A variable metric forward–backward method with extrapolation*

S. Bonettini, F. Porta, V. Ruggiero

August 14, 2018

Abstract

Forward-backward methods are a very useful tool for the minimization of a functional given by the sum of a differentiable term and a nondifferentiable one and their investigation has experienced several efforts from many researchers in the last decade. In this paper we focus on the convex case and, inspired by recent approaches for accelerating first-order iterative schemes, we develop a scaled inertial forward-backward algorithm which is based on a metric changing at each iteration and on a suitable extrapolation step. Unlike standard forward-backward methods with extrapolation, our scheme is able to handle functions whose domain is not the entire space. Both an $\mathcal{O}(1/k^2)$ convergence rate estimate on the objective function values and the convergence of the sequence of the iterates are proved. Numerical experiments on several test problems arising from image processing, compressed sensing and statistical inference show the effectiveness of the proposed method in comparison to well performing state-of-the-art algorithms.

1 Introduction

In this paper we are interested in solving the optimization problem

$$\min_{x\in\mathbb{R}^n} F(x) \equiv f(x) + g(x) \tag{1}$$

where f and g are proper, convex and lower semicontinuous functions from $\mathbb{R}^n$ to $\mathbb{R}\cup\{\infty\}$. Moreover, we assume that f is differentiable with Lipschitz continuous gradient on a suitable closed, convex set $Y \subseteq \mathrm{dom}(f) = \{x \in \mathbb{R}^n : f(x) < \infty\}$, such that

$$\mathrm{dom}(f) \supseteq Y \supseteq \mathrm{dom}(g)$$

We also suppose that g is bounded from below over its domain and problem (1) admits at least a solution. Formulation (1) includes also constrained problems over a closed convex set $\Omega \subset \mathbb{R}^n$ where f has Lipschitz continuous gradient: indeed, the constraints defined by Ω can be inserted into the model by adding to g the indicator function of the feasible set itself, i.e.

$$\min_{x\in\Omega} f(x) + g(x) = \min_{x\in\mathbb{R}^n} f(x) + g(x) + \iota_\Omega(x)$$

where

$$\iota_\Omega(x) = \begin{cases} 0 & \text{if } \ x \in \Omega \\ +\infty & \text{otherwise} \end{cases}$$

*This work has been partially supported by MIUR under the project FIRB - Futuro in Ricerca 2012, contract RBFR12M3AC. The Italian GNCS - INdAM is also acknowledged.

Problem (1) is relevant in various domains of applied science such as signal and image processing, statistical inference and machine learning. A typical feature of these applications is the large number of variables, which makes the class of first order methods very attractive. In this class, forward-backward methods [12, 15] are especially suited for problem (1), since they exploit the decomposition of the objective function in a differentiable term and a nondifferentiable one. The general forward-backward iteration is given by

$$x^{(k+1)} = x^{(k)} + \lambda_k(\mathrm{prox}_{\alpha_k g}(x^{(k)} - \alpha_k \nabla f(x^{(k)})) - x^{(k)}),$$

where λ_k, α_k are positive parameters controlling the steplength and $\mathrm{prox}_{\phi}(\cdot)$ is the proximity operator associated to the convex function ϕ, defined as

$$\mathrm{prox}_{\phi}(y) = \underset{x\in\mathbb{R}^n}{\mathrm{argmin}} \;\; \phi(x) + \frac{1}{2}\|x-y\|^2$$

Forward-backward methods are easy to implement and have well studied convergence properties. On the other hand, it is well known that they can exhibit a poor convergence rate, especially when a high accuracy is required.

In the recent literature, we can find two different approaches aiming to improve the convergence speed of forward-backward methods. They are both described below.

Inertial/estrapolation techniques. This approach consists in adding an extrapolation step to the basic forward-backward iteration, yielding a multistep algorithm, called also heavy ball or inertial method [26, p.65]. The idea of inertial methods became very popular in the last decade, in view of Nesterov's work [24] and it has been further developed in [3], where the authors propose the following variant

$$y^{(k)} = x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}) \quad (2)$$

$$x^{(k+1)} = \mathrm{prox}_{\alpha_k g}(y^{(k)} - \alpha_k \nabla f(y^{(k)})). \quad (3)$$

In [3, 6], the convergence of method (2)–(3) is investigated by showing that for suitable sequences of parameters $\{\alpha_k\}$ and $\{\beta_k\}$ (with $\lim_k \beta_k = 1$) one has $F(x^{(k)}) - F^* = \mathcal{O}\left(\frac{1}{k^2}\right)$, where F^* is the optimal value of the objective function. Recently, under additional assumptions on the sequences $\{\alpha_k\}$ and $\{\beta_k\}$, Chambolle and Dossal in [10] proved the convergence of the iterates $\{x^{(k)}\}$, while in [29] the authors propose a variant of the inertial scheme where the proximal point (3) can be computed inexactly. We also mention the recent work [25] where inertial forward-backward algorithms are analyzed when $f(x)$ is not convex.

A drawback in the use of method (2)–(3) is that it may be unfeasible when $\mathrm{dom}(f)$ does not coincide with the whole space $\mathbb{R}^n$, since the point $y^{(k)}$ computed in (2) does not necessarily belong to $\mathrm{dom}(f)$.

Variable metric/scaling techniques. In a variable metric forward-backward algorithm, the underlaying metric may change at each iteration by means of suitable symmetric positive definite scaling matrices multiplying the gradient of f and also involved in the definition of the proximity operator. The expected advantage in using a variable metric is an improved capability to capture the local features of problem (1), possibly leading to an improvement of the convergence speed

(think for example to the Newton's method). In [13, 14], the authors propose and analyze the following variable metric forward-backward algorithm

$$x^{(k+1)} = x^{(k)} + \lambda_k(\text{prox}_{\alpha_k g}^{D_k}(x^{(k)} - \alpha_k D_k^{-1}\nabla f(x^{(k)})) - x^{(k)}) \tag{4}$$

where $\{D_k\}$ is a user supplied sequence of symmetric positive definite matrices and $\text{prox}_{\alpha_k g}^{D_k}(y)$ is defined as

$$\text{prox}_{\alpha_k g}^{D_k}(y) = \underset{x\in\mathbb{R}^n}{\text{argmin}}\ \ g(x) + \frac{1}{2\alpha_k}(x-y)^T D_k(x-y) \tag{5}$$

The authors also devise conditions on the sequence $\{D_k\}$ ensuring the convergence of $\{x^{(k)}\}$. Method (4), equipped by an Armijo line–search for the computation of λ_k, has been extensively studied in the papers [7, 8, 9] for constrained minimization, when $g(x)$ reduces to the indicator function of a closed convex subset of $\mathbb{R}^n$. The convergence rate on the objective function values in this case is only linear, i.e. $F(x^{(k)}) - F^* = \mathcal{O}\left(\frac{1}{k}\right)$ (see [3, 8]). However, in spite of the theoretical convergence rate, a suitable combination of the stepsize parameter α_k and the scaling matrix D_k can allow method (4) to reach practical performances which are comparable with those of (2)–(3) [8, 27].

Main contribution. In this paper we propose an algorithm combining the two acceleration techniques described above, designing an original variable metric forward-backward method with extrapolation.

In particular, we address the case where $\text{dom}(f)$ is a proper subset of $\mathbb{R}^n$ and we devise suitable conditions on the stepsize parameters and on the scaling matrices sequence to ensure both the convergence of the iterates sequence $\{x^{(k)}\}$ and the $\mathcal{O}\left(\frac{1}{k^2}\right)$ rate for the objective function values.

The effectiveness of the proposed method is evaluated by means of a comparison with other state-of-the-art algorithms, on several optimization problems of the form (1), arising from different real-life applications such as image deblurring, compressed sensing and probability density estimation.

The plan of the paper is the following. In Section 2 we collect some definitions and introductory results. Section 3 is devoted to the description of the proposed algorithm while the convergence rate analysis is performed in Section 3.1 and the convergence of the iterates to a minimizer of the optimization problem is proved in Section 3.2. This last section is strongly inspired from a recent paper by Chambolle and Dossal [10]. Section 4 deals with the results of the numerical experiments we performed on some test problems arising in image and signal processing and statistical inference. Finally, our conclusions are given in Section 5.

2 Notation, definitions and basic results

We denote by $\|\cdot\|$ the Euclidean norm of a vector while $\|\cdot\|_D$ indicates the norm induced by the symmetric positive definite matrix D, i.e. $\|x\|_D^2 = x^T D x$. Furthermore, in the subspace $\mathcal{S}_n(\mathbb{R})$ of the symmetric real matrices of order n, we consider the following Loewner partial ordering

$$\forall D_1, D_2 \in \mathcal{S}_n(\mathbb{R}) \quad D_1 \succeq D_2 \Leftrightarrow x^T D_1 x \geq x^T D_2 x \quad \forall x \in \mathbb{R}^n$$

For any $\eta \in \mathbb{R}$, $\eta > 0$ we define the set $\mathcal{D}_\eta \subset \mathcal{S}_n(\mathbb{R})$ as the set of all positive definite matrices D such that $D \succeq \eta I$. Clearly, if $D \in \mathcal{D}_\eta$, the eigenvalues of D are lower bounded by η and for

each $u \in \mathbb{R}^n$, the following inequality holds

$$\eta \|u\|^2 \leq u^T D u = \|u\|_D^2 \tag{6}$$

The following lemma states a well known property of the projection operator, whose proof runs exactly as in [20, p.48]

LEMMA 2.1 *Let $\Omega \subseteq \mathbb{R}^n$ be a closed convex set and define the scaled Euclidean projection operator associated to $D \in \mathcal{D}_\eta$ as*

$$P_{\Omega,D}(x) = \operatorname*{argmin}_{y \in \Omega} \|y - x\|_D^2 \tag{7}$$

for any $x \in \mathbb{R}^n$. Then, the operator (7) is nonexpansive with respect to the norm induced by the matrix D, i.e.

$$\|P_{\Omega,D}(x) - P_{\Omega,D}(z)\|_D \leq \|x - z\|_D \tag{8}$$

for all $x, z \in \mathbb{R}^n$.

For every $x \in Y$ and $y \in \mathbb{R}^n$ we define

$$\ell(y; x) = f(x) + \nabla f(x)^T (y - x) \tag{9}$$

and

$$q(y; x) = \ell(y; x) + g(y) \tag{10}$$

DEFINITION 2.1 *A smooth function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ has L-Lipschitz continuous gradient on the set $\Omega \subseteq dom(f)$ if there exists $L > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \Omega.$$

We recall also the following result for smooth functions with L-Lipschitz continuous gradient.

LEMMA 2.2 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a continuously differentiable function with L-Lipschitz continuous gradient on $Y \subseteq \mathbb{R}^n$ and $D \in \mathcal{D}_\eta$. Then, for every $x, y \in Y$ we have*

$$f(y) \leq \ell(y; x) + \frac{1}{2\alpha}\|x - y\|_D^2 \tag{11}$$

for all $\alpha \leq \eta/L$.

Proof. From (6) we obtain

$$\ell(y; x) + \frac{1}{2\alpha}\|x - y\|_D^2 \geq \ell(y; x) + \frac{\eta}{2\alpha}\|x - y\|^2 \geq f(y)$$

where the rightmost inequality holds for $\frac{\alpha}{\eta} \leq 1/L$, thanks to Lemma 6.9.1 in [6] (see also [3, Lemma 2.1]). □

DEFINITION 2.2 *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Then, the subdifferential of g at $x \in \mathbb{R}^n$ is the set*

$$\partial g(x) = \{w \in \mathbb{R}^n : g(y) \geq g(x) + (y - x)^T w, \quad \forall y \in \mathbb{R}^n\}$$

A point x is a minimizer of g if and only if $0 \in \partial g(x)$.

As a consequence of the previous definition, we have that $x \in \mathbb{R}^n$ is a solution of problem (1) if and only if $-\nabla f(x) \in \partial g(x)$.

Given a positive number α and a matrix $D \in \mathcal{D}_\eta$, we now define the following function

$$Q_{\alpha,D}(y; x) = q(y; x) + \frac{1}{2\alpha}\|x - y\|_D^2 \tag{12}$$

for $y \in \mathbb{R}^n$, $x \in Y$. The function $Q_{\alpha,D}(\cdot; x)$ admits a unique minimizer, which will be denoted by

$$p_{\alpha,D}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} Q_{\alpha,D}(y; x) \tag{13}$$

Clearly, the point $p_{\alpha,D}(x)$ belongs to $\mathrm{dom}(g)$ and, when $p_{\alpha,D}(x) = x$, x is a minimizer of F. A simple computation shows that

$$p_{\alpha,D}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} g(y) + \frac{1}{2\alpha}\left\|y - x + \alpha D^{-1}\nabla f(x)\right\|_D^2 \tag{14}$$

where it is more evident that the introduction of a matrix D in (12) induces a scaling of $\nabla f(y)$ by D^{-1}. Clearly, according to (5), we have the equivalence $p_{\alpha,D}(x) = \mathrm{prox}_{\alpha g}^D(x - \alpha D^{-1}\nabla f(x))$. In the next lemmas, two useful properties of the operator (13) are proved.

LEMMA 2.3 *Let $f : Y \to \mathbb{R}$ be a continuously differentiable function with L-Lipschitz continuous gradient, $D \in \mathcal{D}_\eta$ and g be a convex function. Let $x \in Y$ and $y = p_{\alpha,D}(x)$. Then, for any $z \in \mathbb{R}^n$, we have*

$$q(y; x) + \frac{1}{2\alpha}\|x - y\|_D^2 \leq q(z; x) + \frac{1}{2\alpha}\|z - x\|_D^2 - \frac{1}{2\alpha}\|z - y\|_D^2 \tag{15}$$

Proof. From the optimality conditions of the problem (13), we have that there exists a vector $w \in \partial g(y)$ such that

$$\nabla f(x) + \frac{1}{\alpha}D(y - x) + w = 0 \tag{16}$$

Since $w \in \partial g(y)$, for all $z \in \mathbb{R}^n$ we have that $g(z) - g(y) \geq w^T(z - y)$, which, together with (16), implies

$$g(z) - g(y) \geq \left(\nabla f(x) + \frac{1}{\alpha}D(y - x)\right)^T (y - z) \tag{17}$$

From the definition of $\ell(y;x)$ in (9), we can write

$$\begin{aligned}
\ell(y;x)+\frac{1}{2\alpha}\|y-x\|_D^2 &= f(x)+\nabla f(x)^T(y-z+z-x)+\frac{1}{2\alpha}\|y-z+z-x\|_D^2 \\
&= \ell(z;x)+\frac{1}{2\alpha}\|z-x\|_D^2+\frac{1}{2\alpha}\|y-z\|_D^2+ \\
&\quad +\nabla f(x)^T(y-z)+\frac{1}{\alpha}(z-x)^TD(y-z) \\
&= \ell(z;x)+\frac{1}{2\alpha}\|z-x\|_D^2-\frac{1}{2\alpha}\|y-z\|_D^2+ \\
&\quad +\left(\nabla f(x)+\frac{1}{\alpha}D(y-x)\right)^T(y-z) \\
&\leq \ell(z;x)+\frac{1}{2\alpha}\|z-x\|_D^2-\frac{1}{2\alpha}\|y-z\|_D^2+g(z)-g(y)
\end{aligned}$$

where the third equality is obtained by adding and subtracting $\frac{1}{\alpha}y^TD(y-z)$ and the final inequality is a consequence of (17). Finally, inequality (15) follows by rearranging terms and recalling the definition (10). □

A direct consequence of the previous lemma is the following result.

LEMMA 2.4 *Let $f:Y\to\mathbb{R}$ be a convex, continuously differentiable function with L-Lipschitz continuous gradient and g be a convex function. Let $F(x)\equiv f(x)+g(x)$, $D\in\mathcal{D}_\eta$ and $y=p_{\alpha,D}(x)$, $x\in Y$. If α is such that y satisfies the condition* (11)*, then, for any $z\in dom(f)$, we have*

$$F(y)+\frac{1}{2\alpha}\|z-y\|_D^2\leq F(z)+\frac{1}{2\alpha}\|z-x\|_D^2 \tag{18}$$

Proof. Since f is convex, the following inequality holds:

$$F(z)\geq q(z;x)\quad\forall x,z$$

Therefore, from Lemma 2.3 and from (11) we have the result. □

3 A variable metric inertial forward-backward method with backtracking

In this section we describe and analyze the proposed method, whose generic scheme is detailed in Algorithm 1. It consists in a variable metric forward–backward iteration (Step 4) combined with an extrapolation–projection step (Step 1).

The steplength α_k is adaptively computed via a backtracking procedure, while suitable choices for the extrapolation parameter β_k and the scaling matrix D_k will be described during the convergence analysis in sections 3.1 and 3.2.

Algorithm 1 can be considered a generalization of the Fast Iterative Soft Tresholding Algorithm (FISTA, [3]), whose iteration has the form (2)–(3). The main novelties we introduce with respect to FISTA are the possibility to employ at each iteration the variable metric induced by the matrix D_k and the projection of the extrapolated point $x^{(k)}+\beta_k(x^{(k)}-x^{(k-1)})$, which

Algorithm 1 Scaled inertial forward-backward method with backtracking

Choose $\alpha_0 > 0$, $\delta < 1$, $x^{(0)} \in Y$. Set $x^{(-1)} = x^{(0)}$ and define a sequence of nonnegative numbers $\{\beta_k\}$ and a sequence of matrices $\{D_k\}$, with $D_k \in \mathcal{D}_\eta$, such that $\gamma = \sup_{k\in\mathbb{N}} \|D_k\| < \infty$.

FOR $k = 0, 1, 2, ...$

STEP 1. Extrapolation: $y^{(k)} = P_{Y,D_k}(x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}))$

STEP 2. Set $\alpha_k = \alpha_{k-1}$, $i_k = 0$

STEP 3. Set $x_+^{(k)} = p_{\alpha_k,D_k}(y^{(k)})$

STEP 4. If

$$f(x_+^{(k)}) \leq \ell(x_+^{(k)}; y^{(k)}) + \frac{1}{2\alpha_k}\|y^{(k)} - x_+^{(k)}\|_{D_k}^2$$

go to Step 5.
else set

$$i_k \leftarrow i_k + 1 \qquad \alpha_k = \delta^{i_k}\alpha_{k-1}$$

and go to Step 3.

STEP 5. Set the new iterate $x^{(k+1)} = x_+^{(k)}$

END

allows to handle problems where $\mathrm{dom}(f) \supseteq Y$ does not coincide with the entire space $\mathbb{R}^n$. When $Y = \mathbb{R}^n$, FISTA is recovered by setting $D_k = I$ for all $k \geq 0$.

For convenience, we restate below our hypotheses on problem (1), which we assume to be fulfilled throughout this section.

(A1) $f, g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are proper, convex and lower semicontinuous;

(A2) f is differentiable with L-Lipschitz continuous gradient on $Y \subseteq \mathrm{dom}(f)$, Y is closed and convex and $\mathrm{dom}(f) \supseteq Y \supseteq \mathrm{dom}(g)$;

(A3) g is bounded from below over its domain;

(A4) problem (1) admits at least a solution.

Moreover, we will indicate hereafter by $\{x^{(k)}\}$ the sequence generated by Algorithm 1, while x^* will denote any of the solutions o (1).

First, we observe that, thanks to assumption (A2), Algorithm 1 is well defined, i.e. the backtracking loop between Steps 3 and 4 terminates in a finite number of steps. Indeed, from Lemma 2.2, observing that the sequence $\{\alpha_k\}$ is non–increasing and that the reducing factor is $\delta < 1$, we obtain the following inequalities

$$0 < \frac{\delta\eta}{L} \leq \alpha_k \leq \alpha_{k-1} \leq \alpha_0 \tag{19}$$

In particular, the backtracking condition implies that the new iterate $x^{(k+1)}$ satisfies

$$f(x^{(k+1)}) \leq \ell(x^{(k+1)}; y^{(k)}) + \frac{1}{2\alpha_k}\|y^{(k)} - x^{(k+1)}\|_{D_k}^2 \tag{20}$$

In the following sections we will show that Algorithm 1 with a proper parameters setting has a $\mathcal{O}(1/k^2)$ convergence rate with respect to the objective function values, i.e.

$$F(x^{(k)}) - F(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

which is the same as FISTA, and, moreover, the iterate sequence $\{x^{(k)}\}$ converges to a minimizer of (1).

3.1 Convergence rate analysis

In the rest of the paper we will assume that the extrapolation parameter β_k has the form

$$\beta_k = \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}} \tag{21}$$

where $\{\theta_k\} \subset (0,1]$ is a given sequence of parameters. Moreover, we will adopt the following notation

$$\begin{aligned} v_k &= F(x^{(k)}) - F(x^*) \\ z^{(k)} &= x^{(k)} + \frac{1-\theta_{k-1}}{\theta_{k-1}}(x^{(k)} - x^{(k-1)}) = x^{(k-1)} + \frac{1}{\theta_{k-1}}(x^{(k)} - x^{(k-1)}) \\ u^{(k)} &= z^{(k)} - x^* \\ t_k &= \frac{1}{\theta_k} \end{aligned} \tag{22}$$

Before giving the main result, we need to prove some technical lemmas. The first of them establishes a key inequality which is crucial for the subsequent analysis and it is analogous to Lemma 4.1 in [3].

LEMMA 3.1 *Let $\{D_k\} \subset \mathcal{D}_\eta$ be a sequence of scaling matrices and assume that $\{\theta_k\}$ satisfies*

$$\frac{1-\theta_k}{\theta_k^2} \leq \frac{1}{\theta_{k-1}^2} \qquad 0 < \theta_k \leq 1 \tag{23}$$

Then, we have

$$2\alpha_{k+1}t_k^2 v_{k+1} + \|u^{(k+1)}\|_{D_k}^2 \leq 2\alpha_k t_{k-1}^2 v_k + \|u^{(k)}\|_{D_k}^2 \tag{24}$$

Proof. Let us define the point $y^* = (1-\theta_k)x^{(k)} + \theta_k x^*$. We have $y^* \in \mathrm{dom}(g)$. From (18) in Lemma 2.4 with $y = x^{(k+1)}$, $x = y^{(k)}$ and $z = y^*$, we have

$$\begin{aligned} F(x^{(k+1)}) + \frac{1}{2\alpha_k}\|y^* - x^{(k+1)}\|_{D_k}^2 &\leq F(y^*) + \frac{1}{2\alpha_k}\|y^* - y^{(k)}\|_{D_k}^2 \\ &\leq (1-\theta_k)F(x^{(k)}) + \theta_k F(x^*) + \frac{1}{2\alpha_k}\|y^* - y^{(k)}\|_{D_k}^2 \\ &\leq (1-\theta_k)F(x^{(k)}) + \theta_k F(x^*) + \frac{1}{2\alpha_k}\|y^* - (x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}))\|_{D_k}^2 \end{aligned}$$

where the first inequality follows from the definition of y^* and the convexity of F and the second one from (8). From (21) and the definition of y^*, we obtain

$$F(x^{(k+1)}) + \frac{1}{2\alpha_k}\|(1-\theta_k)x^{(k)} + \theta_k x^* - x^{(k+1)}\|^2_{D_k} \leq (1-\theta_k)F(x^{(k)}) + \theta_k F(x^*) +$$
$$+\frac{1}{2\alpha_k}\left\|(1-\theta_k)x^{(k)} + \theta_k x^* - \left(x^{(k)} + \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}(x^{(k)} - x^{(k-1)})\right)\right\|^2_{D_k}$$

Rearranging terms, we have

$$F(x^{(k+1)}) + \frac{\theta_k^2}{2\alpha_k}\|z^{(k+1)} - x^*\|^2_{D_k} \leq (1-\theta_k)F(x^{(k)}) + \theta_k F(x^*) + \frac{\theta_k^2}{2\alpha_k}\|z^{(k)} - x^*\|^2_{D_k}$$

Subtracting $F(x^*)$ from both sides leads to

$$v_{k+1} + \frac{\theta_k^2}{2\alpha_k}\|z^{(k+1)} - x^*\|^2_{D_k} \quad \leq \quad (1-\theta_k)v_k + \frac{\theta_k^2}{2\alpha_k}\|z^{(k)} - x^*\|^2_{D_k}$$

Multiplying both sides by $2\alpha_k/\theta_k^2$ and rearranging terms gives

$$2\frac{\alpha_k}{\theta_k^2}v_{k+1} + \|z^{(k+1)} - x^*\|^2_{D_k} \leq 2\alpha_k\frac{1-\theta_k}{\theta_k^2}v_k + \|z^{(k)} - x^*\|^2_{D_k} \tag{25}$$

Finally, observing that $\alpha_{k+1} \leq \alpha_k$, we obtain

$$2\frac{\alpha_{k+1}}{\theta_k^2}v_{k+1} + \|z^{(k+1)} - x^*\|^2_{D_k} \leq 2\alpha_k\frac{1-\theta_k}{\theta_k^2}v_k + \|z^{(k)} - x^*\|^2_{D_k} \tag{26}$$

In view of (23), we obtain (24). □

An example of sequence $\{\theta_k\}$ and corresponding $\{\beta_k\}$ satisfying (21)-(23) is the following one

$$\theta_k = \begin{cases} 1 & k=0 \\ \frac{a}{k+a} & k \geq 1 \end{cases} \qquad \beta_k = \begin{cases} 0 & k=0 \\ \frac{k-1}{k+a} & k \geq 1 \end{cases} \tag{27}$$

with $a \geq 2$. Indeed, since $\theta_k = \frac{1}{t_k}$, condition (23) writes also as

$$t_{k-1}^2 + t_k - t_k^2 \geq 0$$

Since (27) implies $t_k = \frac{k+a}{a}$, for all $k \geq 0$, $a \geq 2$ we have

$$\begin{aligned} t_{k-1}^2 + t_k - t_k^2 &= \frac{(k-1+a)^2}{a^2} + \frac{(k+a)}{a} - \frac{(k+a)^2}{a^2} \\ &= \frac{(k+a)(a-2)+1}{a^2} \geq 0 \end{aligned}$$

The choice $a = 2$ has been proposed in [3] for computing FISTA's extrapolation parameters, while the more general case $a \geq 2$ is considered in [10].

Our aim is now to show that the sequence $\{v_k\}$ is bounded. To this end, we recall the following lemma on summable nonnegative sequences.

LEMMA 3.2 *[26] Let $\{a_k\}$, $\{\zeta_k\}$ and $\{\xi_k\}$ be nonnegative sequences of real numbers such that $a_{k+1} \leq (1+\zeta_k)a_k + \xi_k$ and $\sum_{k=0}^{\infty} \zeta_k < \infty$, $\sum_{k=0}^{\infty} \xi_k < \infty$. Then, $\{a_k\}$ converges.*

In the next lemma, we introduce a crucial assumption on the sequence of matrices $\{D_k\}$ (see also [13, 14]).

LEMMA 3.3 *Let $\{\theta_k\}$ satisfy (23) and define $a_k = 2\alpha_k t_{k-1}^2 v_k + \|u^{(k)}\|_{D_k}^2$. Assume that the sequence of matrices $\{D_k\} \subset \mathcal{D}_\eta$, satisfies*

$$D_{k+1} \preceq (1+\eta_k)D_k \quad \forall k \geq 0 \quad \text{with } \eta_k \in \mathbb{R}, \eta_k \geq 0 \text{ such that } \sum_{k=0}^{\infty} \eta_k < \infty \tag{28}$$

Then, $\{a_k\}$ is a convergent sequence.

Proof. Setting s_k as

$$s_k = 2\alpha_k t_{k-1}^2 v_k \tag{29}$$

in view of (28), we obtain

$$\begin{aligned} a_{k+1} = s_{k+1} + \|u^{(k+1)}\|_{D_{k+1}}^2 &\leq s_{k+1} + (1+\eta_k)\|u^{(k+1)}\|_{D_k}^2 \\ &\leq (1+\eta_k)(s_{k+1} + \|u^{(k+1)}\|_{D_k}^2) \\ &\leq (1+\eta_k)(s_k + \|u^{(k)}\|_{D_k}^2) \\ &= (1+\eta_k)a_k \end{aligned}$$

where the second inequality follows from the fact that $\eta_k \geq 0$ for any $k \geq 0$ and the third one from inequality (24). Thus, Lemma 3.2 implies that $\{a_k\}$ converges. □

We are now ready to give the main result of this section, establishing the convergence rate of Algorithm 1.

THEOREM 3.1 *Let $\{D_k\} \subset \mathcal{D}_\eta$ be a sequence of matrices satisfying (28) and assume that $\{\theta_k\}$, $\{\beta_k\}$ are chosen as in (27) with $a \geq 2$. Then, there exists a constant C such that*

$$F(x^{(k)}) - F(x^*) \leq \frac{C}{(k-1+a)^2} \tag{30}$$

for all $k \geq 1$.

Proof. Lemma 3.3 guarantees in particular that there exists a constant $K > 0$ such that $a_k = 2\alpha_k t_{k-1}^2 v_k + \|u^{(k)}\|_{D_k}^2 \leq K$. Since $2\alpha_k t_{k-1}^2 v_k \leq a_k$, we also have $2\alpha_k t_{k-1}^2 v_k \leq K$. Formula (27) implies that $t_{k-1}^2 = \frac{(k-1+a)^2}{a^2}$. Thus, recalling the definition of v_k and the lower bound in (19) for the parameter α_k, we can write

$$v_k = F(x^{(k)}) - F(x^*) \leq \frac{a^2 LK}{2\eta\delta(k-1+a)^2}$$

obtaining (30) with $C = \frac{a^2 LK}{2\eta\delta}$. □

Theorem 3.1 is a generalization of Theorem 4.4 in [3], which is recovered when $D_k = I$, $a = 2$, $Y = \mathbb{R}^n$. An analogous result for the case $D_k = I$, $a \geq 2$ can be found in [10].

It can be observed that the optimal value of the constant C in (30) is obtained with $a = 2$. However, as pointed out in [10] and as we will see in the following section, selecting $a > 2$ allows to prove the convergence of the iterates $\{x^{(k)}\}$ to a solution of (1).

Remark When the sequence of matrices $\{D_k\}$ satisfies the assumption (28) and, in addition, $\sup_{k\in\mathbb{N}}\|D_k\| = \gamma < \infty$, then there exists a matrix $\mathfrak{D} \in \mathcal{D}_\eta$ such that $D_k \to \mathfrak{D}$ pointwise [13, Lemma 2.3]. A similar result holds also for a sequence of matrices $\{D_k\} \subset \mathcal{D}_\eta$, such that $\sup_{k\in\mathbb{N}}\|D_k\| = \gamma < \infty$, satisfying the following condition

$$D_k \preceq (1+\nu_k)D_{k+1} \quad \forall k \geq 0 \quad \text{with } \nu_k \in \mathbb{R},\ \nu_k \geq 0 \text{ such that } \sum_{k=0}^{\infty} \nu_k < \infty \tag{31}$$

A sufficient condition ensuring both (28) and (31) is the following one

$$\frac{1}{\gamma_k} \leq \|D_k\| \leq \gamma_k \quad \gamma_k^2 = 1+\zeta_k \quad \text{where} \quad \zeta_k \geq 0 \quad \text{and} \quad \sum_{k=0}^{\infty} \zeta_k < \infty \tag{32}$$

with $\gamma_k < \gamma$, $\gamma > 1$. Indeed, in this case $\eta = \frac{1}{\gamma}$ and for any $x \in \mathbb{R}^n$ we have

$$\begin{aligned} x^T D_{k+1} x &\leq \frac{\gamma_{k+1}\gamma_k}{\gamma_k}\|x\|^2 \leq \gamma_{k+1}\gamma_k x^T D_k x \\ x^T D_k x &\leq \frac{\gamma_k\gamma_{k+1}}{\gamma_{k+1}}\|x\|^2 \leq \gamma_k\gamma_{k+1} x^T D_{k+1} x \end{aligned}$$

Let us define $\eta_k = \nu_k = \gamma_{k+1}\gamma_k - 1 = \sqrt{(1+\zeta_{k+1})(1+\zeta_k)} - 1$ and observe that the series $\sum \eta_k$ and $\sum \zeta_k$ have the same behavior, since the known limit $\lim_{z\to 0}(\sqrt{1+z}-1)/z = 1/2$. Therefore, for any $x \in \mathbb{R}^n$, we can write

$$\begin{aligned} x^T D_{k+1} x &\leq (1+\eta_k) x^T D_k x \\ x^T D_k x &\leq (1+\nu_k) x^T D_{k+1} x \end{aligned}$$

with $\eta_k = \nu_k$ for any $k \geq 0$ and $\sum \eta_k < \infty$; thus a sequence of matrices chosen according to the rule (32) satisfies both the assumptions (28) and (31).

For the sequence $\{D_k\}$ satisfying (32) and $\{\theta_k\}$ as in (27) with $a = 2$, the convergence rate estimate established in Theorem 3.1 becomes

$$F(x^{(k)}) - F(x^*) \leq \frac{2L\gamma K}{\delta(k+1)^2}$$

which is very similar to that in [3].

3.2 Convergence of the iterates to a minimizer

In this section, borrowing the ideas in [10], we prove that the iterates $\{x^{(k)}\}$ converge to a minimizer of F, when the parameters sequences are chosen as in (27) with $a > 2$ and the scaling matrices sequence satisfies some additional assumption. Before giving the main result, we need to prove some technical lemmas.

Under the same assumptions of Theorem 3.1, we first prove the boundedness of the sequence $\{\|u^{(k)}\|\}$. To this end, we first recall an useful lemma on summable nonnegative sequences.

LEMMA 3.4 *[8] Let $\{\gamma_k\}$ be a sequence of positive numbers such that $\gamma_k^2 = 1+\eta_k$, $\eta_k \geq 0$, where $\sum_{k=0}^{\infty}\eta_k < \infty$. Let $\tau_k = \prod_{j=0}^{k}\gamma_j^2$ for any $k \geq 0$. Then the sequence $\{\tau_k\}$ is bounded.*

LEMMA 3.5 *Let $\{\theta_k\}$, $\{\beta_k\}$ be defined such that* (21) *and* (23) *hold and assume that the sequence $\{D_k\} \subset \mathcal{D}_\eta$ satisfies* (28)*. Then, the sequence $\{\|u^{(k)}\|\}$ is bounded, i.e. $\|u^{(k)}\| \leq U$ for $k \geq 0$.*

Recalling definition (29), we obtain

$$\begin{aligned}
\|u^{(k+1)}\|^2_{D_{k+1}} &\leq (1+\eta_k)\|u^{(k+1)}\|^2_{D_k} \\
&\leq (1+\eta_k)(s_k - s_{k+1} + \|u^{(k)}\|^2_{D_k}) \\
&\leq (1+\eta_k)(s_k + \|u^{(k)}\|^2_{D_k}) \\
&\leq (1+\eta_k)(s_k + (1+\eta_{k-1})\|u^{(k)}\|^2_{D_{k-1}}) \\
&\leq (1+\eta_k)(1+\eta_{k-1})\left(s_k + \left(s_{k-1} - s_k + \|u^{(k-1)}\|^2_{D_{k-1}}\right)\right) \\
&\leq (1+\eta_k)(1+\eta_{k-1})(s_{k-1} + (1+\eta_{k-2})\|u^{(k-1)}\|^2_{D_{k-2}}) \\
&\leq (1+\eta_k)(1+\eta_{k-1})(1+\eta_{k-2})(s_{k-2} + \|u^{(k-2)}\|^2_{D_{k-2}}) \\
&\leq \tau_k(s_0 + \|u^{(0)}\|^2_{D_0})
\end{aligned}$$

where τ_k is defined as in Lemma 3.4 and we repeatedly applied the inequalities (28) and (24) in the following ones, together with the fact that $\eta_i \geq 0$ for $i \geq 0$. Since τ_k is bounded, $\|u^{(k+1)}\|^2_{D_{k+1}}$ is bounded. Furthermore, from (6) we have the result. □

The following lemma generalizes the results in [10, Theorem 2].

LEMMA 3.6 *Let $\{D_k\} \subset \mathcal{D}_\eta$ be a sequence of matrices satisfying* (28)*, with $\sup_{k\in\mathbb{N}} \|D_k\| = \gamma < \infty$. Assume that $\{\theta_k\}$, $\{\beta_k\}$ are chosen as in* (27)*, with $a > 2$. Then the sequence $\{kv_k\}$ is summable.*

Proof. First we recall that with these settings for θ_k and β_k, (23) and (21) are satisfied. In view of $t_k^2 = \frac{1}{\theta_k^2}$, we can write the inequality (26) as follows:

$$\alpha_{k+1}t_k^2 v_{k+1} - \alpha_k(t_k^2 - t_k)v_k \leq \frac{\|u^{(k)}\|^2_{D_k}}{2} - \frac{\|u^{(k+1)}\|^2_{D_k}}{2}$$

Summing up from $k = 0, ..., K$, since $t_0 = 1$, we have

$$\begin{aligned}
\alpha_{K+1}t_K^2 v_{K+1} + \sum_{k=1}^{K} \alpha_k(t_{k-1}^2 - t_k^2 + t_k)v_k \leq \\
&\leq \frac{1}{2}\sum_{k=1}^{K} \|u^{(k)}\|^2_{D_k} - \|u^{(k)}\|^2_{D_{k-1}} + \frac{\|u^{(0)}\|^2_{D_0}}{2} - \frac{\|u^{(K+1)}\|^2_{D_K}}{2} \\
&\leq \frac{1}{2}\sum_{k=1}^{K} (1+\eta_{k-1})\|u^{(k)}\|^2_{D_{k-1}} - \|u^{(k)}\|^2_{D_{k-1}} + \frac{\|u^{(0)}\|^2_{D_0}}{2} \\
&= \frac{1}{2}\sum_{k=1}^{K} \eta_{k-1}\|u^{(k)}\|^2_{D_{k-1}} + \frac{\|u^{(0)}\|^2_{D_0}}{2} \\
&\leq \frac{\gamma U^2}{2}\sum_{k=0}^{K-1} \eta_k + \frac{\|u^{(0)}\|^2_{D_0}}{2}
\end{aligned}$$

where the second inequality follows from (28) and the last one from the boundedness of $\|D_k\|$ and Lemma 3.5. Furthermore, using (19), we obtain

$$\sum_{k=1}^{K}(t_{k-1}^2 - t_k^2 + t_k)v_k \leq \frac{L\gamma U^2}{2\delta\eta}\sum_{k=0}^{K-1}\eta_k + \frac{\|u^{(0)}\|_{D_0}^2}{2}$$

Then, since η_k is a summable sequence, in view of (23), $\{(t_{k-1}^2 - t_k^2 + t_k)v_k\}$ is nonnegative and summable. Finally, observing that for $a > 2$, we have

$$0 \leq t_{k-1}^2 - t_k^2 + t_k = \frac{k(a-2) + (a-1)^2}{a^2}$$

we can conclude that also $\{kv_k\}$ is summable. □

The following lemma is a consequence of the previous one. It requires an additional condition on the scaling matrix sequence $\{D_k\}$. Indeed we will assume that the sequence $\{\eta_k\}$ in (28) is given by

$$\{\eta_k\} = \mathcal{O}\left(\frac{1}{k^p}\right) \quad \text{with } p > 2 \tag{33}$$

This assumption guarantees also that $\{k\eta_k\}$ is summable. When $\{D_k\}$ is chosen according to (32), the condition (33) is satisfied when $\zeta_k = \frac{b}{k^p}$ for any positive scalar b and $p > 2$.

LEMMA 3.7 *Let the assumption of Lemma 3.6 be fullfilled with $\{\eta_k\} = \mathcal{O}(\frac{1}{k^p})$, $p > 2$ in (28). Then, setting $\delta_k = \|x^{(k)} - x^{(k-1)}\|_{D_k}^2/2$, the sequence $\{k\delta_k\}$ is summable. In addition, there exists $D > 0$ such that for all $k \geq 1$, $\delta_k \leq \frac{D}{k^2}$.*

Proof. From (18) with $x = y^{(k)}$, $y = x^{(k+1)}$ and $z = x^{(k)}$, it follows that

$$F(x^{(k+1)}) + \frac{\|x^{(k)} - x^{(k+1)}\|_{D_k}^2}{2\alpha_k} \leq F(x^{(k)}) + \frac{\|x^{(k)} - y^{(k)}\|_{D_k}^2}{2\alpha_k} \tag{34}$$

From Lemma 2.1, since $x^{(k)} \in \mathrm{dom}(g) \subseteq Y$, we have

$$\|x^{(k)} - y^{(k)}\|_{D_k}^2 \leq \beta_k^2\|x^{(k)} - x^{(k-1)}\|_{D_k}^2$$

Then, subtracting $F(x^*)$ from both sides of (34), we can write

$$v_{k+1} + \frac{\|x^{(k)} - x^{(k+1)}\|_{D_k}^2}{2\alpha_k} \leq v_k + \beta_k^2\frac{\|x^{(k)} - x^{(k-1)}\|_{D_k}^2}{2\alpha_k} \tag{35}$$

From (28) we have

$$\delta_{k+1} = \frac{1}{2}\|x^{(k+1)} - x^{(k)}\|_{D_{k+1}}^2 \leq (1+\eta_k)\frac{1}{2}\|x^{(k+1)} - x^{(k)}\|_{D_k}^2$$

Then, since $\eta_k \geq 0$, from (35), it follows that

$$\delta_{k+1} \quad \leq \quad (1+\eta_k)(\alpha_k(v_k - v_{k+1}) + \beta_k^2\delta_k) \tag{36}$$

Since $\theta_k = \frac{a}{k+a}$ and $\beta_k = \frac{k-1}{k+a}$, (36) writes also as

$$\begin{aligned}(k+a)^2\delta_{k+1} - (1+\eta_k)(k-1)^2\delta_k &\leq (1+\eta_k)\alpha_k(k+a)^2(v_k - v_{k+1}) \\ &\leq \alpha_0(1+\eta_k)(k+a)^2(v_k - v_{k+1}) \end{aligned} \tag{37}$$

where the second inequality follows from (19).

Since $\{\eta_k\} = \mathcal{O}(\frac{1}{k^p})$ with $p > 2$, $\lim_{k\to\infty} \eta_k k^2 = 0$; therefore, given a scalar $\epsilon > 0$ such that $a^2 - 2a \geq \epsilon$, there exists an index ℓ such that for any $k > \ell$ we can write

$$\eta_k(k-1)^2 < \eta_k k^2 < \epsilon \leq a^2 - 2a \tag{38}$$

Summing up the inequality (37) for $k = \ell, ..., K$ yields

$$\begin{aligned}(K+a)^2\delta_{K+1} + \sum_{k=\ell+1}^{K} ((k-1+a)^2 - (1+\eta_k)(k-1)^2)\delta_k \leq (1+\eta_\ell)(\ell-1)^2\delta_\ell + \\ +\alpha_0((\ell+a)^2(1+\eta_\ell)v_\ell - (K+a)^2(1+\eta_K))v_{K+1} + \\ +\alpha_0 \sum_{k=\ell+1}^{K} ((k+a)^2(1+\eta_k) - (k-1+a)^2(1+\eta_{k-1}))v_k \end{aligned} \tag{39}$$

For all k we have

$$\begin{aligned}(k+a)^2(1+\eta_k) - (k-1+a)^2(1+\eta_{k-1}) &= \eta_k(k+a)^2 + 2(k+a) - 1 - (k-1+a)^2\eta_{k-1} \\ &\leq \eta_k(k+a)^2 + 2(k+a)\end{aligned}$$

and, in view of (38), for $k > \ell$ we also have

$$(k-1+a)^2 - (1+\eta_k)(k-1)^2 = a^2 + 2ka - 2a - \eta_k(k-1)^2 > 2ka > 0$$

Then, ignoring negative terms on the right hand side, we obtain

$$\begin{aligned}(K+a)^2\delta_{K+1} + \sum_{k=\ell+1}^{K} 2ka\delta_k \leq (1+\eta_\ell)(\ell-1)^2\delta_\ell + \\ +\alpha_0\left((\ell+a)^2(1+\eta_\ell)v_\ell + \sum_{k=\ell+1}^{K} 2(k+a)v_k + \eta_k(k+a)^2 v_k\right)\end{aligned} \tag{40}$$

Since $\eta_k(k+a)^2$ is bounded, by Lemma 3.6, the right hand side of the previous inequality is uniformly bounded independently on K. This ensures that $k\delta_k$ is summable. Furthermore $K^2\delta_{K+1}$ is globally bounded. □

We are now able to prove the first, weak, convergence result about the sequence generated by Algorithm 1, as stated in the following Corollary.

COROLLARY 3.1 *Let the assumptions of Lemma 3.7 be satisfied. Then, $\{x^{(k)}\}$ is bounded and any of its limit point is a solution of problem* (1).

Proof. A direct consequence Lemma 3.7 is that the sequence $\{k(x^{(k)} - x^{(k-1)})\}$ is bounded. From Lemma 3.5, it follows that the sequence $\{z^{(k)}\}$ defined in (22) is also bounded. These two facts imply that the sequence $\{x^{(k)}\}$ is bounded.

Assume that $\tilde{x}$ is a limit point of $\{x^{(k)}\}$, i.e. there exists a subsequence $\{x^{(k)}\}_{k\in\mathcal{K}}$ of $\{x^{(k)}\}$ such that $x^{(k)} \to \tilde{x}$, $k \in \mathcal{K}$ as $k \to \infty$. This element $\tilde{x}$ of $\mathrm{dom}(g)$ is also a limit point of $\{y^{(k)}\}$. Indeed, from Lemma 2.1, the definition of $y^{(k)}$ and the boundedness of β_k and $\{D_k\}$, we have for any $k \geq 1$ and, in particular for $k \in \mathcal{K}$:

$$\|y^{(k)} - x^{(k)}\|^2 \leq \frac{1}{\eta}\|y^{(k)} - x^{(k)}\|^2_{D_k} \leq \frac{2}{\eta}\beta_k^2\delta_k \leq \frac{2}{\eta}\delta_k$$

Under the assumption of Lemma 3.7, we have $\delta_k \to 0$, as $k \to \infty$, thus $\tilde{x}$ is a limit point of $\{y^{(k)}\}$. From assumption (28) we have that $\{D_k\}$ converges pointwise to some matrix $\mathfrak{D} \in \mathcal{D}_\eta$ [13, Lemma 2.3] and, since $\alpha_k \in [\delta\eta/L, \alpha_0]$ we can assume without loss of generality that α_k converges to some $\alpha > 0$ as k diverges, $k \in \mathcal{K}$. Therefore, $\tilde{x}$ is a fixed point of the operator $p_{\alpha,\mathfrak{D}}$ and, consequently, it is a minimizer of F. □

Before giving the main result stating that the whole sequence converges to a minimizer, we need to prove the following technical lemma, which holds when the matrices sequence $\{D_k\}$ satisfies both (28) and (31).

LEMMA 3.8 *Let the assumption of Lemma 3.6 be fullfilled and $\{D_k\}$ be a sequence of matrices satisfying both the conditions* (28) *and* (31). *Then, denoting $\Phi_k = \frac{\|x^{(k)}-x^*\|^2_{D_k}}{2}$, for $k \geq 1$ we have*

$$\Phi_{k+1} - \Phi_k \leq \beta_k(\Phi_k - \Phi_{k-1}) + 2\beta_k\delta_k(1+\eta_k) + \eta_k(1+\beta_k)\Phi_k + \beta_k\nu_{k-1}\frac{\|x^* - x^{(k-1)}\|^2_{D_k}}{2} \quad (41)$$

Proof. Let x^* be a solution of the problem (1). Using (18) in Lemma 2.4 with $y = x^{(k+1)}$, $x = y^{(k)}$, $z = x^*$, we have

$$\begin{aligned} F(x^{(k+1)}) + \frac{1}{2\alpha_k}\|x^* - x^{(k+1)}\|^2_{D_k} &\leq F(x^*) + \frac{1}{2\alpha_k}\|x^* - y^{(k)}\|^2_{D_k} \\ &\leq F(x^*) + \frac{1}{2\alpha_k}\|x^* - x^{(k)}\|^2_{D_k} + \frac{\beta_k^2}{2\alpha_k}\|x^{(k)} - x^{(k-1)}\|^2_{D_k} + \frac{\beta_k}{\alpha_k}(x^{(k)} - x^{(k-1)})^T D_k(x^{(k)} - x^*) \end{aligned} \quad (42)$$

where the last inequality follows from Lemma 2.1, since $y^{(k)} = P_{Y,D_k}(x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}))$. We observe that

$$(x^{(k)} - x^{(k-1)})^T D_k(x^{(k)} - x^*) = \frac{1}{2}\|x^{(k)} - x^{(k-1)}\|^2_{D_k} + \frac{1}{2}\|x^{(k)} - x^*\|^2_{D_k} - \frac{1}{2}\|x^{(k-1)} - x^*\|^2_{D_k}$$

Using this equality in (42) we obtain

$$\begin{aligned} F(x^{(k+1)}) + \frac{1}{2\alpha_k}\|x^* - x^{(k+1)}\|^2_{D_k} &\leq F(x^*) + \frac{1}{2\alpha_k}\|x^* - x^{(k)}\|^2_{D_k} + \frac{\beta_k^2}{2\alpha_k}\|x^{(k)} - x^{(k-1)}\|^2_{D_k} + \\ &+ \frac{\beta_k}{\alpha_k}\left(\frac{1}{2}\|x^{(k)} - x^{(k-1)}\|^2_{D_k} + \frac{1}{2}\|x^{(k)} - x^*\|^2_{D_k} - \frac{1}{2}\|x^{(k-1)} - x^*\|^2_{D_k}\right) \end{aligned}$$

Then we have

$$\frac{1}{2\alpha_k}\|x^* - x^{(k+1)}\|^2_{D_k} - \frac{1}{2\alpha_k}\|x^* - x^{(k)}\|^2_{D_k} \leq F(x^*) - F(x^{(k+1)}) + \frac{\beta_k^2 + \beta_k}{2\alpha_k}\|x^{(k)} - x^{(k-1)}\|^2_{D_k} +$$
$$+\frac{\beta_k}{\alpha_k}(\frac{1}{2}\|x^{(k)} - x^*\|^2_{D_k} - \frac{1}{2}\|x^{(k-1)} - x^*\|^2_{D_k})$$

Since $F(x^*) - F(x^{(k+1)}) \leq 0$ and $\beta_k^2 + \beta_k \leq 2\beta_k$, we obtain

$$\frac{1}{2}\|x^* - x^{(k+1)}\|^2_{D_k} - \frac{1}{2}\|x^* - x^{(k)}\|^2_{D_k} \leq \beta_k(\frac{1}{2}\|x^* - x^{(k)}\|^2_{D_k} - \frac{1}{2}\|x^* - x^{(k-1)}\|^2_{D_k}) + 2\beta_k\delta_k$$

Multiplying the last inequality by $(1 + \eta_k)$, from (28), we obtain

$$\begin{aligned}
&\Phi_{k+1} - (1+\eta_k)\Phi_k \leq \\
&\quad \leq \beta_k(1+\eta_k)(\Phi_k - \frac{\|x^* - x^{(k-1)}\|^2_{D_k}}{2}) + 2\beta_k(1+\eta_k)\delta_k \\
&\quad = \beta_k(\Phi_k - \Phi_{k-1}) + \beta_k\eta_k\Phi_k + \beta_k(\Phi_{k-1} - (1+\eta_k)\frac{\|x^* - x^{(k-1)}\|^2_{D_k}}{2}) + 2\beta_k(1+\eta_k)\delta_k
\end{aligned}$$

Thus, in view of the assumption (31), since

$$\Phi_{k-1} = \frac{\|x^* - x^{(k-1)}\|^2_{D_{k-1}}}{2} \leq (1+\nu_{k-1})\frac{\|x^* - x^{(k-1)}\|^2_{D_k}}{2}$$

we obtain (41). □

Now, as in [10], we introduce the notation

$$\begin{aligned}
\beta_{j,k} &= \Pi_{\ell=j}^k \beta_\ell = \Pi_{\ell=j}^k \frac{\ell - 1}{\ell + a} \quad j \geq 1, k \geq j \\
\beta_{j,k} &= 1 \quad j > k
\end{aligned} \tag{43}$$

Since $\beta_1 = 0$, $\beta_{1,k} = 0$ for $k \geq 1$. Moreover, in [10] it is proved that, for $a > 2$, the following inequality holds for $j \geq 2$

$$\sum_{k=j}^{\infty} \beta_{j,k} \leq \frac{j+5}{2} \tag{44}$$

This inequality is exploited in the proof of the following convergence theorem, whose line is very similar to that of Theorem 3 in [10]. This result requires that the sequence of matrices $\{D_k\}$ satisfies both the assumptions (28) and (31), where $\{\eta_k\}$ and $\{\nu_k\}$ are $\mathcal{O}(\frac{1}{k^p})$ with $p > 2$.

THEOREM 3.2 *Assume that $\{\theta_k\}$ and $\{\beta_k\}$ are chosen as in (27) with $a > 2$ and let $\{D_k\} \subset \mathcal{D}_\eta$ be a sequence of matrices satisfying (28) and (31) with $\sup_{k\in\mathbb{N}} \|D_k\| = \gamma < \infty$, $\{\eta_k\} = \mathcal{O}(\frac{1}{k^p})$ and $\{\nu_k\} = \mathcal{O}(\frac{1}{k^p})$ with $p > 2$. Then, the sequence $x^{(k)}$ converges to a minimizer of F.*

Proof. The proof follows [10, Theorem 3] and [23]. We first prove that $\Phi_k = \|x^{(k)} - x^*\|^2_{D_k}/2$ converges. From Corollary 3.1, we have that $\{x^{(k)}\}$ is a bounded sequence. Then, there exists

a positive scalar M such that $\frac{\|x^{(k)}-x^*\|^2}{2} \leq M$, for all $k \geq 0$. From the inequality (41), since $\sup_{k\in\mathbb{N}} \|D_k\| = \gamma$, we obtain

$$\Phi_{k+1} - \Phi_k \leq \beta_k(\Phi_k - \Phi_{k-1}) + 2\beta_k\delta_k(1+\eta_k) + \eta_k(1+\beta_k)\gamma M + \beta_k\nu_{k-1}\gamma M \tag{45}$$

Now, defining $p_k = \max(0, \Phi_k - \Phi_{k-1})$ and recalling that $\beta_k \leq 1$, we obtain

$$p_{k+1} \leq \beta_k p_k + 2\beta_k\delta_k(1+\eta_k) + \beta_k\gamma M(\eta_k + \nu_{k-1}) + \gamma M\eta_k \tag{46}$$

By applying (46) recursively and using (43) and $\beta_1 = 0$, it follows that

$$p_{k+1} \leq 2\sum_{j=2}^{k} \beta_{j,k}\delta_j(1+\eta_j) + \gamma M\sum_{j=2}^{k} \beta_{j,k}(\eta_j + \nu_{j-1}) + \gamma M\sum_{j=2}^{k} \beta_{j,k}\eta_{j-1} + \gamma M\eta_k$$

for all $k \geq 2$. Hence,

$$\begin{aligned}
\sum_{k=2}^{+\infty} p_{k+1} &\leq 2\sum_{k=2}^{+\infty}\sum_{j=2}^{k} \beta_{j,k}\delta_j(1+\eta_j) + \gamma M\sum_{k=2}^{+\infty}\sum_{j=2}^{k} \beta_{j,k}(\eta_j + \nu_{j-1}) + \\
&\quad + \gamma M\sum_{k=2}^{+\infty}\sum_{j=2}^{k} \beta_{j,k}\eta_{j-1} + \gamma M\sum_{k=2}^{+\infty} \eta_k \\
&\leq 2\sum_{j=2}^{+\infty} \delta_j(1+\eta_j)\sum_{k=j}^{+\infty} \beta_{j,k} + \gamma M\sum_{j=2}^{+\infty}(\eta_j + \nu_{j-1})\sum_{k=j}^{+\infty} \beta_{j,k} + \\
&\quad + \gamma M\sum_{j=2}^{+\infty} \eta_{j-1}\sum_{k=j}^{+\infty} \beta_{j,k} + \gamma M\sum_{k=2}^{+\infty} \eta_k \\
&\leq 2\sum_{j=1}^{+\infty} \delta_j(1+\eta_j)\frac{j+5}{2} + \gamma M\sum_{j=1}^{+\infty}(\eta_j + \nu_{j-1})\frac{j+5}{2} + \\
&\quad + \gamma M\sum_{j=1}^{+\infty} \eta_{j-1}\frac{j+5}{2} + \gamma M\sum_{k=1}^{+\infty} \eta_k
\end{aligned}$$

where the last inequality follows form (44). From the assumption on $\{\eta_j\}$ and $\{\nu_j\}$, $\{j\eta_j\}$ and $\{j\nu_j\}$ are summable; from Lemma 3.7, $\{j\delta_j\}$ is also summable. This implies that the right side of the last inequality is finite, therefore $\{p_k\}$ is summable. We set $q_k = \Phi_k - \sum_{i=1}^{k} p_i$ and since $\Phi_k \geq 0$ and $\sum_{i=1}^{\infty} p_i$ is bounded, we have that q_k is bounded from below. On the other hand

$$q_{k+1} = \Phi_{k+1} - p_{k+1} - \sum_{i=1}^{k} p_i \leq \Phi_{k+1} - \Phi_{k+1} + \Phi_k - \sum_{i=1}^{k} p_i = q_k$$

Therefore $\{q_k\}$ is a non-increasing sequence and it is convergent. This implies that $\Phi_k = s_k + \sum_{i=1}^{k} p_i$ is convergent.

Assume now that $\tilde{x} \in \mathrm{dom}(g)$ is a limit point of $\{x^{(k)}\}$, i.e. there exists a subsequence $\{x^{(k_i)}\}$ of $\{x^{(k)}\}$ such that $\lim_{i\to\infty} x^{(k_i)} = \tilde{x}$. By Corollary 3.1, $\tilde{x}$ is a minimizer of F. Thus, the first

part of the proof applies also to $\tilde{x}$ and we can conclude that $\{\|x^{(k)} - \tilde{x}\|^2_{D_k}\}$ converges. As a consequence of this we have

$$\lim_{k\to\infty} \|x^{(k)} - \tilde{x}\|^2_{D_k} = \lim_{i\to\infty} \|x^{(k_i)} - \tilde{x}\|^2_{D_{k_i}} = 0$$

Since $\eta\|x^{(k)} - x^*\|^2 \leq \|x^{(k)} - x^*\|^2_{D_k}$, the last equality implies that the whole sequence converges to the minimizer $\tilde{x}$. □

As a consequence of the analysis performed in this section, we can conclude that when the sequence of matrices $\{D_k\}$ is chosen so that the condition (32) holds with $\{\zeta_k\} = \mathcal{O}\left(\frac{1}{k^p}\right)$ with $p > 2$, and the extrapolation parameters θ_k and β_k are defined as in (27), Algorithm 1 generates a sequence $\{x^{(k)}\}$ convergent to a minimizer of F with a $\mathcal{O}\left(\frac{1}{k^2}\right)$ convergence rate for the objective function values.

4 Numerical experiments

In this section we present the results of several numerical experiments which aim at evaluating the effectiveness of the proposed scaled forward-backward method with extrapolation (SFBEM) by comparison with other state-of-the-art algorithms. The numerical experiments concern three different optimization problems which can be formalized as in (1) and arise from some relevant real-life applications.

4.1 Image deblurring with Poisson noise

We consider the inverse problem of recovering an unknown image x_{true} from a given data corrupted by noise. Bayesian approaches suggest to address this problem by minimizing a functional which can be expressed as the sum of a discrepancy function, typically depending on the noise type affecting the data, and a regularization term adding a priori information and possible constraints. In particular, in the case of Poisson noise, the discrepancy function measuring the distance from the data $b \in \mathbb{R}^n$ is the generalized Kullback-Leibler (KL) divergence of the form

$$\text{KL}(x) = \sum_{i=1}^{n} \left\{ b_i \ln \frac{b_i}{(Ax + bg)_i} + (Ax + bg)_i - b_i \right\} \tag{47}$$

where $A \in \mathbb{R}^{n\times n}$ is a linear operator modeling the distortion due to the image acquisition system and $bg \in \mathbb{R}^n$ is a known positive background radiation constant. A typical assumption for the matrix A is that it has nonnegative elements and each row and column has at least one positive entry. We refer the interested reader to [5] for a detailed survey on the image deblurring problem in presence of Poisson noise and the properties of the KL function (47).

As for the regularization term, we consider a smooth discrete version of the total variation, also known in the literature as *hypersurface potential* (HS) [1, 11, 31], that, for a square $m \times m$ image with $m^2 = n$, is defined as

$$\text{HS}(x) = \sum_{i,j=1}^{m} \sqrt{((\mathcal{D}x)_{i,j})^2_1 + ((\mathcal{D}x)_{i,j})^2_2 + \delta^2},$$

where $\mathcal{D} : \mathbb{R}^{m^2} \longrightarrow \mathbb{R}^{2m^2}$ is the discrete gradient operator with periodic boundary conditions

$$(\mathcal{D}x)_{i,j} = \begin{pmatrix} ((\mathcal{D}x)_{i,j})_1 \\ ((\mathcal{D}x)_{i,j})_2 \end{pmatrix} = \begin{pmatrix} x_{i+1,j} - x_{i,j} \\ x_{i,j+1} - x_{i,j} \end{pmatrix}, \quad x_{n+1,j} = x_{1,j}, \quad x_{i,n+1} = x_{i,1}. \tag{48}$$

In conclusion, a way to recover the true image from the corrupted data is to find a solution of the following optimization problem

$$\min_{x \in \mathbb{R}^n} F(x) \equiv \mathrm{KL}(x) + \rho\, \mathrm{HS}(x) + \iota_{x \geq 0}(x), \tag{49}$$

where ρ is a positive parameter balancing the role of the regularization term and $\iota_{x \geq 0}$ denotes the indicator function of the nonnegative orthant; indeed, the unknown (the pixels of the image) have to be nonnegative.

Problem (49) can be naturally cast in the form (1) by setting $f(x) = \mathrm{KL}(x) + \rho\, \mathrm{HS}(x)$ and $g(x) = \iota_{x \geq 0}(x)$.

In this case $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : Ax + bg > 0\}$ and ∇f is Lipschitz continuous on $Y = \{x \in \mathbb{R}^n : x \geq 0\} = \mathrm{dom}(g) \subseteq \mathrm{dom}(f)$. However, only an estimation from above of the Lipschitz constants of both ∇KL and ∇HS is known (see [19] and [21], respectively).

The numerical tests have been performed by solving the optimization problem (49) on two different datasets. The original images are the 128×128 micro [30] and the 256×256 Cameraman, both used in several papers and reported in the first row of Figure 1. The values of the micro original image are in the range $[0, 69]$, while the values of the Cameraman lay in the interval $[0, 1000]$. The corrupted data (Figure 1, second row) have been generated by convolving the objects with a suitable point spread function (the psf proposed in [30] for the micro image and a Gaussian psf, with standard deviation equal to 1.3, for the Cameraman one), adding a constant background equal to 1 and perturbing the result of the convolution with Poisson noise (simulated through the `imnoise` function of the Matlab Image Processing Toolbox). For both test problems we assume periodic boundary conditions, thus A is block-circulant with circulant blocks and the matrix vector products involving A can be performed via the Fast Fourier Transform.

We chose the regularization parameter ρ equal to 0.09 for the first test problem and 0.045 for the second one and the parameter δ for the HS functional equal to 0.05 for both datasets. We compare the SFBEM approach with some other recent methods:

- FISTA with backtracking [3], which can be considered as a special case of SFBEM where the scaling matrix is chosen at each iterate as the identity matrix. Actually our implementation slightly differs from the standard FISTA by the presence of the projection after the extrapolation step, which is needed when solving (49) since $\mathrm{dom}(f)$ does not coincide with $\mathbb{R}^n$;

- the scaled gradient projection method (SGP) [9, 31], which is a well known algorithm for differentiable constrained optimization problems. The SGP iteration has the form (4), where the proximity operator reduces to the projection operator onto the constraints set. Here the selection of steplength parameter α_k is based on the adaptive alternation of the Barzilai-Borwein rules proposed in [9] and the value of λ_k is computed by a line search procedure. We point out that there exists a more recent variant of SGP [27] which employs a different steplength selection rule. In order to avoid redundant results, we prefer to consider only the standard SGP approach as a comparative tool;

Micro Cameraman

Figure 1: First row: original images for the two image deblurring test problems. Second row: blurred and noisy images for the two image deblurring test problems.

- the nonscaled version of SGP, hereafter indicated by GP.

The scaling matrix for SFBEM and SGP has been selected by exploiting the split gradient idea suggested in [22] and based on a decomposition of the gradient into a nonnegative part and a negative one. In particular, in [31] the authors show that the gradient of $f(x) = \mathrm{KL}(x) + \rho\,\mathrm{HS}(x)$ can be decomposed in the form

$$-\nabla f(x) = U_{\mathrm{KL}}(x) + \rho U_{\mathrm{HS}}(x) - V_{\mathrm{KL}}(x) - \rho V_{\mathrm{HS}}(x)$$

with $U_{\mathrm{KL}}, U_{\mathrm{HS}} \geq 0$ and $V_{\mathrm{KL}}, V_{\mathrm{HS}} > 0$. Then a possible scaling matrix is given by

$$D_k = \mathrm{diag}\left(\max\left(\frac{1}{\gamma_k}, \min\left(\gamma_k, \frac{w^{(k)}}{V_{\mathrm{KL}}(w^{(k)}) + \rho V_{\mathrm{HS}}(w^{(k)})}\right)\right)\right)^{-1} \tag{50}$$

where the quotient is componentwise and $w^{(k)}$ is equal to the previous iterate $x^{(k)}$ for SGP and to the extrapolated point $y^{(k)}$ for SFBEM. Moreover we set $\gamma_k = \sqrt{1 + \dfrac{10^{13}}{(k+1)^p}}$, $p = 2.1$.

We remark that in this case the projection at Step 1 of SFBEM is independent on the scaling matrix D_k, since Y is the nonnegative orthant, i.e. $P_{Y,D} \equiv P_Y$. Thus we are allowed to first compute $y^{(k)} = P_Y(x^{(k)} + \beta_k(x^{(k)} - x^{(k-1)}))$ then choosing D_k depending on $y^{(k)}$ for updating $x^{(k+1)}$.

Finally, the sequence $\{\beta_k\}$ employed in the definition of the extrapolation step for both FISTA and SFBEM has been chosen as in (27) with $a = 2.1$ in order to ensure the convergence of the sequence of the iterates.

The performance of the algorithms has been compared by evaluating their ability in reducing the objective function: in particular we computed an approximate solution x^* of (49) by performing 20000 SGP iterations and, for any method, we consider at each iterate the relative difference between the objective function and the minimum value

$$\frac{F(x^{(k)}) - F(x^*)}{F(x^*)} \tag{51}$$

Table 1 and Table 2 report the number of iterations and the computational time needed by each method to reduce the relative difference (51) below a certain tolerance *tol*. Since in both test problems the solution x^* is unique, the relative minimization error $\mathrm{RME}(x^{(k)}) = \frac{\|x^{(k)}-x^*\|}{\|x^*\|}$ is also reported. The computational time presented is the average execution time (in seconds) over ten runs.

Figure 2 shows the decrease of the relative differences (51) with respect to both the iteration number and the computational time. We also observed that once the quantity (51) is below the tolerance 10^{-7}, all algorithms provide the same relative reconstruction error $\mathrm{RRE}(x^{(k)}) = \frac{\|x^{(k)}-x_{\mathrm{true}}\|}{\|x_{\mathrm{true}}\|}$, which measures the quality of the computed solution as an approximation of x_{true}. More precisely, this value is 0.088 for the micro test problem $\big(\mathrm{RRE}(b_{\,\mathrm{micro}}) = 0.195\big)$ and 0.087 for the Cameraman one $\big(\mathrm{RRE}(b_{\,\mathrm{Cameraman}}) = 0.121\big)$.

	Micro								
	$tol = 10^{-3}$			$tol = 10^{-5}$			$tol = 10^{-7}$		
	It.	RME	Time	It.	RME	Time	It.	RME	Time
GP	585	0.0414	5.69	2100	0.0077	21.31	3459	0.0014	34.70
SGP	75	0.0261	0.72	203	0.0061	2.14	336	0.0016	4.48
FISTA	223	0.0410	4.12	888	0.0048	15.26	3298	0.0005	56.05
SFBEM	64	0.0202	0.84	188	0.0038	2.17	515	0.0004	6.72

Table 1: Number of iterations and computational time required by each algorithm to reduce the relative difference (51) below given tolerances for the micro test problem. The corresponding RME and computational time (average over 10 runs) are also reported.

	Cameraman								
	$tol = 10^{-3}$			$tol = 10^{-5}$			$tol = 10^{-7}$		
	It.	RME	Time	It.	RME	Time	It.	RME	Time
GP	1730	0.0134	76.76	4046	0.0021	164.07	5637	0.0003	229.16
SGP	241	0.0102	9.37	1178	0.0014	47.48	1671	0.0001	76.24
FISTA	226	0.0105	13.13	858	0.0011	54.46	3332	0.0001	220.71
SFBEM	42	0.0103	2.90	163	0.0012	10.31	705	0.0001	48.41

Table 2: Number of iterations and computational time required by each algorithm to reduce the relative difference (51) below given thresholds for the Cameraman test problem. The corresponding RME and computational time (average over 10 runs) are also reported.

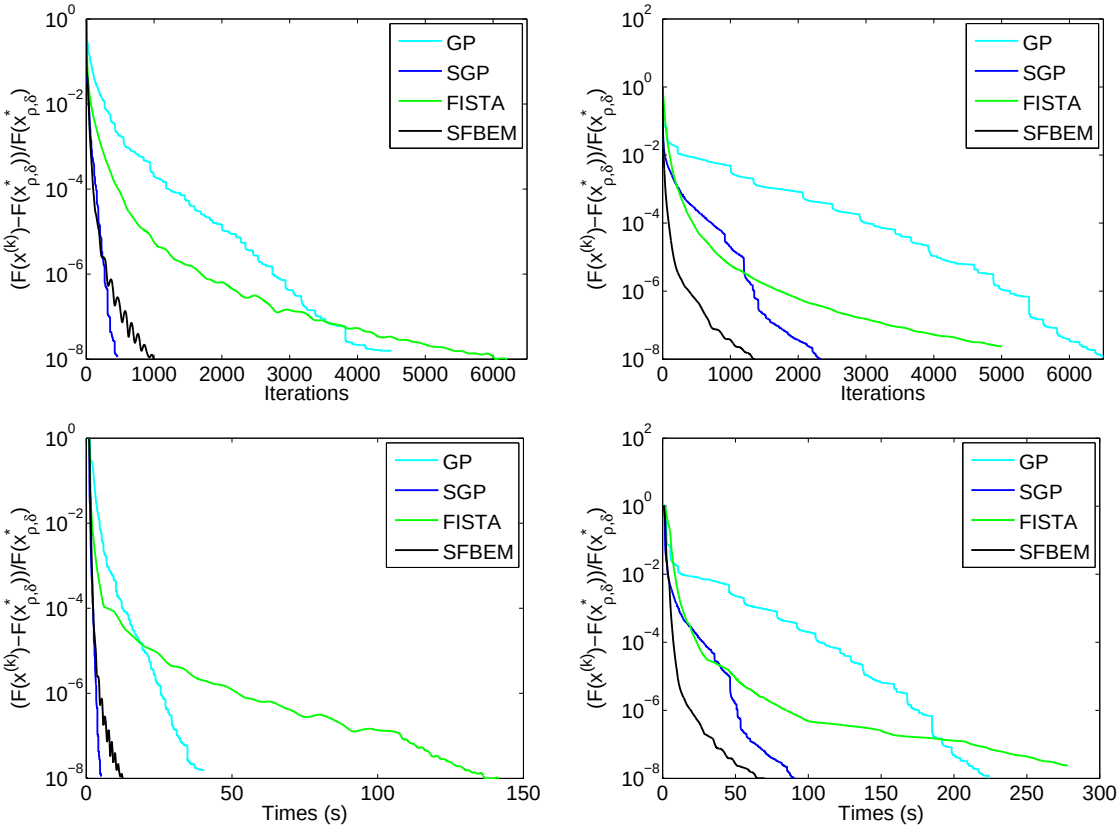

Figure 2: Plots of the relative difference (51) with respect to the iterations number (top) and computational time (bottom) for the micro (left) and Cameraman (right) test problems.

From the numerical results shown in Tables 1 and 2 and in Figure 2, it is possible to conclude that the presence of a non trivial scaling matrix makes the performances of SFBEM always superior to those of the nonscaled FISTA approach in terms of both number of iterations and computational time, while providing the same RRE and RME. Moreover, the comparison with SGP also supports the effectiveness of SFBEM: indeed, the performances of SFBEM are as good as those provided by SGP which is known in the literature as one of the most competitive algorithms to deal with image deblurring problems.

4.2 Compressed sensing with Poisson noise

As a second benchmark framework, we consider a compressed sensing problem which consists in recovering a sparse vector of nonnegative values starting from noisy measurements. More in detail, we assume that the observed data $b \in \mathbb{R}^m$ is the realization of a Poisson random variable with expected value given by $Ax_{\text{true}} + bg$, where $x_{\text{true}} \in \mathbb{R}^n$ is the signal of interest, $A \in \mathbb{R}^{m \times n}$ is the measurement matrix and bg is a known background. As suggested in [28], the true signal

can be reconstructed by addressing a minimization problem of the form

$$\min_{x\in\mathbb{R}^n} \mathrm{KL}(x) + \rho\|x\|_1 + \iota_{x\geq 0}(x) \tag{52}$$

where KL is the generalized Kullback-Leibler divergence (47), the ℓ_1 norm induces sparsity on the solution, ρ is the positive regularization parameter and $\iota_{x\geq 0}$ is the indicator function of the nonnegative orthant.

In this case, we set $f(x) = \mathrm{KL}(x)$ and $g(x) = \rho\|x\|_1 + \iota_{x\geq 0}(x)$ and Y is the nonnegative orthant. The operator $p_{\alpha,D}(x)$ associated to $g(x)$ can be computed in closed form [19, Section II].

The numerical experiments are carried out on a test problem which has been generated with the following steps:

(i) a matrix $A \in \mathbb{R}^{1000\times 5000}$ has been generated as detailed in [28] so that A preserves both the positivity and the flux of any signal (i.e. if $z \geq 0$ then $Az \geq 0$ and $\sum_{i=1}^m (Az)_i \leq \sum_{i=1}^n z_i$);

(ii) the signal to recover $x_{\mathrm{true}} \in \mathbb{R}^{5000}$ has all zeros except for 20 non-zero entries drawn uniformly in the interval $[0, 10^5]$;

(iii) the observed signal $b \in \mathbb{R}^{1000}$ has been obtained by corrupting the vector $Ax_{\mathrm{true}} + bg$ ($bg = 10^{-10}$) by means of the Matlab `imnoise` function.

The regularization parameter ρ has been fixed equal to 10^{-3}.

We compare SFBEM, FISTA with backtracking and the SPIRAL method developed in [19] and designed to solve problems of the type (52). The SPIRAL approach is a forward-backward algorithm that employs a steplength selection strategy based on the Barzilai–Borwein rules [2]; the convergence is guaranteed by a proper linesearch on the values of the objective function. The Matlab code of SPIRAL is available on-line [18]. The scaling matrix for SFBEM has been selected by exploiting the already mentioned decomposition idea of the gradient of the differentiable part of the objective function: in particular, by writing the gradient of the KL functional as

$$-\nabla \mathrm{KL}(x) = U_{\mathrm{KL}}(x) - V_{\mathrm{KL}}(x)$$

with $U_{\mathrm{KL}} \geq 0$ and $V_{\mathrm{KL}} > 0$, D_k is defined as

$$D_k = \mathrm{diag}\left(\max\left(\frac{1}{\gamma_k}, \min\left(\gamma_k, \frac{y^{(k)}}{V_{\mathrm{KL}}(y^{(k)})}\right)\right)\right)^{-1}$$

with $\gamma_k = \sqrt{1 + \dfrac{10^6}{(k+1)^p}}$, $p = 2.1$. The sequence $\{\beta_k\}$ used to update the extrapolation point has been chosen as in (27) with $a = 10$. The considered methods have been stopped when the relative distance between two successive iterations is less than 10^{-7}. Table 3 shows the performance of FISTA, SFBEM and SPIRAL in solving problem (52) in terms of number of iterations and computational time (average over ten runs) to make the relative distance (51) smaller than prefixed thresholds. For this test problem, the minimum point x^* has been computed by the SPIRAL method in 10000 iterations. Moreover, we report the relative reconstruction error and the relative minimization error. In order to better appreciate the results, Figure 3 depicts the decreasing behavior of the objective function with respect to both the number of iterations and

	Compressed sensing problem								
	$tol = 10^{-3}$			$tol = 10^{-5}$			$tol = 10^{-7}$		
	It.	RME	Time	It.	RME	Time	It.	RME	Time
FISTA	637	0.0111	15.75	933	0.0011	23.62	1412	0.0001	35.87
SFBEM	369	0.0096	8.77	568	0.0011	14.37	806	0.0001	20.04
SPIRAL	1180	0.0120	15.87	1309	0.0013	17.33	1379	0.0001	18.10

Table 3: Number of iterations and computational time required by each algorithm to reduce the relative difference (51) below given tolerances for the compressed sensing test problem. The corresponding RME and computational time (average over 10 runs) are also reported.

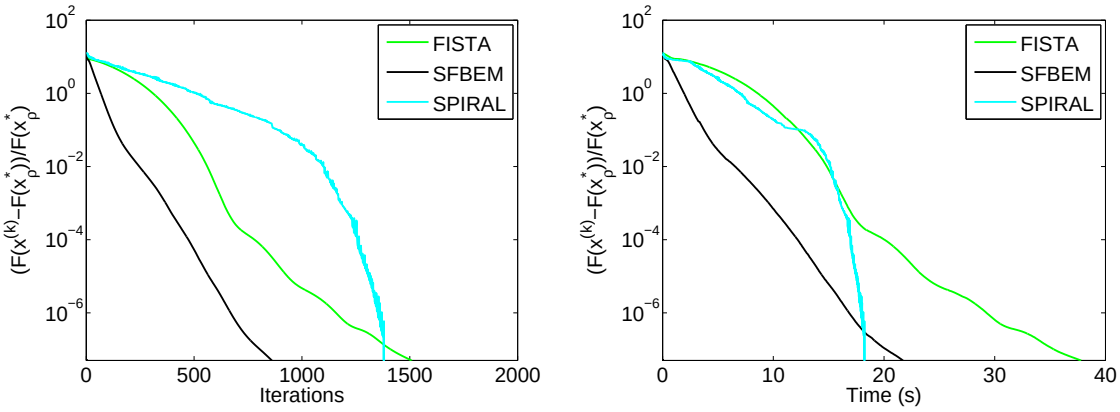

Figure 3: Compressed sensing problem: relative difference (51) with respect to the iterations number (left) and computational time (right).

the computational time. All the considered algorithms yield to the same value of the RRE equal to 0.075.

The results obtained on the compressed sensing problem confirm the same conclusions in the image deblurring framework: the benefit of applying the SFBEM instead of FISTA is evident from the significant reduction of the number of iterations and computational time, as reported in Table 3.

4.3 Probability density estimation

The last optimization problem we considered concerns the estimation of an unknown Gaussian mixture probability density. More in detail, if the sample $\{\tau_1, \tau_2, ..., \tau_n \,|\, \tau_i \in \mathbb{R}\}$ has been drawn from an unknown probability density function $\mu(\tau)$ which can be expressed as a Gaussian mixture, then a possible estimator has the form [17]

$$\hat{\mu}(\tau) = \sum_{i=1}^{n} x_i \kappa_\sigma(\tau, \tau_i) \tag{53}$$

where $\kappa_\sigma(\cdot, \tau_i)$ is a Gaussian kernel with variance σ and center τ_i and x_i is a suitable coefficient. In [17] the authors proved that the weight vector x can be computed as a solution of the following

minimization problem

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{2}x^TCx - p^Tx + \iota_{\Delta_1^+}(x) \tag{54}$$

where

(a) the element $C_{i,j}$ of the matrix $C \in \mathbb{R}^{n\times n}$ is the Gaussian kernel of variance 2σ, i.e. $C_{i,j} = \kappa_{2\sigma}(\tau_i, \tau_j)$;

(b) the i-th component of the vector $p \in \mathbb{R}^n$ is defined as $p_i = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} \kappa_\sigma(\tau_i, \tau_j)$;

(c) $\iota_{\Delta_1^+}$ is the indicator function of the simplex $\Delta_1^+ = \{x \in \mathbb{R}^n \,|\, x_i \geq 0 \ \forall i, \ \ \sum_{i=1}^n x_i = 1\}$.

In this case we set $f(x) = \frac{1}{2}x^TCx - p^Tx$, $g(x) = \iota_{\Delta_1^+}(x)$ and $Y = \mathbb{R}^n$. Thus, the operator $p_{\alpha,D}(x)$ consists in the projection onto the simplex Δ_1^+. Such projection can be formulated as a root-finding problem and effectively computed by the secant-like algorithm proposed in [16].

For the numerical experiments we analyzed the following Gaussian mixture

$$\mu(\tau) = \frac{1}{5}\sum_{i=1}^{5} \kappa_{\sigma_i}(\tau, c_i)$$

with $\sigma_i = \sqrt[4]{\left(\dfrac{7}{9}\right)^{i-1}}$ and $c_i = 14\left(\left(\dfrac{7}{9}\right)^{i-1} - 1\right)$, $i = 1, ..., 5$. The matrix C and the vector p have been generated with a sample of 1000 points drawn from μ by using the `gmdistribution` function of the Matlab Statistics and Machine Learning Toolbox.

The effectiveness of SFBEM has been evaluated in a comparison with FISTA with backtracking, SGP and GP.

The scaling matrix for SFBEM and SGP has been selected in the form

$$D_k = \operatorname{diag}\left(\max\left(\frac{1}{\gamma_k}, \min\left(\gamma_k, \frac{w^{(k)}}{Cw^{(k)}}\right)\right)\right)^{-1}$$

with $\gamma_k = \sqrt{1 + \dfrac{10^{10}}{(k+1)^p}}$, $p = 2.1$ and $w^{(k)}$ equal to $y^{(k)}$ for SFBEM and $x^{(k)}$ for SGP. This choice of the scaling matrix mimics the split gradient based scaling proposed in [4] for quadratic problems. The extrapolation parameters sequence $\{\beta_k\}$ has been chosen as in (27) with $a = 2.1$. Table 4 reports the number of iterations and the computational time (an average value over ten runs) needed by the four methods to ensure a sufficient decrease of the distances (51), where x^* has been computed by means of 25000 FISTA iterations. GP and SGP do not succeed in satisfying the more restrictive threshold within the prefixed maximum number of iterations (25000). In Figure 4 we can appreciate the decrease of the objective function obtained by the considered algorithms and in Figure 5 the reconstruction of the probability density function (pdf) through the estimator (53) where for simplicity we assume $\sigma = 1$.

The numerical experiments performed in the probability density estimation setting reinforce the validity of the proposed SFBEM scheme in comparison with the other state-of-the-art approaches we tested.

	Probability density estimation problem					
	$tol = 10^{-3}$		$tol = 10^{-5}$		$tol = 10^{-7}$	
	It.	Time	It.	Time	It.	Time
GP	111	0.51	20479	114.62	-	-
SGP	90	0.98	4305	26.01	-	-
FISTA	54	0.94	2141	16.54	21885	165.00
SFBEM	53	0.67	810	6.46	3883	28.88

Table 4: Number of iterations and computational time required by each algorithm to bring the relative difference (51) below given thresholds for the probability density estimation test problem.

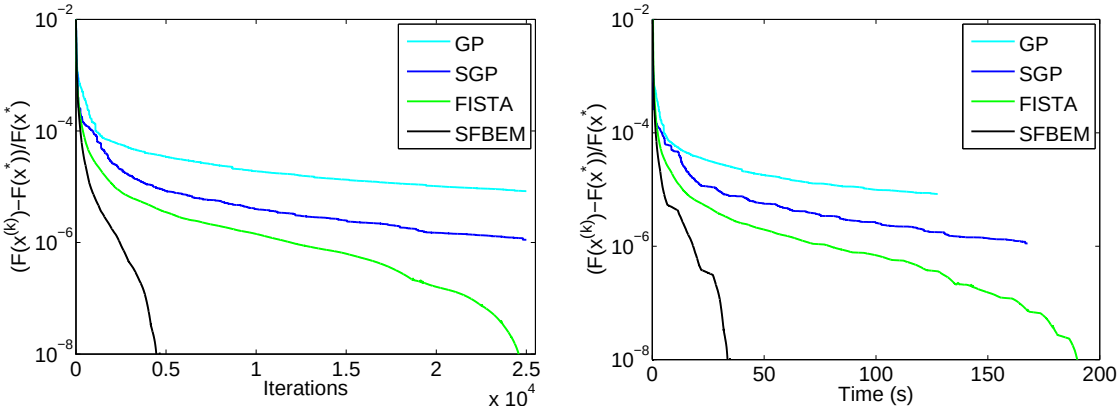

Figure 4: Relative difference between the objective function values $F(x^{(k)})$ provided by the different methods and the minimum value $F(x^*)$.

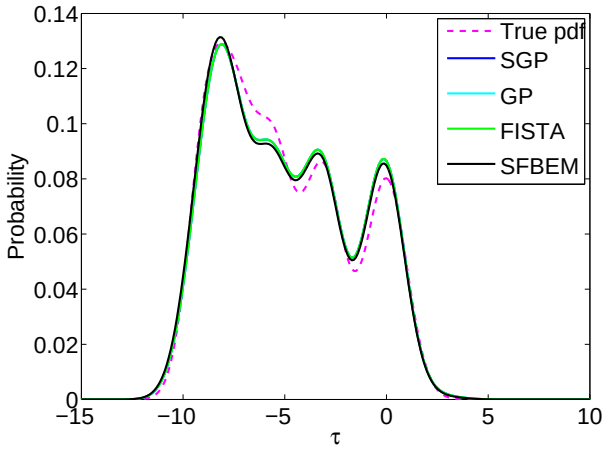

Figure 5: Density estimation results.

5 Conclusions

In this paper we proposed a variable metric forward-backward method with extrapolation based on two fundamental ingredients: a symmetric and positive definite scaling matrix multiplying the gradient of the differentiable part of the objective function, possibly capturing useful features of the problem to handle, and an inertial step which employs the information of the two last iterations in order to compute the new one. A proper backtracking strategy ensuring a sufficient decrease of the objective function and suitable adaptive bounds on the scaling matrix allow to prove the convergence of the scheme to a minimizer of the considered problem. We also provided a convergence rate estimate which is similar to existing convergence rate results for nonscaled forward-backward algorithms with extrapolation. Numerical experiments, carried out on optimization problems of different nature, showed very promising results in comparison with other algorithms which have already gained a great popularity in the literature. Future work will be addressed to analyze the possibility of introducing an inexact solution of the minimization problem which characterizes the backward step.

References

[1] R. Acar and C. R. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Problems*, 10(6):1217–1229, 2004.

[2] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8(1):141–148, January 1988.

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

[4] F. Benvenuto, R. Zanella, L. Zanni, and M. Bertero. Nonnegative least-squares image deblurring: improved gradient projection approaches. *Inverse Probl.*, 26(2):025004, February 2010.

[5] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini. Image deblurring with Poisson data: From cells to galaxies. *Inverse Problems*, 25(12), 2009.

[6] D. Bertsekas. *Convex Optimization Theory*, chapter 6 on Convex Optimization Algorithms, pages 251–489. 5 november 2012 edition.

[7] E. G. Birgin, J. M. Martinez, and M. Raydan. Inexact spectral projected gradient methods on convex sets. *IMA J. Numer. Anal.*, 23(4):539–559, October 2003.

[8] S. Bonettini and M. Prato. New convergence results for the scaled gradient projection method. *submitted*, available on http://arxiv.org/abs/1406.6601, 2015.

[9] S. Bonettini, R. Zanella, and L. Zanni. A scaled gradient projection method for constrained image deblurring. *Inverse Probl.*, 25(1):015002, January 2009.

[10] A. Chambolle and Ch. Dossal. On the convergence of the iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm". *J. Optim. Theory Appl.*, 2015.

[11] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing*, 6(2):298–311, 1997.

[12] P.L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-point algorithms for inverse problems in science and engineering*, Springer Optimization and Its Applications, pages 185–212. Springer, New York NY, 2011.

[13] P.L. Combettes and B.C. Vũ. Variable metric quasi-Féjer monotonicity. *Nonlinear Analysis: Theory, Methods, and Applications*, 78:17–31, feb 2013.

[14] P.L. Combettes and B.C. Vũ. Variable metric forward-backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[15] P.L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[16] Y. H. Dai and R. Fletcher. New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds. *Math. Program., Ser. A*, 106:403–421, 2006.

[17] M. Girolami and C. He. Probability density estimation from optimally condensed data samples. *IEEE Trans. Pattern Anal. Mach. Intell*, 25(10):1253–1264, 2003.

[18] Z.T. Harmany, R.F. Marcia, and R.M. Willett. The Sparse Poisson Intensity Reconstruction ALgorithms (SPIRAL) toolbox, 2012. http://drz.ac/code/spiraltap/.

[19] Z.T. Harmany, R.F. Marcia, and R.M. Willett. This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms - Theory and practice. *IEEE Transactions on Image Processing*, 21(3):1084–1096, 2012.

[20] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer Verlag, Heidelberg, 2001.

[21] T.L. Jensen, J. Jrgensen, P. Hansen, and S.H. Jensen. Implementation of an optimal first-order method for strongly convex total variation regularization. *BIT*, 52(2):329–356, 2012.

[22] H. Lantéri, M. Roche, O. Cuevas, and C. Aime. A general method to devise maximum likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Process.*, 81(5):945–974, May 2001.

[23] D. Lorentz and T. Pock. An inertial forward–backward algorithm for monotone inclusion. *J. Math. Imaging Vis.*, 2014.

[24] Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103:127–152, 2005.

[25] P. Ochs, Y. Chen, T. Brox, and T. Pock. iPiano: Inertial proximal algorithm for non-convex optimization. *SIAM J. Imaging Sci.*, 7(2):1388–1419, 2014.

[26] B. Polyak. *Introduction to Optimization*. Optimization Software - Inc., Publication Division, N.Y., 1987.

[27] F. Porta, M. Prato, and L. Zanni. A new steplength selection for scaled gradient methods with application to image deblurring. *Journal of Scientific Computing*, page in press, 2015.

[28] M. Raginsky, R.M. Willett, Z.T. Harmany, and R.F. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010.

[29] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23(3):1607–1633, 2013.

[30] R. M. Willett and R. D. Nowak. Platelets: A multiscale approach for recovering edges and surfaces in photon limited medical imaging. *IEEE Transactions on Medical Imaging*, 22:332–350, 2003.

[31] R. Zanella, P. Boccacci, L. Zanni, and M. Bertero. Efficient gradient projection methods for edge-preserving removal of Poisson noise. *Inverse Probl.*, 25(4):045010, April 2009.