, $\lambda > 0$ is the regularization parameter, and $R(u)$ is a penalty term encoding some a-priori information on the object to be reconstructed. In the following, we consider an approximation of the KL divergence that is obtained by truncating its Taylor expansion around the vector z at the second order. Introducing the diagonal positive definite matrix $W := \text{diag}(1/z)$, whose elements on the diagonal are defined as $W_{i,i} = 1/z_i$ for $i = 1, \ldots, d$, the desired KL approximation can be written as the following weighted least squares data fidelity term

$$\frac{1}{2} \|Hu - z\|_W^2 := \frac{1}{2} \langle Hu - z, W(Hu - z) \rangle = \frac{1}{2} \sum_{i=1}^{d} \frac{((Hu)_i - z_i)^2}{z_i}, \qquad (44)$$

where the weighted norm is defined as $\|u\|_W = \sqrt{\langle u, Wu \rangle}$. The data term (44) promotes high fidelity in low-intensity pixels and large regularization in high-intensity pixels (i.e. in presence of high levels of noise). Such an approximation has already been considered for applications such as positron emission tomography and super-resolution microscopy, see e.g. [12, 27, 36, 40].

As we want to impose edge-preserving regularization on the unknown image, we couple the data fidelity function (44) with the classical isotropic Total Variation (TV) function [38]

$$TV(u) = \|\nabla u\|_{2,1} = \sum_{i=1}^{d} \|\nabla_i u\|_2, \qquad (45)$$

where $\nabla_i u \in \mathbb{R}^2$ represents the standard forward difference discretization of the image gradient at pixel i. In conclusion, we aim at solving the following minimization problem

$$\min_{u \in \mathbb{R}^d} \frac{1}{2} \|Hu - z\|_W^2 + \lambda TV(u). \tag{46}$$

Problem (46) can be framed in (5) by setting

$$f(u) := \frac{1}{2} \|Hu - z\|_W^2 \tag{47}$$

$$g(v) := \sum_{i=1}^d \lambda \left\| \begin{pmatrix} v_{2i-1} \\ v_{2i} \end{pmatrix} \right\|_2$$
$$A := (\nabla_1^T \ \nabla_2^T \ \cdots \ \nabla_d^T)^T \in \mathbb{R}^{2d \times d} \tag{48}$$

where $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^{2d}$. Note that g is convex and continuous, and ∇f is L–Lipschitz continuous on $\mathbb{R}^d$ with the constant L being estimated as $L \leq \|A^T W A\|_2 \leq \|A\|_2^2 \|W\|_2$, where $\|W\|_2 = 1/\min\{z_i : i = 1, \ldots, d\}$. Furthermore, problem (46) admits at least one solution under mild assumptions on H, for instance when H is such that $H_{ij} \geq 0 \ i, j = 1, \ldots, d$ and has at least one strictly positive component for each row and column [1]. Hence Assumptions 1 are satisfied, Theorem 2 holds and the iterates $\{u_n\}_{n \in \mathbb{N}}$ of Algorithm 1 are guaranteed to converge to a solution of problem (46), provided that the inertial parameters $\{\gamma_n\}_{n \in \mathbb{N}}$ satisfy condition (31).

Note that each step of Algorithm 1 can be computed in closed form, if applied to problem (46): indeed A is simply the gradient operator (48) and A^T its associated divergence, the spectral norm $\|A\|$ can be explicitly upper bounded with the value $\sqrt{8}$, and prox_{g^*} is the projection operator onto the cartesian product $B(0, \lambda) \times B(0, \lambda) \times \cdots \times B(0, \lambda)$, where $B(0, \lambda) \subseteq \mathbb{R}^2$ is a ball of center 0 and radius λ (see [13]).

4.1.1 Implementation and parameters choice

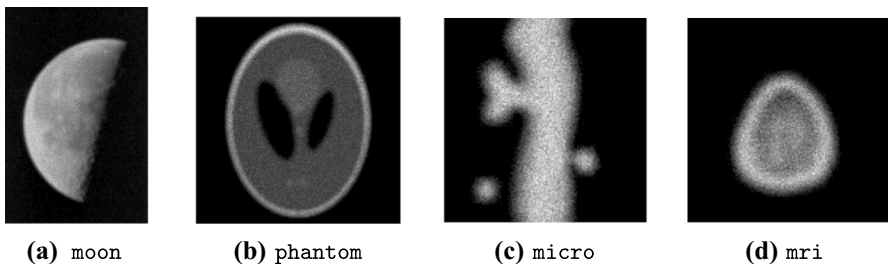For our numerical tests, we consider four grayscale images, which are displayed in Fig. 1:

(a) `moon` **(b)** `phantom` **(c)** `micro` **(d)** `mri`

Fig. 1 Blurred and noisy test images used our numerical experiments

- `moon`, `phantom`, `mri` are demo images of size 537×358, 256×256, and 128×128 respectively, which are taken from MATLAB Image Processing Toolbox;
- `micro` is the 128×128 confocal microscopy phantom described in [46].

The regularization parameter has been fixed equal to $\lambda = 0.15$ for `moon`, $\lambda = 0.12$ for `phantom`, $\lambda = 0.09$ for `micro`, and $\lambda = 0.12$ for `mri`. The blurring matrix H is associated to a Gaussian blur with window size 9×9 and standard deviation equal to 4, and products with H (respectively H^T) are performed by assuming reflexive boundary conditions. We simulate numerically the blurred and noisy acquired images by convolving the true object with the given PSF and then using the `imnoise` function of the MATLAB Image Processing Toolbox to simulate the presence of Poisson noise.

We implement several versions of Algorithm 1, each one equipped with a different rule for the computation of the inertial parameter γ_n at STEP 1 and a different strategy for the initialization of the inner loop. The resulting algorithms are denoted as follows.

- FB (warm): the nested non-inertial primal–dual algorithm formerly presented in [17], which is obtained as a special case of Algorithm 1 by setting $\gamma_n \equiv 0$;
- FB (no-warm): a nested non-inertial primal–dual algorithm obtained from Algorithm 1 by setting $\gamma_n \equiv 0$ and replacing STEP 2 with the initialization $v_n^0 = 0$; due to the lack of the "warm-start" strategy, Theorem 2 does not apply to this algorithm, and hence its convergence is not guaranteed;
- FISTA (warm): a particular instance of Algorithm 1 obtained by choosing γ_n as the FISTA inertial parameter γ_n^{FISTA} defined in (37); note that we do not have any information on whether γ_n^{FISTA} complies with condition (31) or not, hence it is unclear a priori whether the iterates will converge or not to a minimum point;
- FISTA (no-warm): here we set $\gamma_n = \gamma_n^{FISTA}$ and use the null dual vector $v_n^0 = 0$ as initialization of the inner loop at STEP 2;
- FISTA-like (warm): this is Algorithm 1 with the inertial parameter γ_n computed according to the rule (36), in which we set $C = 10\|u_1 - u_0\|$ and $\rho_n = 1/n^{1.1}$; as explained in Remark 3, such a practical rule ensures that condition (31) holds, therefore by Theorem 2 we are guaranteed that the iterates of the algorithm will converge to a solution of (5);
- FISTA-like (no-warm): this is just as "FISTA-like (warm)" but with the "warm-start" strategy at STEP 2 replaced by $v_n^0 = 0$; again no theoretical guarantees are available in this case.

For all algorithms, we set the initial primal iterate as the observed image, namely $u_0 = u_{-1} = z$, the initial dual iterate as the null vector, i.e. $v_{-1}^{k_{\max}} = 0$, the primal steplength as $\alpha = (1 - \epsilon)/(8\|W\|_2)$, and the dual steplength as $\beta = (1 - \epsilon)/8$, where $\epsilon > 0$ is the machine precision. Regarding the number of primal–dual iterations, we will test different values of $k_{\max}$ in the next section.

4.1.2 Results

As a first test, we run all previously described algorithms with $k_{\max} = 1$ for 2000 iterations. In Fig. 2, we show the decrease of the primal function values $F(u_n)$ with respect to the iteration number for all test images. We observe that the "FB" and "FISTA–like" implementations of Algorithm 1 that make use of the "warm-start" strategy converge to a smaller function value than the ones provided by the other competitors. These two instances of Algorithm 1 are indeed the only ones whose convergence to a minimum point is guaranteed by Theorem 2. On the contrary, "FISTA (warm)" is the only warm–started algorithm that tends to a higher function value, and in fact it provides the worst reconstructions for all test problems, as can be seen in the comparison of the reconstructed `phantom` images in Fig. 4. The other algorithms using a null initialization for the inner loop all fail to converge to the smallest function value. Such numerical failures are typical of nested inertial primal–dual algorithms that do not employ a "warm-start" initialization in the inner loop, see e.g. [4, 18].
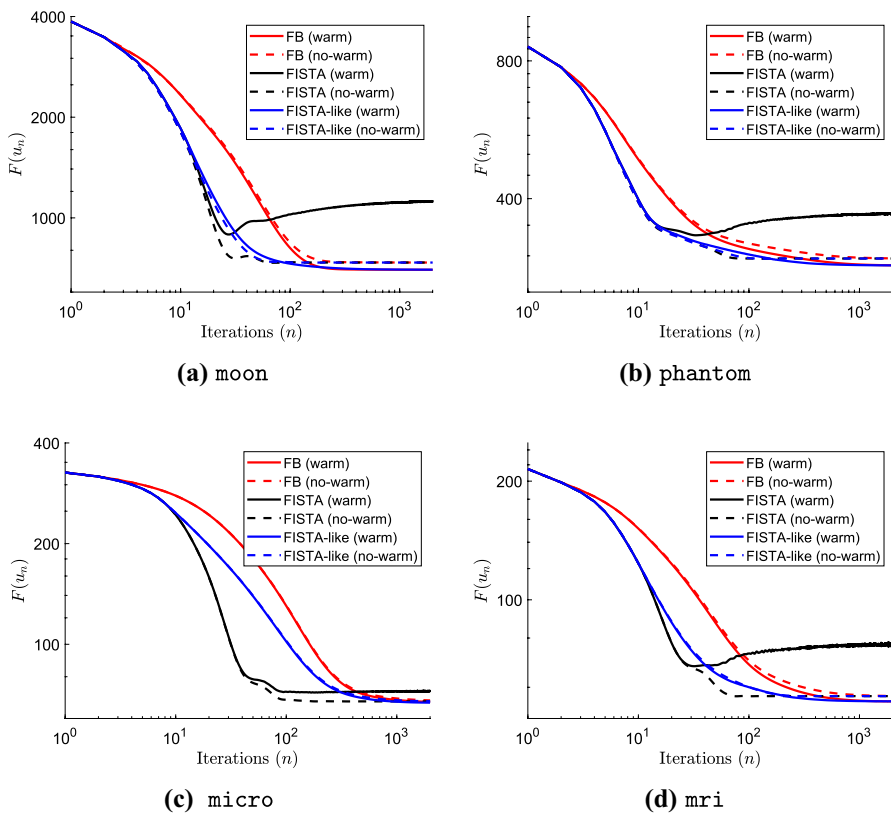
Fig. 2 Decrease of the objective function values $F(u_n)$ with respect to the iteration number in the case $k_{\max} = 1$

(a) `moon`

(b) `phantom`

(c) `micro`

(d) `mri`
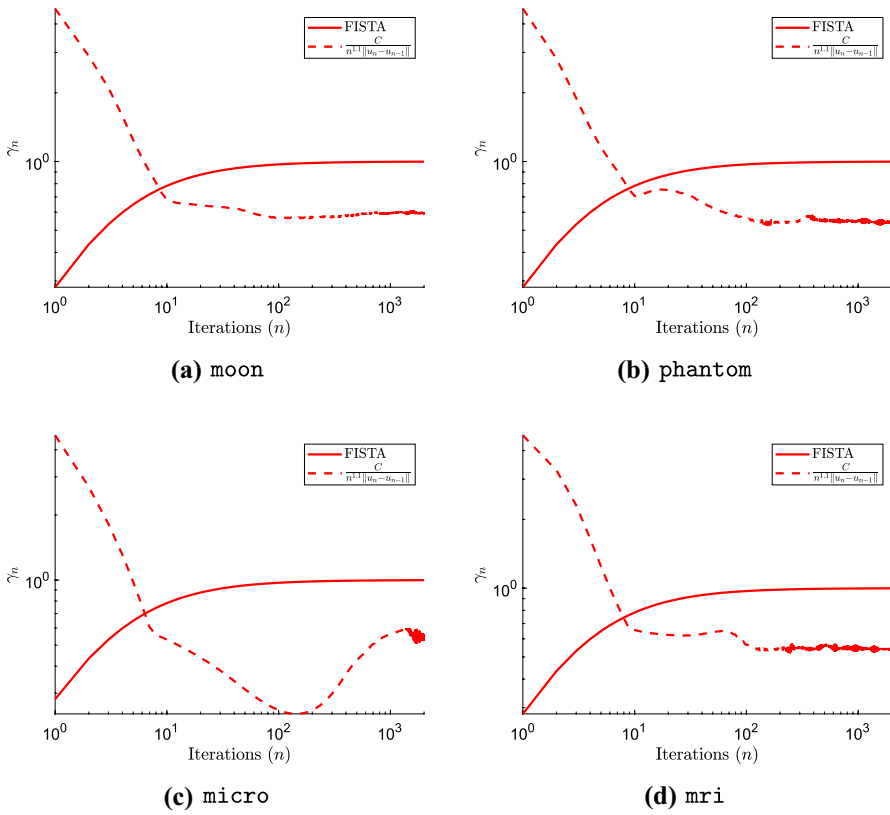
Fig. 3 Comparison between FISTA inertial parameter and the sequence $\{C/(n^{1.1}\|u_n - u_{n-1}\|)\}_{n \in \mathbb{N}}$ in the case $k_{\max} = 1$

(a) FISTA (warm)   **(b)** FISTA (no-warm)   **(c)** FISTA-like (warm)
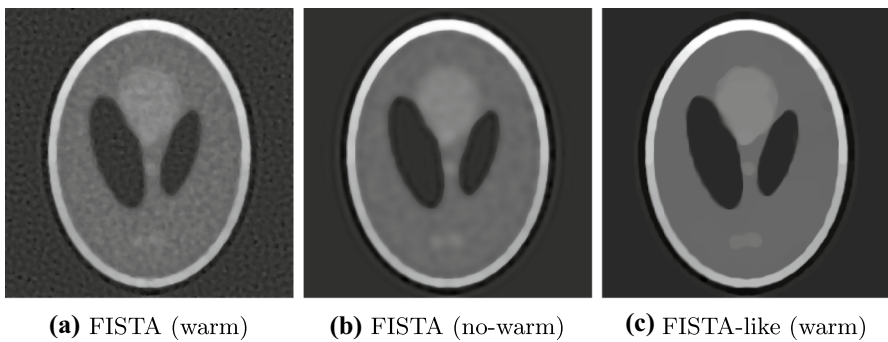
Fig. 4 Reconstructions provided by FISTA (warm), FISTA (no-warm), FISTA-like (warm) with $k_{\max} = 1$ after 2000 iterations

Table 1 Test image `phantom`. Function values after 2000 iterations attained by the algorithms for different values of $k_{\max}$

	$k_{\max} = 1$	$k_{\max} = 5$	$k_{\max} = 10$	$k_{\max} = 20$
FB (warm)	286.05	286.05	286.04	286.04
FB (no-warm)	296.23	289.69	288.06	287.10
FISTA (warm)	370.49	303.39	294.12	289.42
FISTA (no-warm)	295.98	289.41	287.76	286.79
FISTA-like (warm)	285.93	285.88	285.82	285.80
FISTA-like (no-warm)	295.98	289.41	287.76	286.79

(a) `moon`

(b) `moon`
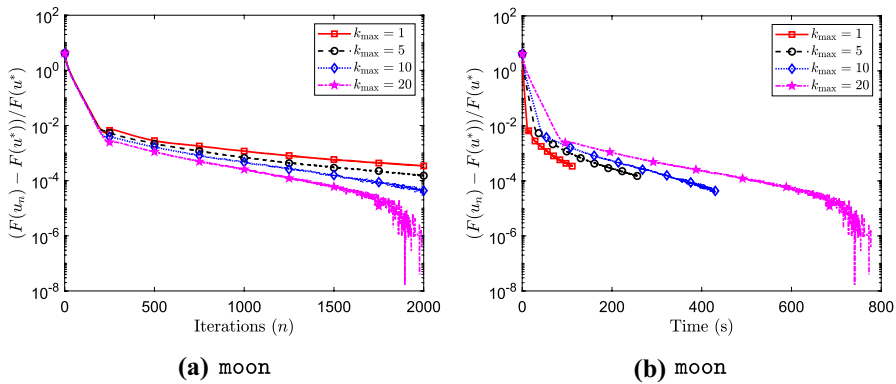
Fig. 5 Test image `moon`. Relative decrease of the function values of FISTA-like (warm) with respect to the iteration number (left) and time (right) for different values of $k_{\max}$

Focusing on the "FISTA–like" variants, we observe that these algorithms perform identically to the "FISTA" variants in the first iterations, then move away towards smaller function values. This is due to the computational rule (36), which has the practical effect of shrinking the inertial parameter after a certain number of iterations. Indeed Fig. 3 shows that the parameter $C/(n^{1.1}\|u_n - u_{n-1}\|)$ becomes smaller than γ_n^{FISTA} after the first $10 - 20$ iterations, and then seemingly stabilizes around a value smaller than 1. However, we also remark that condition (36) alone is not enough for ensuring convergence to the true minimum value, as the "FISTA–like (no-warm)" implementation stabilizes to a higher function value than the one provided by the warm-started FISTA-like algorithm for all test images. Hence, we conclude that the combination of condition (36) with the "warm-start" strategy on the inner loop is strictly necessary in order to prevent the divergence of Algorithm 1 from the true solution of (5).

We now evaluate the impact of the number of inner primal–dual iterations on the overall performance of Algorithm 1. In Table 1 we report the function value provided by the different algorithms after 2000 iterations on the test image `phantom`, as $k_{\max}$ varies from 1 to 20. On the one hand, we see that all algorithms yield roughly the same function value if $k_{\max}$ is large, however at the expense of a much higher computational cost per outer iteration. On the other hand, the warm-started FISTA–like implementation provides the best accuracy level already with $k_{\max} = 1$,
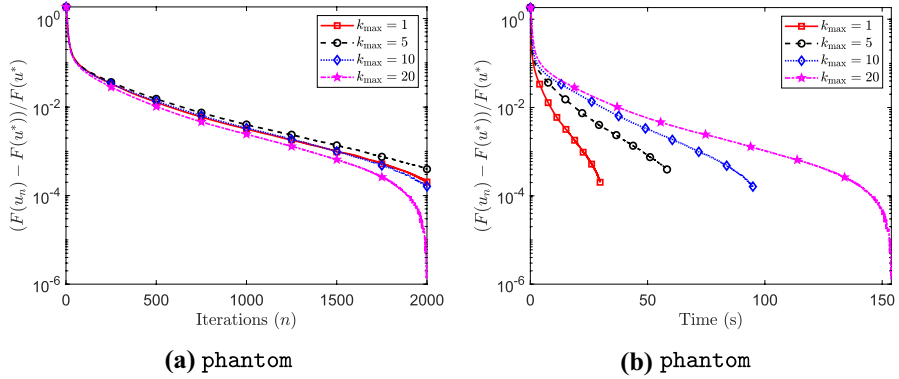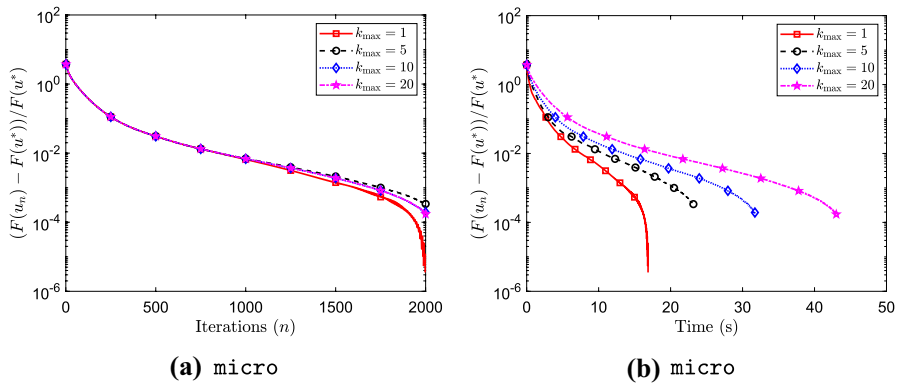
(a) `phantom`   **(b)** `phantom`

Fig. 6 Test image `phantom`. Relative decrease of the function values of FISTA-like (warm) with respect to the iteration number (left) and time (right) for different values of k_{max}
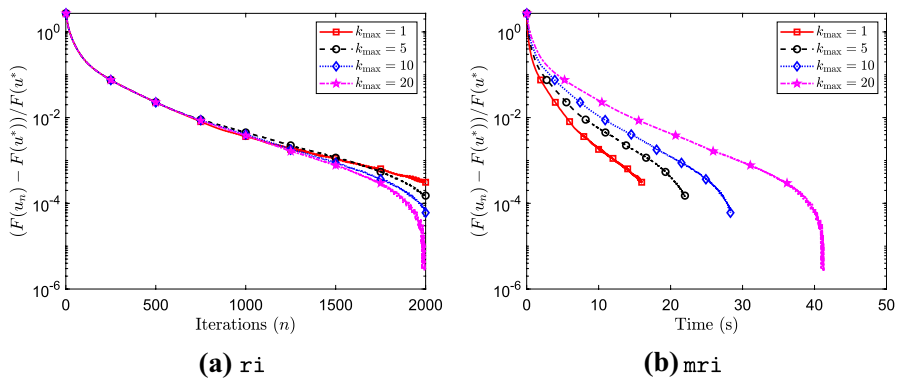

(a) `micro`   **(b)** `micro`

Fig. 7 Test image `micro`. Relative decrease of the function values of FISTA-like (warm) with respect to the iteration number (left) and time (right) for different values of k_{max}

(a) `ri`   **(b)** `mri`

Fig. 8 Test image `mri`. Relative decrease of the function values of FISTA-like (warm) with respect to the iteration number (left) and time (right) for different values of k_{max}

and increasing $k_{\max}$ does not seem to be particularly beneficial for this variant. In Figs. 5, 6, 7 and 8, we show the relative decrease $(F(u_n) - F(u^*))/F(u^*)$ of the warm-started FISTA–like algorithm for several values of $k_{\max}$, with respect to both iterations and time, where $F(u^*)$ is the minimum value among the ones attained after 2000 iterations for the different values of $k_{\max}$. While the plots on the left show a similar performance in terms of iterations as $k_{\max}$ varies, the plots on the right confirm that a higher value of $k_{\max}$ often results in an increased computational effort to achieve the desired accuracy level.

4.2 KL-TV image restoration under Poisson noise

In this section, we consider again the problem of image restoration under Poisson noise, i.e. the problem of recovering the original image $u \in \mathbb{R}^d$ from a distorted image $z \in \mathbb{R}^d$, $z_i \geq 0$ $i = 1, \ldots, d$ that is a realization of a Poisson random vector of the form

$$z = \mathcal{P}(Hu + b\boldsymbol{e}),$$

where $H \in \mathbb{R}^{d \times d}$ is again the blurring matrix, $\boldsymbol{e}$ is the vector of all ones, and $b > 0$ is a constant term taking into account the existence of some background emission. This time, we follow the Maximum a Posteriori approach without approximating the KL divergence, and imposing both TV regularization and non-negativity on the image pixels. Therefore, we address the following optimization problem

$$\min_{u \in \mathbb{R}^d} KL(Hu + b\boldsymbol{e};z) + \lambda TV(u) + \iota_\Omega(u), \tag{49}$$

where the functions KL and TV are defined in (43) and (45), respectively, $\Omega = \{u \in \mathbb{R}^d : u_i \geq 0, i = 1, \ldots, d\}$ denotes the non-negative orthant, and ι_Ω is the indicator function of Ω, i.e.

$$\iota_\Omega(u) = \begin{cases} 0, & u \in \Omega \\ \infty, & \text{otherwise.} \end{cases}$$

Note that problem (49) can be included in (5) by setting

$$f(u) := 0 \tag{50}$$

$$g(v) := \sum_{i=1}^d \lambda \left\| \begin{pmatrix} v_{2i-1} \\ v_{2i} \end{pmatrix} \right\|_2 + \iota_\Omega \begin{pmatrix} v_{2d+1} \\ \vdots \\ v_{3d} \end{pmatrix} + KL \left(\begin{pmatrix} v_{3d+1} \\ \vdots \\ v_{4d} \end{pmatrix} + b\boldsymbol{e};z \right) \tag{51}$$

$$A := (\nabla_1^T \ \nabla_2^T \ \cdots \ \nabla_d^T \ I_d \ H^T)^T \in \mathbb{R}^{4d \times d}$$

where $u \in \mathbb{R}^d$, $v \in \mathbb{R}^{4d}$, and $I_d \in \mathbb{R}^{d \times d}$ is the identity matrix of size d. We remark that f is convex, continuously differentiable, and its gradient is L–Lipschitz continuous for any constant $L > 0$, whereas g is convex and lower semicontinuous. In addition, the problem admits at least one solution under mild assumptions on H, for instance the same ones assumed for problem (46) [24, Proposition 3]. Thus,

Theorem 2 guarantees once again the convergence of the iterates $\{u_n\}_{n\in\mathbb{N}}$ of Algorithm 1 to a solution of problem (49).

As in the previous test, we remark that each step of Algorithm 1 can be performed explicitly. Indeed, the blocks of the matrix A are the discrete gradient operator ∇, the identity matrix I, and the blurring matrix H, all of which can be easily applied in closed form. Furthermore, the spectral norm of A can be upper bounded as $\|A\|^2 \le \|\nabla\|^2 + \|H^T H\| + 1$, where $\|\nabla\|^2 \le 8$ and $\|H^T H\| \le \|H\|_1 \|H\|_\infty$. Finally, the term g can be written as the sum of separable functions $g = g_1 + g_2 + g_3$, where $g_1(v_1,\dots,v_{2d}) = \sum_{i=1}^d \lambda\|(v_{2i-1}\ v_{2i})\|_2$, $\quad g_2(v_{2d+1},\dots,v_{3d}) = \iota_\Omega(v_{2d+1},\dots,v_{3d})$, and $g_3(v_{3d+1},\dots,v_{4d}) = KL((v_{3d+1},\dots,v_{4d}) + b\boldsymbol{e};z)$, so that prox_{g^*} can be block-decomposed as

$$\mathrm{prox}_{g^*}(v) = \Big(\mathrm{prox}_{g_1^*}(v_1,\dots,v_{2d}), \mathrm{prox}_{g_2^*}(v_{2d+1},\dots,v_{3d}), \mathrm{prox}_{g_3^*}(v_{3d+1},\dots,v_{4d})\Big),$$

where $\mathrm{prox}_{g_1^*}$ is the projection operator onto the cartesian product $B(0,\lambda) \times B(0,\lambda) \times \cdots \times B(0,\lambda)$, $\mathrm{prox}_{g_2^*}$ is the projection operator onto the orthant $\Omega^* = \{v \in \mathbb{R}^d : v_i \le 0,\ i = 1,\dots,d\}$, and $\mathrm{prox}_{g_3^*}$ can be computed according to a closed-form formula [19, Table 2]. Therefore, prox_{g^*} is overall computable in closed form.

4.2.1 Implementation and parameters choice

For the following tests, we consider the grayscale images `phantom`, `mri`, `micro` used in the tests of section 4.1, and two different types of blur: a Gaussian blur with window size 9×9, and an out-of-focus blur with radius 4. The products with the blurring matrix H (respectively H^T) are performed by assuming periodic boundary conditions. For each test image and blur, we convolve the true object with the corresponding PSF, add a constant background term $b = 10^{-6}$, and then apply the `imnoise` function of the MATLAB Image Processing Toolbox to simulate the presence of Poisson noise. We consider two different Signal-to-Noise (SNR) ratios (66 and 72 dB), which are imposed by suitably pre-scaling the images before applying the `imnoise` function. We recall that the SNR in presence of Poisson noise is estimated by [6]

$$\mathrm{SNR} = 10\log_{10}\left(\frac{n_{\text{exact}}}{\sqrt{n_{\text{exact}} + n_{\text{background}}}}\right),$$

where n_{exact} and $n_{\text{background}}$ represent the total number of photons in the image and in the background term, respectively.

We compare Algorithm 1 with some well-known optimization methods suited for the deblurring of Poissonian images. We detail the implemented algorithms below.

- Algorithm 1: we equip our proposed algorithm with $k_{\max} \in \{1, 5, 10\}$, γ_n computed according to the rule (36) with the same choices of C, ρ_n adopted in section 4.1, the initial dual iterate as $v_{-1}^{k_{\max}} = 0$, the primal steplength as $\alpha = 10^{-2}$,

and the dual steplength as $\beta = 0.9/(9 + \|H\|_1\|H\|_\infty)$. We note that, although the sequence $\{u_n\}_{n\in\mathbb{N}}$ is guaranteed to converge to a solution of problem (49), some of the iterates might fall outside the domain Ω, as the primal sub-iterates u_n^k in STEP 3 are not projected onto Ω. As a partial remedy to this issue, we report the results for the projected sequence $\{P_\Omega(u_n)\}_{n\in\mathbb{N}}$, where P_Ω denotes the projection operator onto Ω. Such a sequence is both feasible and convergent to a solution of (49), even though we can not state anything regarding its convergence rate, as Theorem 3 is applicable only to the non-projected sequence $\{u_n\}_{n\in\mathbb{N}}$.

- CP: this is the popular Chambolle-Pock primal-dual algorithm [15], whose general iteration is reported in (17). Here we apply CP to problem (49) by setting $f(u) = \iota_\Omega(u)$, $g(v) = \sum_{i=1}^d \lambda\|(v_{2i-1}\ v_{2i})\|_2 + KL((v_{3d+1},\dots,v_{4d}) + b\boldsymbol{e};z)$, and $A = (\nabla_1^T\ \nabla_2^T\ \cdots\ \nabla_d^T\ H^T)^T \in \mathbb{R}^{3d\times d}$. Since the steplength parameters τ, σ need to satisfy the condition $\tau\sigma < 1/\|A\|^2$ to guarantee the convergence of the sequence, we choose the same primal steplength as in Algorithm 1, i.e., $\tau = 10^{-2}$, and $\sigma = 0.9\tau^{-1}/(8 + \|H\|_1\|H\|_\infty)$.

- PIDSplit+: this is an efficient algorithm based on an alternating split Bregman technique, which was first proposed in [42]. The PIDSplit+ algorithm depends on a parameter $\gamma > 0$, which is set here as $\gamma = 50/\lambda$, as suggested in [42]. Note that there exist also adaptive strategies for computing this parameter, see e.g. [47].

For all algorithms, we set the initial primal iterate as the observed image, namely $u_0 = z$. For each test problem, the regularization parameter λ is tuned by a trial-and-error strategy. In particular, we run the CP algorithm by varying λ in the set $\{10^{-3}, 5\cdot10^{-3}, 10^{-2}, 5\cdot10^{-2}, 10^{-1}, 5\cdot10^{-1}\}$, performing 5000 iterations and using $u_0 = z$ at each run. The value of λ providing the smallest relative error $\|u_n - u^\dagger\|/\|u^\dagger\|$ is retained as the regularization parameter, being $u^\dagger$ the ground truth and u_n the last iterate.

4.2.2 Results

We apply Algorithm 1, PIDSplit+, and CP on 12 test problems, which are generated by varying the test image (phantom, mri, micro), the blur (Gaussian or out-of-focus), and the noise level (66 or 72 dB). For each test problem, we run all algorithms for 1000 iterations.

In Figs. 9, 10, 11, and 12, we report the relative decrease of the function values $(F(u_n) - F(u^*))/F(u^*)$, where u^* is a precomputed solution obtained by running the CP algorithm for 5000 iterations. We can appreciate the acceleration obtained by Algorithm 1 with respect to the iteration number, whereas a higher $k_{\max}$ usually increases the computational time without a significant gain in terms of accuracy. With $k_{\max} = 1$, the computational times of Algorithm 1 are comparable to the ones of CP, even though our proposed algorithm is sometimes faster in the first seconds
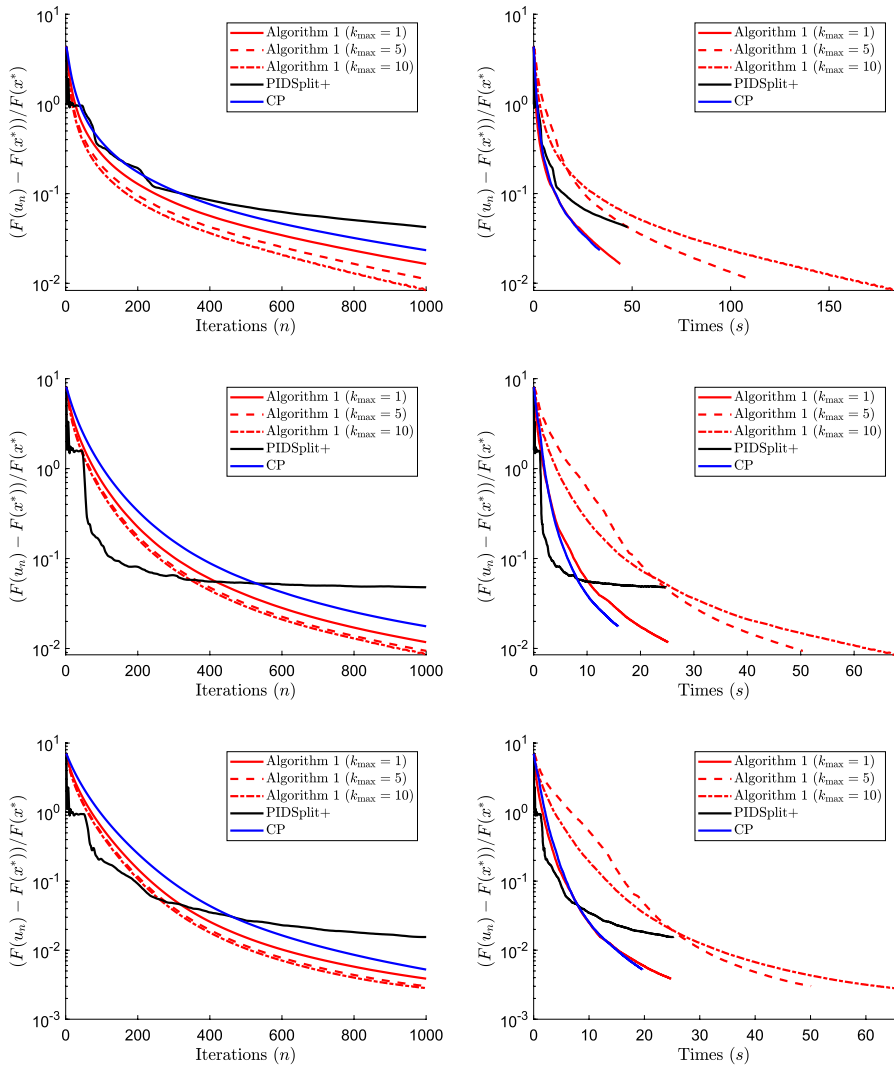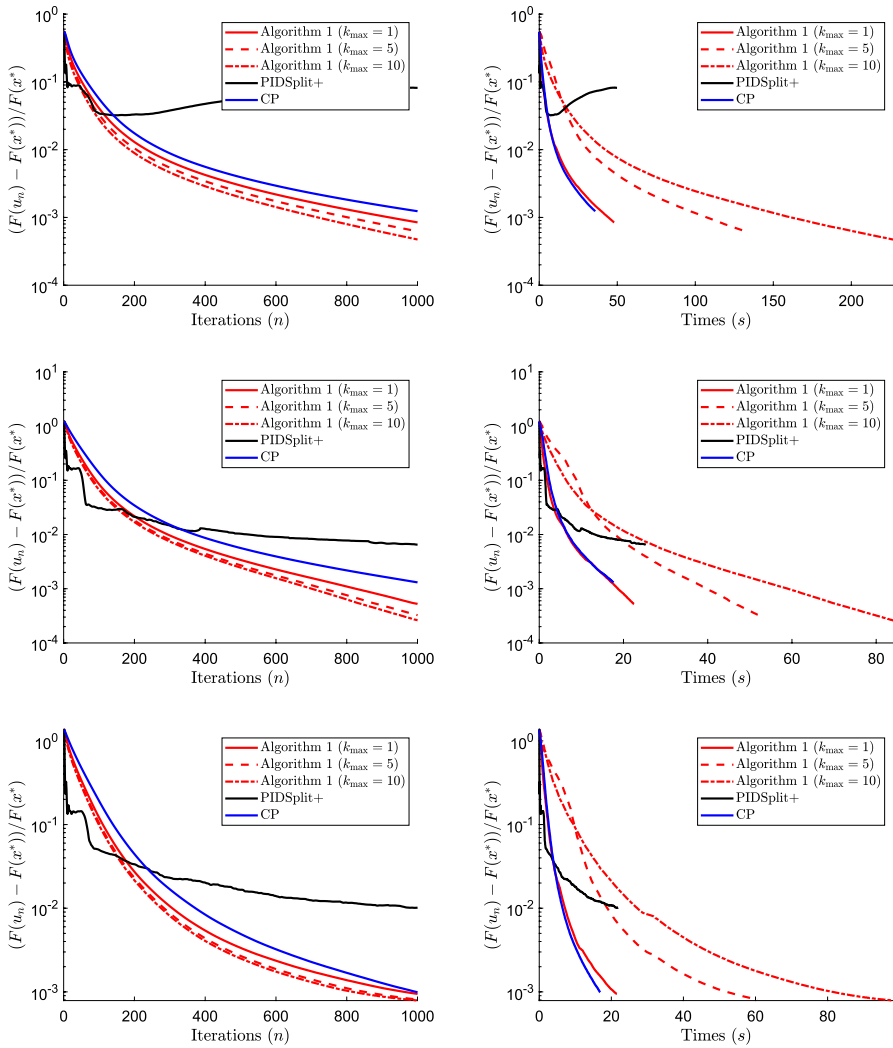
Fig. 9 Test problems with Gaussian blur, SNR = 66 dB. From top to bottom: test images `phantom`, `mri`, `micro`. Relative decrease of the function values with respect to the iteration number (left) and time (right)

(see e.g. Fig. 11). Overall, Algorithm 1 exhibits a rate of convergence towards the minimum point that is similar for each combination of noise level and blur.

In Table 2, we show the values for the structural similarity (SSIM) index [45] as a measure of the quality of the reconstructions provided by the algorithms. As it can be seen, all methods provide acceptable SSIM values, and are quite stable with respect to the noise level and blur. The quality of the reconstructions

Fig. 10 Test problems with Gaussian blur, SNR = 72 dB. From top to bottom: test images `phantom`, `mri`, `micro`. Relative decrease of the function values with respect to the iteration number (left) and time (right)
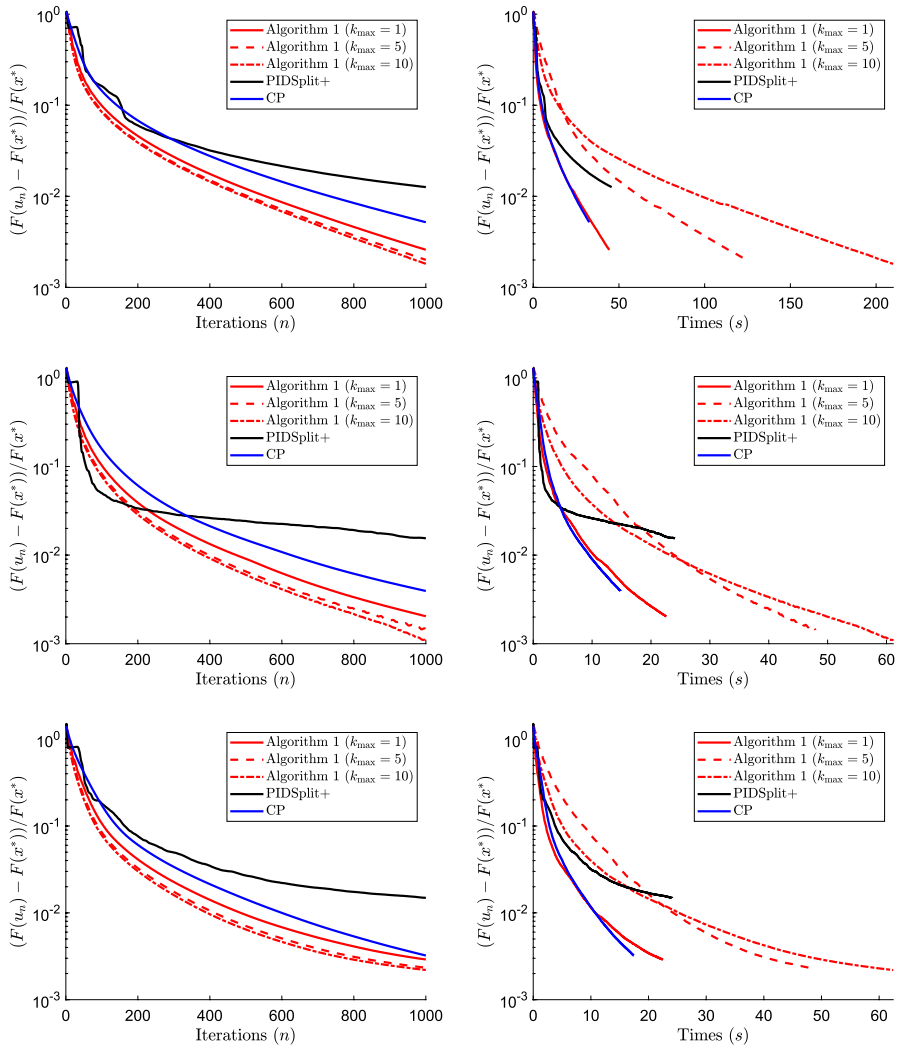
Fig. 11 Test problems with out-of-focus blur, SNR = 66 dB. From top to bottom: test images `phantom`, `mri`, `micro`. Relative decrease of the function values with respect to the iteration number (left) and time (right)
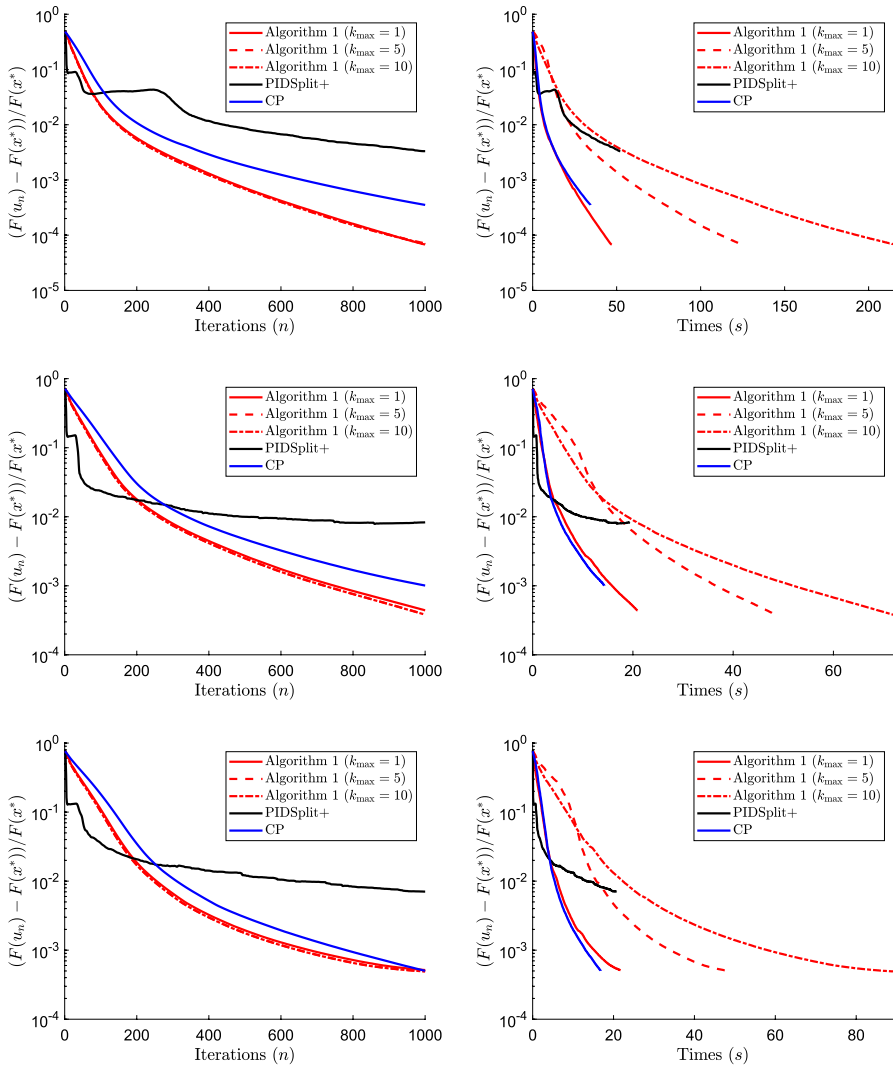
Fig. 12 Test problems with out-of-focus blur, SNR = 72 dB. From top to bottom: test images `phantom`, `mri`, `micro`. Relative decrease of the function values with respect to the iteration number (left) and time (right)

Table 2 SSIM values obtained after running the algorithms for 1000 iterations on each test problem

Blur	SNR	λ	Alg. 1 $k_{max} = 1$	Alg. 1 $k_{max} = 5$	Alg. 1 $k_{max} = 10$	PID Split+	CP
			phantom				
Gaussian	66	0.01	0.882	0.885	0.887	0.926	0.878
	72	0.05	0.812	0.813	0.814	0.784	0.810
Out-of-focus	66	0.05	0.958	0.959	0.959	0.955	0.955
	72	0.1	0.884	0.884	0.884	0.879	0.882
			mri				
Gaussian	66	0.01	0.927	0.927	0.927	0.911	0.928
	72	0.1	0.886	0.885	0.885	0.887	0.888
Out-of-focus	66	0.05	0.924	0.923	0.923	0.917	0.926
	72	0.1	0.903	0.903	0.903	0.9	0.905
			micro				
Gaussian	66	0.05	0.953	0.952	0.951	0.939	0.954
	72	0.1	0.905	0.904	0.903	0.890	0.909
Out-of-focus	66	0.05	0.963	0.962	0.961	0.955	0.965
	72	0.1	0.912	0.911	0.911	0.905	0.915

provided by Algorithm 1 seems to be not particularly sensitive to the choice of k_{max}, although a much wider experimentation is needed to confirm this remark.

5 Conclusions

We have presented a nested primal–dual algorithm with an inertial step on the primal variable for solving composite convex optimization problems. At each outer iteration, the proposed method starts the inner loop with the outcome of the previous one, employing the so-called "warm-start" strategy, and then performs a prefixed number of primal–dual iterations for inexactly computing the proximal–gradient point. We have proved the convergence of the primal iterates towards a minimum point, as well as an $O(1/n)$ convergence rate for the ergodic sequences associated to the primal and dual iterates, under the assumption that the inertial parameter satisfies an appropriate (implementable) condition. Numerical experiments on some Poisson image deblurring problems show that the proposed inertial primal–dual method preserves the typical acceleration effect of FISTA-like algorithms in the first iterations, while ensuring convergence to the true minimum value as the iterations proceed. Possible future developments include the introduction of linesearches and variable steplengths and metrics inside the proposed scheme, as well as the study of a suitable practical rule for choosing the "optimal" number k_{max} of inner primal–dual iterations.

Data availability statement The test image `micro` is included in the published article [46], available at`https://ieeexplore.ieee.org/document/1199635`The test images `moon`, `phantom`, `mri` are part of the MATLAB Image Processing Toolbox, which can be downloaded from`https://it.mathworks.com/products/image.html`

Declarations

Conflict of interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Acar, R., Vogel, C.R.: Analysis of bounded variation penalty methods for ill-posed problems. Inverse Probl. **10**(6), 1217–1229 (1994)
2. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. SIAM J. Optim. **26**(3), 1824–1834 (2016)
3. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Structured sparsity through convex optimization. Stat. Sci. **27**(4), 450–468 (2012)
4. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Processing **18**(11), 2419–34 (2009)
5. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
6. Bertero, M., Boccacci, P., Ruggiero, V.: *Inverse Imaging with Poisson Data*. IOP Publish., (2018)
7. Bertsekas, D.: Nonlinear Programming. Athena Scientific, Nashua, NH (1999)
8. Bonettini, S., Loris, I., Porta, F., Prato, M.: Variable metric inexact line-search based methods for nonsmooth optimization. SIAM J. Optim. **26**(2), 891–921 (2016)
9. Bonettini, S., Rebegoldi, S., Ruggiero, V.: Inertial variable metric techniques for the inexact forward-backward algorithm. SIAM J. Sci. Comput. **40**(5), A3180–A3210 (2018)
10. Bottou, L., Curtis, F.C., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)
11. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. **3**, 492–526 (2010)
12. Burger, M., Müller, J., Papoutsellis, E., Schönlieb, C.-B.: Total variation regularisation in measurement and image space for pet reconstruction. Inverse Probl. **30**, 105003 (2014)
13. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**, 89–97 (2004)
14. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm". J. Optim. Theory Appl. **166**(3), 968–982 (2015)
15. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**, 120–145 (2011)

16. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. **159**, 253–287 (2016)
17. Chen, J., Loris, I.: On starting and stopping criteria for nested primal-dual iterations. Numer. Algorithms **82**, 605–621 (2019)
18. Cloquet, C., Loris, I., Verhoeven, C., Defrise, M.: GISTA reconstructs faster with a restart strategy and even faster with a FISTA-like reconstruction. In *Proceedings of the 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference*, pp. 2334–2338, (2012)
19. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (Eds.) Fixed-point algorithms for inverse problems in science and engineering. Springer Optimization and Its Applications, pp. 185–212. Springer, New York NY (2011)
20. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
21. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
22. di Serafino, D., Landi, G., Viola, M.: Directional TGV-based image restoration under Poisson noise. J. Imaging **7**(6), 99 (2021)
23. Drori, Y., Sabach, S., Teboulle, M.: A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. Oper. Res. Lett. **43**(2), 209–214 (2015)
24. Figueiredo, M.A.T., Bioucas-Dias, J.M.: Restoration of Poissonian images using alternating direction optimization. IEEE Trans. Image Process. **19**(12), 3133–3145 (2010)
25. Kongskov, R.D., Dong, Y.: Directional total generalized variation regularization for impulse noise removal. In: F. Lauze, Y. Dong, and A. Dahl, (Eds.) Scale Space and Variational Methods in Computer Vision, pp. 221–231, (2017)
26. Kongskov, R.D., Dong, Y., Knudsen, K.: Directional total generalized variation regularization. BIT Numer. Math. **59**, 903–928 (2019)
27. Lazzaretti, M., Calatroni, L., Estatico, C.: Weighted-CEL0 sparse regularization for molecule localisation in super-resolution microscopy with poisson data. In *2021 International Symposium on Biomedical Imaging*, pp. 1751–1754, (2021)
28. Lorenz, D., Pock, T.: An inertial forward-backward algorithm for monotone inclusions. J. Math. Imaging Vis. **51**, 311–325 (2015)
29. Loris, I., Verhoeven, C.: On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. Inverse Probl. **27**, 125007 (2011)
30. Ma, S., Goldfarb, D., Chen, L.: Fixed point and Bregman iterative methods for matrix rank minimization. Math. Program. **128**, 321–353 (2011)
31. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France **93**, 273–299 (1965)
32. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Dokl. **269**, 543–547 (1983)
33. Nesterov, Y.: Introductory lectures on convex optimization: a basic course. Applied optimization. Kluwer Academic Publ, Boston, Dordrecht, London (2004)
34. Polyak, B.: Introduction to Optimization. Optimization Software - Inc., Publication Division, N.Y. (1987)
35. Rasch, J., Chambolle, A.: Inexact first-order primal-dual algorithms. Comput. Optim. Appl. **76**, 381–430 (2020)
36. Rebegoldi, S., Calatroni, L.: Scaled, inexact and adaptive generalized FISTA for strongly convex optimization. arXiv:2101.03915, (2021)
37. Rockafellar, R.T.: Convex Analysis. Princeton University Press, UK (1970)
38. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. J. Phys. D. **60**(1–4), 259–268 (1992)
39. Salzo, S., Villa, S.: Inexact and accelerated proximal point algorithms. J. Convex Anal. **19**(4), 1167–1192 (2012)
40. Sawatzky, A.: (Nonlocal) Total Variation in medical imaging. *Ph.D. Thesis, University of Münster*, (2011)
41. Schmidt, M., Le Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. arXiv:1109.2415v2, (2011)
42. Setzer, S., Steidl, G., Teuber, T.: Deblurring Poissonian images by split Bregman techniques. J. Vis. Commun. Image R. **21**, 193–199 (2010)

43. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward-backward algorithms. SIAM J. Optim. **23**(3), 1607–1633 (2013)
44. Vu, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**(3), 667–681 (2013)
45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**, 600–612 (2004)
46. Willett, R.M., Nowak, R.D.: Platelets: A multiscale approach for recovering edges and surfaces in photon limited medical imaging. IEEE Trans. Med. Imaging **22**, 332–350 (2003)
47. Zanni, L., Benfenati, A., Bertero, M., Ruggiero, V.: Numerical methods for parameter estimation in Poisson data inversion. J. Math. Imaging Vis. **52**(3), 397–413 (2015)