IRIS

INSTITUTIONAL RESEARCH INFORMATION SYSTEM
ARCHIVIO ISTITUZIONALE DEI PRODOTTI DELLA RICERCA

intestazione repositorydell'ateneo

New convergence results for the scaled gradient projection method

(Article begins on next page)

18 September 2017

New convergence results for the scaled gradient projection method

S Bonettini[1] and M Prato[2]

[1] Dipartimento di Matematica e Informatica, Università di Ferrara, Via Saragat 1, 44122 Ferrara, Italy

[2] Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, Via Campi 213/b, 41125 Modena, Italy

E-mail: `silvia.bonettini@unife.it`, `marco.prato@unimore.it`

Abstract. The aim of this paper is to deepen the convergence analysis of the scaled gradient projection (SGP) method, proposed by Bonettini et al. in a recent paper for constrained smooth optimization. The main feature of SGP is the presence of a variable scaling matrix multiplying the gradient, which may change at each iteration. In the last few years, an extensive numerical experimentation showed that SGP equipped with a suitable choice of the scaling matrix is a very effective tool for solving large scale variational problems arising in image and signal processing. In spite of the very reliable numerical results observed, only a weak convergence theorem is provided establishing that any limit point of the sequence generated by SGP is stationary. Here, under the only assumption that the objective function is convex and that a solution exists, we prove that the sequence generated by SGP converges to a minimum point, if the scaling matrices sequence satisfies a simple and implementable condition. Moreover, assuming that the gradient of the objective function is Lipschitz continuous, we are also able to prove the $\mathcal{O}(1/k)$ convergence rate with respect to the objective function values. Finally, we present the results of a numerical experience on some relevant image restoration problems, showing that the proposed scaling matrix selection rule performs well also from the computational point of view.

1. Introduction

Several inverse problems in applied sciences can be addressed by means of a constrained optimization problem

$$\min_{\boldsymbol{x} \in \Omega} f(\boldsymbol{x}), \tag{1}$$

where $\Omega \subseteq \mathbb{R}^n$ is a closed and convex set and f is a continuously differentiable function. First order methods are attractive for solving (1) especially when n is large and when

the Hessian is not available or difficult to exploit. Indeed, the main strengths of these methods are, in general, the low memory requirement and the low computational cost per iteration. When the constraints set Ω has some special structure (e.g. box constraints, simplexes, balls), gradient projection (GP) methods have shown to be a valid tool to solve (1) in a variety of frameworks, such as signal and image processing [1, 2], statistical inference [3, 4] and machine learning [5, 6, 7].

The increasing popularity of first order methods gave rise in the recent literature to several studies aiming to devise suitable approaches for improving the convergence properties. In particular, we mention the extrapolation/inertial techniques [8, 9, 10] and the variable metric approach [11, 12, 13]. In the first case, an extrapolation step ensures the $\mathcal{O}(1/k^2)$ convergence rate of the objective function values to the optimal one, where k is the iteration index. In the latter case, an acceleration of the progress towards the solution is achieved by adopting a variable metric, which, at each iteration, could better capture the local features of the problem.

The focus of this work is on the scaled gradient projection (SGP) method [12], a variable metric algorithm based on a Armijo line–search. The basic SGP iteration is given by

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \lambda^{(k)}\boldsymbol{d}^{(k)} = \boldsymbol{x}^{(k)} + \lambda^{(k)}(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}), \tag{2}$$

where $\boldsymbol{y}^{(k)}$ is the scaled Euclidean projection of $\boldsymbol{x}^{(k)} - \alpha_k D_k \nabla f(\boldsymbol{x}^{(k)})$ onto Ω, i.e.

$$\boldsymbol{y}^{(k)} = \arg\min_{\boldsymbol{x}\in\Omega} \nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{x} - \boldsymbol{x}^{(k)}) + \frac{1}{2\alpha_k}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^T D_k^{-1}(\boldsymbol{x} - \boldsymbol{x}^{(k)}), \tag{3}$$

$\alpha_k > 0$ is the stepsize parameter, D_k is a symmetric positive definite scaling matrix and $\lambda^{(k)} \in (0, 1]$ is computed by means of a linesearch backtracking procedure to guarantee the sufficient decrease of the objective function.

Iteration (2)–(3) is a special case of a variable metric forward–backward method, which applies to any problem of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(x) + g(x) \tag{4}$$

where f is continuously differentiable and g is convex, proper and lower semicontinuous and it is described by the iteration

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \lambda_k(\mathrm{prox}_{\alpha_k g}^{D_k^{-1}}(\boldsymbol{x}^{(k)} - \alpha_k D_k \nabla f(\boldsymbol{x}^{(k)})) - \boldsymbol{x}^{(k)}) \tag{5}$$

where

$$\mathrm{prox}_{\alpha g}^{D}(\boldsymbol{y}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} g(\boldsymbol{x}) + \frac{1}{2\alpha}(\boldsymbol{x} - \boldsymbol{y})^T D(\boldsymbol{x} - \boldsymbol{y}). \tag{6}$$

Clearly, (2)–(3) is recovered when $g(\boldsymbol{x})$ is the indicator function of the set Ω.

Several recent papers [13, 14, 15] deal with method (5), providing convergence results, under the assumption that the gradient of f is Lipschitz continuous, in the convex [13] and nonconvex case for functions satisfying the Kurdyka-Łojasiewicz property [14, 15]. Even if the methods considered in the above mentioned papers are formally identical to

(2)–(3) when $g(\boldsymbol{x})$ is the indicator function of the set Ω, our approach is substantially different due to the different meaning of the parameters α_k, λ_k, D_k involved in the iteration, and it requires a different analysis.

Indeed, in our approach, α_k and D_k have to be considered as 'free' parameters which, when chosen in a clever way, can significantly improve the convergence behaviour of the algorithm (see e.g. [16, 17, 18, 19]). In particular, the recent literature shows that a suitable combination of the stepsize parameter α_k and of the scaling matrix D_k makes SGP a very effective tool in solving convex [12, 20, 21, 22] and nonconvex [23, 24, 25, 26] problems arising in signal and image processing applications.

However, the convergence analysis of SGP available in the literature only establish that, when α_k and the eigenvalues of D_k are bounded above and below away from zero, any limit point of the sequence $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ is stationary for problem (1). This result has been proved in [12, Theorem 2.1] without any further assumption and it is mainly based on the properties of the Armijo linesearch.

In this paper we provide a new, stronger, convergence result for SGP when applied to convex problems, establishing the convergence of the sequence $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ to a solution of (1), provided that the eigenvalues of D_k converge to one as k diverges, at a certain rate. Moreover, if we further assume that the gradient of f is Lipschitz continuous, we provide a new $\mathcal{O}(1/k)$ complexity result on the objective function value.

We observe that the $\mathcal{O}(1/k)$ complexity result is worse than the one obtained for the inertial/extrapolation methods such as the celebrated FISTA [9]. However, we also show that the practical performances of SGP are comparable to FISTA on some significant image restoration problems. Our numerical experience also shows that the condition on the scaling matrix selection ensuring the theoretical convergence of the method is also useful from a computational point of view.

The paper is organized as follows: in section 2 we recall the basic properties of the Armijo linesearch procedure and of descent methods. The analysis of the scaled gradient projection method is performed in section 3, where the relationship between our approach and the related literature is also discussed, while section 4 is devoted to some illustrative numerical examples. Our conclusions are given in section 5.

Notation and basic definitions In the following $\|\cdot\|$ indicates the ℓ_2 norm of a vector while $\|\cdot\|_D$ denotes the norm induced by the symmetric positive definite matrix D, i.e. $\|\boldsymbol{x}\|_D = (\boldsymbol{x}^T D \boldsymbol{x})^{\frac{1}{2}}$; $\mathbb{R}^n_{>0}$ and $\mathbb{R}^n_{\geq 0}$ denote the positive and non-negative orthants of $\mathbb{R}^n$, respectively; $\mu_{\min}(A), \mu_{\max}(A)$ are the minimum and maximum eigenvalue of a square matrix A, respectively. The notation $A \succeq B$, where $A, B \in \mathbb{R}^{n\times n}$ are symmetric positive semidefinite matrices, indicates that $A - B$ is positive semidefinite. Given $\mu \geq 1$, we denote by $\mathcal{M}_\mu$ the set of the symmetric positive definite matrices with all eigenvalues contained in the interval $[\frac{1}{\mu}, \mu]$. For any $D \in \mathcal{M}_\mu$ we have that D^{-1} also belongs to $\mathcal{M}_\mu$

and

$$\frac{1}{\mu}\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{x}\|_D^2 \leq \mu\|\boldsymbol{x}\|^2 \qquad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{7}$$

We also recall the definitions of *stationary point* and *descent direction* for problem (1) (see for example [27]).

Definition 1.1 *A point $\boldsymbol{x} \in \Omega$ is a stationary point for problem (1) if*

$$\nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{y} \in \Omega.$$

Definition 1.2 *Let $\boldsymbol{x}$ be any point of the set Ω.*

(i) *A vector $\boldsymbol{d} \in \mathbb{R}^n$ is a feasible direction at $\boldsymbol{x}$ if $\boldsymbol{x} + \boldsymbol{d} \in \Omega$.*

(ii) *A vector $\boldsymbol{d} \in \mathbb{R}^n$ is a descent direction at $\boldsymbol{x}$ for problem (1) if it is feasible and $\nabla f(\boldsymbol{x})^T \boldsymbol{d} < 0$.*

Finally, we report the definitions of convex, globally and locally Lipschitz and level bounded function.

Definition 1.3 *A continuously differentiable function $f : \Omega \to \mathbb{R}$ is said to be:*

(i) *convex if*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega; \tag{8}$$

(ii) *globally Lipschitz, if*

$$\|f(\boldsymbol{y}) - f(\boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega; \tag{9}$$

(iii) *locally Lipschitz, if for every compact set $K \subseteq \Omega$ there exists $L_K > 0$ such that*

$$\|f(\boldsymbol{y}) - f(\boldsymbol{x})\| \leq L_K\|\boldsymbol{y} - \boldsymbol{x}\| \quad \forall \boldsymbol{x}, \boldsymbol{y} \in K; \tag{10}$$

(iv) *level bounded, if the set $\Omega_\zeta = \{\boldsymbol{x} \in \Omega : f(\boldsymbol{x}) \leq \zeta\}$ is bounded for every $\zeta \in \mathbb{R}$.*

2. General results about Armijo based gradient projection methods

In this section, we recall the basic properties of the most popular linesearch procedure, the *Armijo linesearch*, given in Algorithm 1. These results allow to prove a general convergence result which applies to any method where the objective function over two successive iterates decreases at least as it would decrease by applying the Armijo linesearch procedure along a suitable descent direction.

For the linesearch procedure based on the Armijo rule we recall the following basic theorem, which can be derived from known results [27, 28].

Algorithm 1 Armijo linesearch (LS) algorithm

Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of points in Ω and $\{\boldsymbol{d}^{(k)}\}_{k\in\mathbb{N}}$ a sequence of descent directions. Choose some $\delta, \beta \in (0,1)$ and compute $\lambda^{(k)}$ as follows:

1. Set $\lambda^{(k)} = 1$

2. IF
$$f(\boldsymbol{x}^{(k)} + \lambda^{(k)}\boldsymbol{d}^{(k)}) \leq f(\boldsymbol{x}^{(k)}) + \beta\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} \tag{11}$$
    THEN go to step 3
    ELSE set $\lambda^{(k)} = \delta\lambda^{(k)}$ and go to step 2

3. END

Proposition 2.1 *Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of points in Ω. Assume that $\boldsymbol{x}^{(k)}$ converges to some $\bar{\boldsymbol{x}} \in \Omega$ and let $\{\boldsymbol{d}^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of descent directions such that*

$$\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} < 0 \quad \forall k \in \mathbb{N}. \tag{12}$$

Then the LS algorithm is well defined, i.e. for each $k \in \mathbb{N}$ it terminates in a finite number of steps. If, in addition, there exists a number $M > 0$ such that $\|\boldsymbol{d}^{(k)}\| \leq M \ \forall k \in \mathbb{N}$ and

$$\lim_{k\to\infty} f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k)} + \lambda^{(k)}\boldsymbol{d}^{(k)}) = 0, \tag{13}$$

where $\lambda^{(k)}$ is computed with Algorithm 1, then we have

$$\lim_{k\to\infty} \nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} = 0.$$

It is worth stressing that the previous proposition applies to every sequence $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{d}^{(k)}\}_{k\in\mathbb{N}}$ satisfying the assumptions of Proposition 2.1, not only for sequences defined as $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \lambda^{(k)}\boldsymbol{d}^{(k)}$.

A further direct consequence of the Armijo condition is the following lemma which will be used in the next section to prove the convergence of the scaled gradient projection method.

Lemma 2.1 *Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of points in Ω and $\{\boldsymbol{d}^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of descent directions such that condition (12) holds. Suppose that there exists $l \in \mathbb{R}$ such that $f(\boldsymbol{x}) \geq l$ for all $\boldsymbol{x} \in \Omega$ and that*

$$f(\boldsymbol{x}^{(k+1)}) \leq f(\boldsymbol{x}^{(k)} + \lambda^{(k)}\boldsymbol{d}^{(k)}) \quad \forall k \in \mathbb{N}. \tag{14}$$

Then we have

$$0 \leq -\sum_{k=0}^{\infty} \lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} < \infty. \tag{15}$$

Proof. Inequality (11) can be rewritten as

$$-\beta\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} \leq f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)}).$$

Summing the previous inequality for $k = 0, ..., j$ gives

$$\begin{aligned}
-\beta\sum_{k=0}^{j}\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} &\leq \sum_{k=0}^{j}(f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})) \\
&= f(\boldsymbol{x}^{(0)}) - f(\boldsymbol{x}^{(j+1)}) \\
&\leq f(\boldsymbol{x}^{(0)}) - l.
\end{aligned} \tag{16}$$

Thus, inequality (15) follows. □

The previous results hold, in general, for any sequence $\{\boldsymbol{d}^{(k)}\}_{k\in\mathbb{N}}$ of descent directions. In particular, if we choose as descent direction the vector defined in (2)–(3), then a further property of $\{\boldsymbol{d}^{(k)}\}_{k\in\mathbb{N}}$ holds true, as reported in the following lemma whose proof can be found in [12, Lemmata 2.2,2.3].

Lemma 2.2 *Let $\boldsymbol{x}^{(k)} \in \Omega$ and $\boldsymbol{d}^{(k)}$ be defined as in (2)–(3). Then we have*

$$\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} \leq -\frac{\|\boldsymbol{d}^{(k)}\|^2_{D_k^{-1}}}{\alpha_k}. \tag{17}$$

Moreover, $\boldsymbol{d}^{(k)} = \boldsymbol{0}$ if and only if $\boldsymbol{x}^{(k)}$ is stationary for problem (1).

We are now ready to give the more general convergence result based on the above mentioned properties of the descent direction and with the Armijo rule establishing the sufficient decrease of the objective function. Its proof is omitted since it can be easily derived by the analogous results in [11, 12, 23].

Theorem 2.1 *Let $\alpha_{\min}, \alpha_{\max}, \mu$ be three positive constants such that $0 < \alpha_{\min} \leq \alpha_{\max}$ and $\mu \geq 1$. Let $\{\alpha_k\}_{k\in\mathbb{N}} \subset [\alpha_{\min}, \alpha_{\max}]$ be a sequence of parameters and $\{D_k\}_{k\in\mathbb{N}} \subset \mathcal{M}_\mu$. Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}} \subset \Omega$ be any sequence satisfying property (14), where $\boldsymbol{d}^{(k)}$ is defined in (2)–(3) and $\lambda^{(k)}$ is computed with the Armijo linesearch procedure in Algorithm 1. If $\bar{\boldsymbol{x}}$ is a limit point of $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$, then $\bar{\boldsymbol{x}}$ is a stationary point for problem (1).*

It is worth stressing that the previous result applies also to nonconvex problems and the gradient of the objective function is not required to be Lipschitz continuous. Moreover, the only limitations to the algorithms parameters choice are that α_k and the eigenvalues of D_k have to be bounded above and below away from zero.

When ∇f satisfies some Lipschitz property, the next proposition states that the Armijo steplengths are bounded away from zero. This also means that there exists a finite upper bound for the number of backtracking reductions at any iteration (a similar result can be found in [29, Theorem 3.2]). This result will be useful in the convergence rate analysis of the next section.

Proposition 2.2 *Assume that ∇f satisfies one of the following conditions:*

a) *∇f is globally Lipschitz on Ω;*

b) *∇f is locally Lipschitz and f is level bounded on Ω.*

Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ be any sequence satisfying the assumptions of Theorem 2.1 and $\{\lambda^{(k)}\}_{k\in\mathbb{N}}$ the related steplengths computed by Algorithm 1. Then, there exists a positive constant $0 < \lambda_{\min} \le 1$ such that

$$\lambda^{(k)} \ge \lambda_{\min}. \tag{18}$$

Proof. If ∇f is Lipschitz continuous on Ω with Lipschitz constant L, then from the descent lemma [27, p.667] we have

$$f(\boldsymbol{x}^{(k)} + \lambda\boldsymbol{d}^{(k)}) \le f(\boldsymbol{x}^{(k)}) + \lambda\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} + \frac{L}{2}\lambda^2\|\boldsymbol{d}^{(k)}\|^2, \tag{19}$$

where $\lambda \in [0,1]$.

If, instead, ∇f is only locally Lipschitz, by assumption f is level bounded; since (14) implies $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}} \subset \Omega_{f(\boldsymbol{x}^{(0)})}$, we have that $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ is bounded. Equation (17) implies $\|\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}\| \le \frac{2}{\mu\alpha_{\max}}\|\nabla f(\boldsymbol{x}^{(k)})\|$. Then, $\{\boldsymbol{y}^{(k)}\}_{k\in\mathbb{N}}$ is also bounded and there exists a compact set K containing the points $\boldsymbol{x}^{(k)} + \lambda(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})$ for any $\lambda \in [0,1]$ and any $k \in \mathbb{N}$. As a consequence of this, inequality (19) holds with $L = L_K$.

By inequalities (19) and (17) we further obtain

$$f(\boldsymbol{x}^{(k)} + \lambda\boldsymbol{d}^{(k)}) \le f(\boldsymbol{x}^{(k)}) + \lambda\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} - \frac{L}{\gamma}\lambda^2\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)}$$

$$= f(\boldsymbol{x}^{(k)}) + \lambda\left(1 - \frac{L}{\gamma}\lambda\right)\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)},$$

where $\gamma = \mu\alpha_{\max}$. The previous inequality ensures that the Armijo condition

$$f(\boldsymbol{x}^{(k)} + \lambda\boldsymbol{d}^{(k)}) \le f(\boldsymbol{x}^{(k)}) + \lambda\beta\nabla f(\boldsymbol{x}^{(k)})^T\boldsymbol{d}^{(k)} \tag{20}$$

is satisfied, for all $k \in \mathbb{N}$, when $(1 - L\lambda/\gamma) \ge \beta$, that is for all λ such that $\lambda \le (1-\beta)\gamma/L$. If $\lambda^{(k)}$ is the steplength computed by Algorithm 1 and the backtracking loop is performed at least once, then $\lambda = \lambda^{(k)}/\delta$ does not satisfies inequality (20), which means $\lambda^{(k)} > \gamma(1-\beta)\delta/L$. Thus, the steplength sequence $\{\lambda^{(k)}\}_{k\in\mathbb{N}}$ satisfies inequality (18) with $\lambda_{\min} = \min\{1, \gamma(1-\beta)\delta/L\}$. $\qquad\square$

3. Convergence analysis of the scaled gradient projection algorithm

In this section we consider the SGP method whose basic scheme is reported in Algorithm 2. Clearly, Theorem 2.1 applies also to Algorithm 2, establishing that any limit point of the sequence $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ is stationary. Our aim is to propose practical conditions for

selecting the SGP metric, i.e. the parameter μ_k, ensuring the convergence of the sequence $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ to a solution of (1), under the only assumption that f is convex and admits a finite minimum.

The same conditions allow us also to prove a $\mathcal{O}(1/k)$ complexity result for SGP, which holds when the gradient of f satisfies some Lipschitz assumption.

Algorithm 2 Scaled gradient projection (SGP) method

Choose $0 < \alpha_{\min} \le \alpha_{\max}$, $\mu \ge 1$, $\delta, \beta \in (0,1)$, $\boldsymbol{x}^{(0)} \in \Omega$.

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

1. Choose $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$;

2. Choose $\mu_k \le \mu$ and a positive definite matrix $D_k \in \mathcal{M}_{\mu_k}$;

3. Compute $\boldsymbol{y}^{(k)}$ as in (3);

4. Set $\boldsymbol{d}^{(k)} = \boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}$;

5. Compute the steplength parameter $\lambda^{(k)}$ with Algorithm 1;

6. Set $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \lambda^{(k)}\boldsymbol{d}^{(k)}$.

Before to give the main convergence result, we prove the following lemma.

Lemma 3.1 *Let $\{\mu_k\}_{k\in\mathbb{N}}$, $\{\zeta_k\}_{k\in\mathbb{N}}$ be two sequences of numbers such that*

$$\mu_k^2 = 1 + \zeta_k, \quad \zeta_k \ge 0, \quad \sum_{k=0}^{\infty} \zeta_k < \infty. \tag{21}$$

Then the sequence $\{\theta_k\}_{k\in\mathbb{N}}$, with $\theta_k = \prod_{j=0}^{k} \mu_j^2$, is bounded.

Proof. We want to show that there exists a constant $M > 0$ such that $\theta_k \le M$ for all $k \in \mathbb{N}$. By the monotonicity of the logarithm, this is true if and only if $\log(\theta_k) \le \log(M)$ $\forall k \in \mathbb{N}$. By definition of θ_k we have

$$\log(\theta_k) = \sum_{j=0}^{k} \log(\mu_j^2) \le \sum_{j=0}^{\infty} \log(\mu_j^2). \tag{22}$$

Thus, if the series on the right hand side of (22) converges, the quantities θ_k are bounded for all k. We observe that, since $\mu_j^2 = 1 + \zeta_j$, by the known limit $\lim_{\zeta_j \to 0} \log(1+\zeta_j)/\zeta_j = 1$, the series $\sum_{j=0}^{\infty} \log(\mu_j^2)$ and $\sum_{j=0}^{\infty} \zeta_j$ have the same behaviour. Thus, since by hypothesis the latter one is convergent, the theorem follows. $\qquad\square$

The next theorem states that, when f is convex and admits finite minimum, if the scaling matrices D_k asymptotically reduce to the identity matrix at a certain rate, then the sequence generated by SGP converges to a solution of (1). The line of the proof is similar

to that of [30, Theorem 1], which can be considered as a special case of it. After giving the proof of our result, we discuss the relations of our approach with the related work already present in the literature.

Theorem 3.1 *Assume that the objective function of (1) is convex and the solution set X^* is not empty. Let $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ be the sequence generated by SGP where $D_k \in \mathcal{M}_{\mu_k}$ and $\{\mu_k\}_{k\in\mathbb{N}}$ satisfies (21). Then the sequence $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ converges to a solution of (1).*

Proof. We recall first the basic norm equality

$$\|\boldsymbol{x} - \boldsymbol{y}\|_E^2 + \|\boldsymbol{y} - \boldsymbol{z}\|_E^2 - \|\boldsymbol{x} - \boldsymbol{z}\|_E^2 = 2(\boldsymbol{y} - \boldsymbol{x})^T E(\boldsymbol{y} - \boldsymbol{z}) \tag{23}$$

which holds true for any positive definite matrix E. Moreover, it is easy to see that if $D_k \in \mathcal{M}_{\mu_k}$, then $D_k^{-1} \in \mathcal{M}_{\mu_k}$.
Let $\hat{\boldsymbol{x}} \in X^*$. By definition of $\boldsymbol{y}^{(k)}$ we have

$$(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)} + \alpha_k D_k \nabla f(\boldsymbol{x}^{(k)}))^T D_k^{-1}(\boldsymbol{x} - \boldsymbol{y}^{(k)}) \geq 0 \quad \forall \boldsymbol{x} \in \Omega$$

which, for $\boldsymbol{x} = \hat{\boldsymbol{x}}$ gives

$$(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})^T D_k^{-1}(\hat{\boldsymbol{x}} - \boldsymbol{x}^{(k)}) \geq \alpha_k \nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}})$$
$$+ (\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)} + \alpha_k D_k \nabla f(\boldsymbol{x}^{(k)}))^T D_k^{-1}(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})$$
$$\geq \alpha_k(f(\boldsymbol{x}^{(k)}) - f(\hat{\boldsymbol{x}})) + \|\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}\|_{D_k^{-1}}^2 + \alpha_k \nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})$$
$$= \alpha_k(f(\boldsymbol{x}^{(k)}) - f(\hat{\boldsymbol{x}})) + \frac{1}{(\lambda^{(k)})^2}\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|_{D_k^{-1}}^2 + \alpha_k \nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}),$$

where the inequality follows from the convexity of f and the last equality by definition of $\boldsymbol{x}^{(k+1)}$. By equality (23) with $\boldsymbol{x} = \boldsymbol{x}^{(k+1)}$, $\boldsymbol{y} = \boldsymbol{x}^{(k)}$, $\boldsymbol{z} = \hat{\boldsymbol{x}}$, $E = D_k^{-1}$ we obtain

$$\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 = \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 + \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|_{D_k^{-1}}^2 - 2(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k+1)})^T D_k^{-1}(\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}})$$
$$= \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 + \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|_{D_k^{-1}}^2 - 2\lambda^{(k)}(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})^T D_k^{-1}(\hat{\boldsymbol{x}} - \boldsymbol{x}^{(k)})$$
$$\leq \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 + \left(1 - \frac{2}{\lambda^{(k)}}\right)\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|_{D_k^{-1}}^2 - 2\alpha_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})$$
$$- 2\lambda^{(k)}\alpha_k(f(\boldsymbol{x}^{(k)}) - f(\hat{\boldsymbol{x}}))$$

which, since $\lambda^{(k)} \leq 1$, results in

$$\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 \leq \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 - 2\alpha_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}) +$$
$$- 2\lambda^{(k)}\alpha_k(f(\boldsymbol{x}^{(k)}) - f(\hat{\boldsymbol{x}})) \tag{24}$$
$$\leq \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 - 2\alpha_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}).$$

(since $f(\boldsymbol{x}^{(k)}) - f(\hat{\boldsymbol{x}}) \geq 0$). From the last inequality and in view of (7), it follows that

$$\frac{1}{\mu_k}\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 \leq \|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2$$
$$\leq \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 - 2\alpha_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})$$
$$\leq \mu_k\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}),$$

that is

$$\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 \leq \mu_k^2 \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|^2 - 2\mu_k\alpha_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}).$$

Recalling that the scalar product at the right-hand-side is nonpositive, since $\mu_k \geq 1$ and $\alpha_k \leq \alpha_{\max}$ this results in

$$\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 \leq \mu_k^2 \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}\mu_k^2\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}).$$

By repeatedly applying the previous inequality we obtain

$$\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 \leq \theta_0^k\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}\sum_{j=0}^{k}\theta_j^k\lambda^{(j)}\nabla f(\boldsymbol{x}^{(j)})^T(\boldsymbol{y}^{(j)} - \boldsymbol{x}^{(j)}),$$

where $\theta_j^k = \prod_{i=j}^{k}\mu_j^2$. Since $\mu_j^2 \geq 1$, we have $\theta_j^k \leq \theta_0^k$, and by Lemma 3.1 we obtain

$$\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 \leq M\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}M\sum_{j=0}^{k}\lambda^{(j)}\nabla f(\boldsymbol{x}^{(j)})^T(\boldsymbol{y}^{(j)} - \boldsymbol{x}^{(j)}), \quad (25)$$

where $\theta_0^k \leq M$. Now we can apply Lemma 2.1 to conclude that $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ is bounded and, thus, it has at least one limit point. Let us denote such limit point by $\boldsymbol{x}^\infty$. By Theorem 2.1, $\boldsymbol{x}^\infty$ is stationary; in particular, since f is convex, it is a minimum point, i.e. $\boldsymbol{x}^\infty \in X^*$. Let $\{\boldsymbol{x}^{(k_i)}\}_{i\in\mathbb{N}}$ be a subsequence of $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ which converges to $\boldsymbol{x}^\infty$. By applying the same arguments employed to derive (25), for any fixed $i \in \mathbb{N}$ and for all $k \geq k_i$ we obtain

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\infty\|^2 \leq M\|\boldsymbol{x}^{(k_i)} - \boldsymbol{x}^\infty\|^2 - 2\alpha_{\max}M\sum_{j=k_i}^{k}\lambda^{(j)}\nabla f(\boldsymbol{x}^{(j)})^T(\boldsymbol{y}^{(j)} - \boldsymbol{x}^{(j)}). \quad (26)$$

Since $\{\boldsymbol{x}^{(k_i)}\}_{i\in\mathbb{N}}$ converges to $\boldsymbol{x}^\infty$ and $-\sum_{j=0}^{\infty}\lambda^{(j)}\nabla f(\boldsymbol{x}^{(j)})^T(\boldsymbol{y}^{(j)} - \boldsymbol{x}^{(j)})$ is a convergent series, for any $\varepsilon > 0$ there exists a sufficiently large integer k_i such that $\|\boldsymbol{x}^{(k_i)} - \boldsymbol{x}^\infty\|^2 \leq \varepsilon/2M$ and $-\sum_{j=k_i}^{k}\lambda^{(j)}\nabla f(\boldsymbol{x}^{(j)})^T(\boldsymbol{y}^{(j)} - \boldsymbol{x}^{(j)}) \leq \varepsilon/(4M\alpha_{\max})$. Then, it follows from (26) that $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\infty\|^2 \leq \varepsilon$ for all $k \geq k_i$. Since ε can be chosen arbitrarily small, this means that the whole sequence $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ converges to $\boldsymbol{x}^\infty$. $\qquad\square$

The previous theorem gives an easily implementable rule to ensure the theoretical convergence of SGP to a solution. Moreover, as shown in section 4, this fact seems to have a favourable impact also on the practical performances of the method. This result is also coherent with the conclusions drawn from the numerical experience in [31], where the advantages of using a scaling matrix multiplying the gradient were observed mainly at the initial iterations.

Finally, we discuss our contribution with respect to recent, related works. In particular, in [13] the authors analyze the convergence of method (5) applied to problem (4), proving the convergence of the iterates when f is convex and has Lipschitz continuous gradient. In their analysis, they assume that λ_k, α_k can be chosen in suitable intervals whose bounds

depends on the Lipschitz constant of ∇f and, moreover, the matrices sequence $\{D_k\}_{k\in\mathbb{N}}$ satisfies

$$(1+\zeta_k)D_{k+1} \succeq D_k, \tag{27}$$

where ζ_k is a summable sequence (see also [32]).

Another forward–backward method based on variable metric and described by iteration (5) is proposed in [14], where the authors perform the convergence analysis when f has Lipschitz continuous gradient and $f+g$ satisfies the Kurdyka-Łojasiewicz property. The convergence of the iterates to a stationary point is proved when α_k, λ_k are chosen to ensure a suitable descent of the objective function, while D_k fulfills the condition

$$f(\boldsymbol{x}) \leq f(\boldsymbol{x}^{(k)}) + \nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{x}-\boldsymbol{x}^{(k)}) + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|^2_{D_k^{-1}}, \quad \forall \boldsymbol{x} \in \mathrm{dom}(g). \tag{28}$$

An inexact computation of the proximal point (6) is also allowed.

Variable metrics were considered also in [33, Chapter 5] in the context of subgradient methods for nonsmooth, convex, unconstrained minimization. In this case, setting $D_k = B_k B_k^T$, the scaling matrices are assumed to satisfy $\prod_{k=0}^{\infty}\|B_{k+1}^{-1}B_k\|^2 < \infty$ and

$$\|B_{k+1}^{-1}B_k\| \geq 1. \tag{29}$$

We remark that our condition, $D_k \in \mathcal{M}_{\mu_k}$, is quite different from both (28) and (29), since it does not impose a strict constraint on the scaling matrix choice. This freedom of choosing the metric at each iteration allows for example to adopt a suitable adaptation of a well performing scaling technique, based on a gradient splitting [2, 34], which may lead to significant improvements of the convergence behaviour, as we will show in section 4.

As a concluding remark, we observe that in our approach and in the above mentioned ones there are no assumptions about the structure of the matrices D_k. However, the computation of the projection (3) (and of the analogous proximal point (6) in the more general case) may become impractical unless D_k is a "simple" matrix, e.g. diagonal, even if Ω has a favourable shape. Indeed, in our analysis we assume that the exact solution of the problem in (3) is available. Future work will address the possibility to extend our results when an inexact computation of (3) is allowed, as in [11, 14].

In the following we give a complexity result about SGP, showing that it has a $\mathcal{O}(1/k)$ convergence rate on the objective function value. Similar results can be found in [9] for forward–backward methods with linesearch along the projection arc (i.e. of the form (5) with $D_k = I$, $\lambda_k = 1$ for all k and with α_k determined by a backtracking procedure).

Theorem 3.2 *Assume that the hypotheses of Theorem 3.1 hold and, in addition, that assumption a) or b) of Proposition 2.2 is satisfied. Let f^* be the optimal function value for problem (1). Then, we have*

$$f(\boldsymbol{x}^{(k)}) - f^* = \mathcal{O}(1/k).$$

Proof. Setting $a = 2\lambda_{\min}\alpha_{\min}$, where $\lambda_{\min}$ is defined in Proposition 2.2, from (24) we have

$$
\begin{aligned}
\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 &\leq \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 - 2\alpha_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}) + \\
&\quad - 2\lambda^{(k)}\alpha_k(f(\boldsymbol{x}^{(k)}) - f(\hat{\boldsymbol{x}})) \\
&\leq \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 - 2\alpha_{\max}\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}) + \\
&\quad + a(f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^{(k)})),
\end{aligned}
$$

where the second inequality follows from the fact that $\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)})$ and $f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^{(k)})$ are negative quantities. Thanks to inequality (7), we can write

$$
\begin{aligned}
\frac{1}{\mu_k}\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 &\leq \|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 \\
&\leq \|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|_{D_k^{-1}}^2 - 2\alpha_{\max}\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}) + \\
&\quad + a(f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^{(k)})) \\
&\leq \mu_k\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}) + \\
&\quad + a(f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^{(k)})).
\end{aligned}
$$

By multiplying the last inequality by μ_k we obtain

$$
\begin{aligned}
\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 &\leq \mu_k^2\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}\mu_k\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}) + \\
&\quad + \mu_k a(f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^{(k)})) \\
&\leq \mu_k^2\|\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}\mu_k^2\lambda^{(k)}\nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}) + \\
&\quad + a(f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^{(k)})),
\end{aligned}
$$

where the last inequality follows from the fact that $\mu_k \geq 1$. By repeatedly applying the last inequality we obtain

$$
\begin{aligned}
\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 &\leq \theta_0^k\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}\sum_{j=0}^{k}\theta_j^k\lambda^{(j)}\nabla f(\boldsymbol{x}^{(j)})^T(\boldsymbol{y}^{(j)} - \boldsymbol{x}^{(j)}) + \\
&\quad + a\left((k+1)f(\hat{\boldsymbol{x}}) - \sum_{j=0}^{k}f(\boldsymbol{x}^{(j)})\right) \\
&\leq M\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\|^2 - 2\alpha_{\max}M\sum_{j=0}^{k}\lambda^{(j)}\nabla f(\boldsymbol{x}^{(j)})^T(\boldsymbol{y}^{(j)} - \boldsymbol{x}^{(j)}) + \\
&\quad + a\left((k+1)f(\hat{\boldsymbol{x}}) - \sum_{j=0}^{k}f(\boldsymbol{x}^{(j)})\right), \quad\quad (30)
\end{aligned}
$$

where, as in the proof of Theorem 3.1, we set $\theta_j^k = \prod_{i=j}^{k}\mu_j^2$ and M is the upper bound of all θ_j^k. Thanks to inequality (16), we have

$$
\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 \leq M\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\|^2 + \frac{2\alpha_{\max}M}{\beta}(f(\boldsymbol{x}^{(0)}) - f(\hat{\boldsymbol{x}})) +
$$

$$+ a(kf(\hat{\boldsymbol{x}}) - \sum_{j=1}^{k} f(\boldsymbol{x}^{(j)})), \tag{31}$$

where we also added the positive quantity $a(f(\boldsymbol{x}^{(0)}) - f(\hat{\boldsymbol{x}}))$ to the right hand side of (30). Moreover, exploiting the inequality

$$0 \le \sum_{j=0}^{k} j(f(\boldsymbol{x}^{(j)}) - f(\boldsymbol{x}^{(j+1)})) = \sum_{j=1}^{k} f(\boldsymbol{x}^{(j)}) - kf(\boldsymbol{x}^{(k+1)}) \tag{32}$$

gives

$$\|\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}\|^2 \le M\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\|^2 + \frac{2\alpha_{\max}M}{\beta}(f(\boldsymbol{x}^{(0)}) - f(\hat{\boldsymbol{x}})) +$$
$$+ ak(f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^{(k+1)})).$$

Rearranging terms, this finally yields

$$f(\boldsymbol{x}^{(k+1)}) - f(\hat{\boldsymbol{x}}) \le \frac{M}{ak}\left(\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\|^2 + 2\frac{\alpha_{\max}}{\beta}(f(\boldsymbol{x}^{(0)}) - f(\hat{\boldsymbol{x}}))\right),$$

establishing the result. □

In the recent literature, several authors developed the so-called *intertial* methods, which are first order methods including an extrapolation step which allows to prove a $\mathcal{O}(1/k^2)$ convergence rate on the objective function values (see for example [9, 10, 35, 36]). However, as we will show in section 4, the practical performances of SGP can be comparable with those of $\mathcal{O}(1/k^2)$ methods, even if the theoretical convergence rate estimate is only $\mathcal{O}(1/k)$.

4. Numerical illustration

In this section we consider some relevant applications and we show that they can be effectively solved by algorithms which can be framed in the analysis of the previous sections. We give also some hints on how to choose the parameters α_k and D_k at each iteration, even if a specific treatment of this issue is far beyond the scope of this paper. Both sets of numerical tests concern the image deconvolution problem in the presence of Poisson noise. In particular, in the next subsection we will consider a fit-to-data + regularization model with an arbitrarily fixed regularization parameter, while in the following tests we will investigate the same problem combined with an automatic procedure for the choice of this parameter recently proposed by Zanni et al. [37].

4.1. Edge preserving image restoration

Our basic assumption is that the available data $\boldsymbol{g} \in \mathbb{R}^n$ is a realization of a Poisson random variable whose mean is $A\boldsymbol{x}^* + b\boldsymbol{e}$, where $A \in \mathbb{R}^{n \times n}$ is a structured matrix representing

the convolution operator, $b \in \mathbb{R}$ is a positive parameter representing the background radiation, $\boldsymbol{e} \in \mathbb{R}^n$ is the vector of all ones and $\boldsymbol{x}^*$ is the image we would like to recover. In the following, we will assume that $A\boldsymbol{e} = \boldsymbol{e}$, $A^T\boldsymbol{e} = \boldsymbol{e}$, which is not a restrictive assumption, since it can be assured by a simple normalization.

According to the Bayesian approach [38], an approximation $\boldsymbol{x}_\nu$ of $\boldsymbol{x}^*$ can be obtained by solving the following optimization problem

$$\min_{\boldsymbol{x} \geq 0} f(\boldsymbol{x}) \equiv KL(\boldsymbol{x}) + \nu R(\boldsymbol{x}), \tag{33}$$

where $KL(\boldsymbol{x})$ is the generalized Kullback–Leibler divergence

$$KL(\boldsymbol{x}) = \sum_{i=1}^{n} \left\{ g_i \log\left(\frac{g_i}{(A\boldsymbol{x})_i + b} \right) + (A\boldsymbol{x})_i + b - g_i \right\}, \tag{34}$$

$R(\boldsymbol{x})$ is some regularization functional, chosen according to the a priori information on the desired solution, and $\nu > 0$ is the regularization parameter balancing the relative weight of the two terms. In order to preserve the edges in the restored image, a good choice for the regularization term is the following hypersurface (HS) functional [39, 40]

$$HS_\rho(\boldsymbol{x}) = \sum_{i=1}^{n} \sqrt{(\boldsymbol{D}_i^h \boldsymbol{x})^2 + (\boldsymbol{D}_i^v \boldsymbol{x})^2 + \rho^2}, \tag{35}$$

where $\rho > 0$ and $\boldsymbol{D}_i^h$, $\boldsymbol{D}_i^v$ are finite difference approximations of the horizontal and vertical image gradient, respectively. If ρ is small, it can be considered as an approximation of the total variation functional, but it has been shown that better reconstructions can be obtained for large values of the smoothing parameter [41]. Thus, we consider the following convex optimization problem

$$\min_{\boldsymbol{x} \geq 0} f(\boldsymbol{x}) \equiv KL(\boldsymbol{x}) + \nu HS_\rho(\boldsymbol{x}), \tag{36}$$

whose main features have been studied in [42] (see also [43]).

As for the SGP method, borrowing the ideas in [20], at each iteration k we adopt the following diagonal scaling matrix

$$[D_k]_{ii} = \max\left\{ \frac{1}{\mu_k}, \min\left\{ \mu_k, \frac{\boldsymbol{x}^{(k)}}{1 + \nu V_i^R(\boldsymbol{x}^{(k)})} \right\} \right\}, \qquad i = 1, \ldots, n, \tag{37}$$

where $V_i^R(\boldsymbol{x}^{(k)})$ is defined as in [20, formula (25)], while $\mu_k = \sqrt{1 + 10^{10}/k^2}$ so that Theorem 3.1 applies. The steplength parameter α_k is then computed in two different ways:

- the adaptive alternation of the scaled Barzilai–Borwein (BB) rules as proposed in [12];
- the Ritz-like values proposed by Fletcher [19] for a steepest descent method in the case of unconstrained optimization and recently extended to the SGP algorithm applied to a general constrained problem (1) [44, 45].

Besides SGP, we consider also for comparison the "plain" gradient projection (GP) method with Euclidean projection and variable steplength (chosen with the same two rules exploited in the scaled case), the PidSplit+ algorithm [46], which is an alternating direction method of multipliers specific for the minimization of the Kullback–Leibler plus the discrete total variation functional ($\rho = 0$), adapted to the smoothed case with $\rho > 0$, and the accelerated proximal-gradient method with inertial/extrapolation with backtracking (FISTA-b) [9].

As test problems, we consider:

- the Shepp-Logan (SL) phantom of size 256×256, multiplied by a factor of 500, corrupted with Gaussian blur of variance 9 and with Poisson noise simulated using the `imnoise` Matlab function on the blurred image including the additive background. The background constant is $b = 10$;
- the confocal microscopy (CM) phantom of size 128×128 described in [47, section V.C], with values in the range $[0, 68]$ and with a constant background $b = 1$.

We assume periodic boundary conditions, so that the matrix A is block circulant with circulant blocks (BCCB) and the matrix-vector products involving A can be performed with a $\mathcal{O}(n \log(n))$ complexity by means of the fast Fourier transform [48]. In figure 1 we report the original objects, the corrupted images and the solutions $\boldsymbol{x}^*$ of problem (33) for both test problems. The parameters (ν, ρ) in (36) have been empirically tuned to obtain a visually satisfactory solution and have been set equal to $(0.0415, 1)$ for SL and $(0.06, 1)$ for CM. Moreover, the 'γ' parameter of PidSplit+ has been set equal to $50/\nu$ (SL) and $1/\nu$ (CM) and the initial steplength parameter for FISTA-b is 100 in both cases. In order to illustrate the convergence behaviour of the methods, we first compute a ground truth solution $\boldsymbol{x}_\nu$ (see figure 1, right panel) by running 1500 iterations of SGP. Then, we evaluate the progress towards this solution by computing at each iterate the relative difference of the objective function value with respect to the estimated minimum $f(\boldsymbol{x}_\nu)$ (see figure 2). We include in our comparison also the version of SGP with fixed bounds on the scaling matrix $\mu_k = \mu = 10^5$, which is denoted by SGP*. From figure 2 we can observe what follows:

- the choice of a suitable projection operator can have a significant impact on the practical performances of the gradient projection methods, since GP is outperformed by SGP with both choices for the steplength parameters;
- SGP with variable bounds on the scaling matrix gives the best performances: in particular, condition (21), which is employed in Theorem 3.1 to prove the convergence of the method on convex problems, seems to significantly improve its practical performances by exploiting the benefits of the scaling matrix at the beginning of the iterative procedure. On the contrary, when the iterates are close to the limit point, the method behaves like a nonscaled algorithm allowing a further faster decrease of
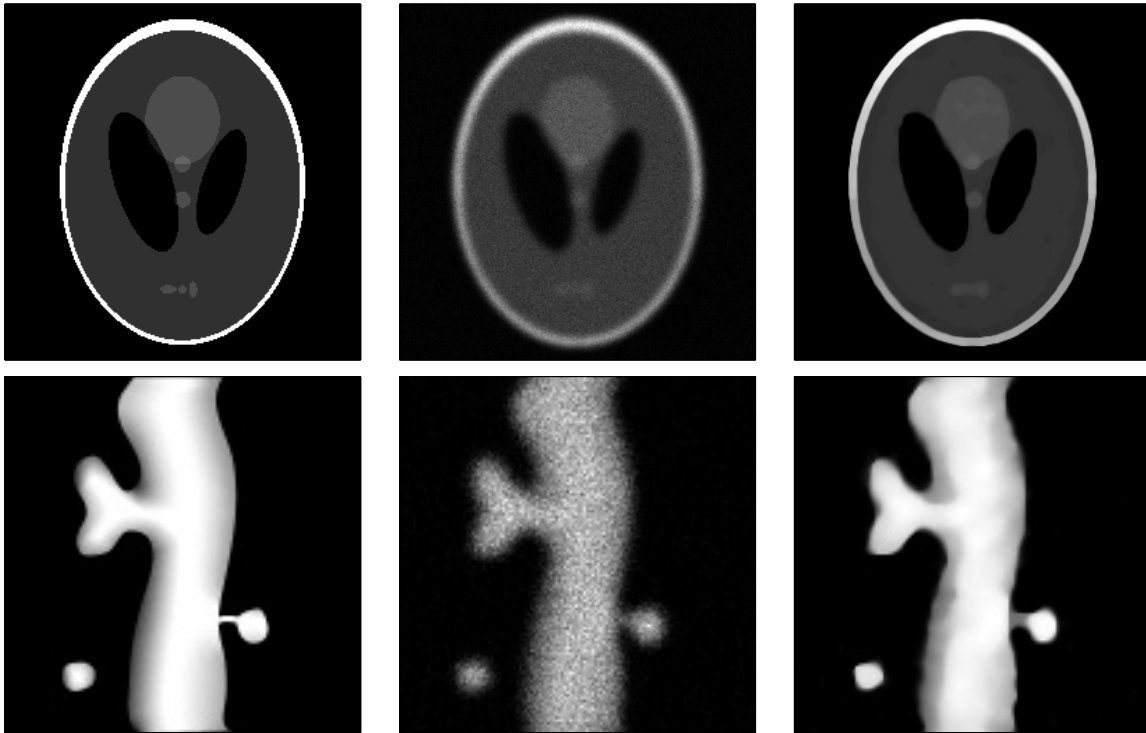
Figure 1. Test problems: Shepp-Logan (top) and confocal microscopy (bottom) phantoms. Original image (left), noisy blurred image (middle) and optimal solution (right).

the objective function with respect to the one endowed with a variable metric with fixed bounds;

- in spite of the theoretical convergence rate given in Theorem 3.2, the practical behaviour of SGP is comparable with the $\mathcal{O}(1/k^2)$ method FISTA-b.

4.2. Automatic parameter estimation

The choice of the regularization parameter in Poisson data inversion is an active field and several different strategies have been proposed in the last years [49, 40, 50, 51, 52]. Here we consider that proposed by Bertero et al. [40], which consists of selecting the value of ν in (33) such that

$$\mathcal{D}_A(\boldsymbol{x}_\nu) = \eta, \tag{38}$$

where

$$\mathcal{D}_A(\boldsymbol{x}_\nu) = \frac{2}{n} KL(\boldsymbol{x}_\nu)$$

and η is a given number close to 1 [40, 53] (here we will assume $\eta = 1$). In particular, in [37] the authors introduced an effective secant-type solver for the discrepancy equation
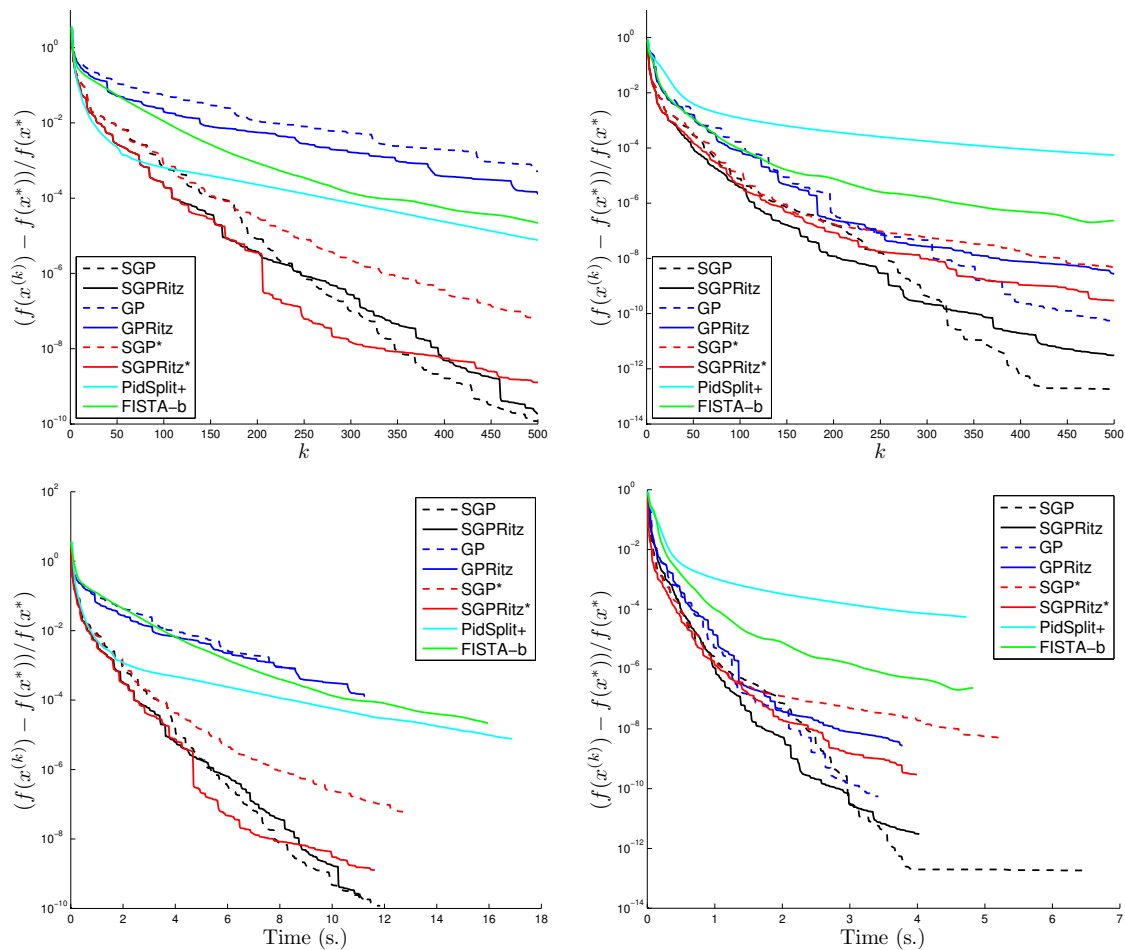
Figure 2. Image deconvolution: objective function decrease in logarithmic scale versus the iterations number (top) and computational time (bottom) for the SL (left) and CM (right) datasets.

(38), called modified Dai-Fletcher (MDF) method, able to reduce the number of required solutions of problems (33). At each step of the secant method, an approximation of the solution of problem (33) for a given value of ν is provided by running an optimization method until the stopping criterium

$$|f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k-1)})| \leq \varepsilon |f(\boldsymbol{x}^{(k)})|, \tag{39}$$

where $\varepsilon = 5 \times 10^{-8}$, is satisfied or when a maximum number of iterations equal to 5000 is reached. In this section we consider again the KL + HS model (36) and we investigate the impact of (some of) the strategies used for the previous tests within this automatic scheme for the choice of ν. In particular, we restrict our analysis to the GP, SGP* and SGP methods equipped with the Ritz-like steplengths, and the PidSplit+ algorithm with the adaptive choice of its parameter γ described in [37, equation (24)], which resulted to

be less dependent on the parameter settings than the standard approach.

The test problems we considered are based on the Satellite dataset already used in several papers and available at www.mathcs.emory.edu/∼nagy/RestoreTools/index.html. The original image is sized 256×256 and assumes values in the range $[0, 2550]$. The blurred image has been obtained by convolving the object with a point spread function simulating a ground-based telescope response, and a constant background $b = 10$ has been added to the resulting image before introducing Poisson noise. Two further datasets have been obtained by multiplying object and background by factors of 10 and 100 before the blurring step. The three test sets will be denoted by S2550, S25500 and S255000 and the corrupted images are shown in figure 3 together with the original one. As concerns the parameter ρ defining the HS regularization term, we followed the suggestion in [37] and set $\rho = 10^{-4} \max(\boldsymbol{g})$.
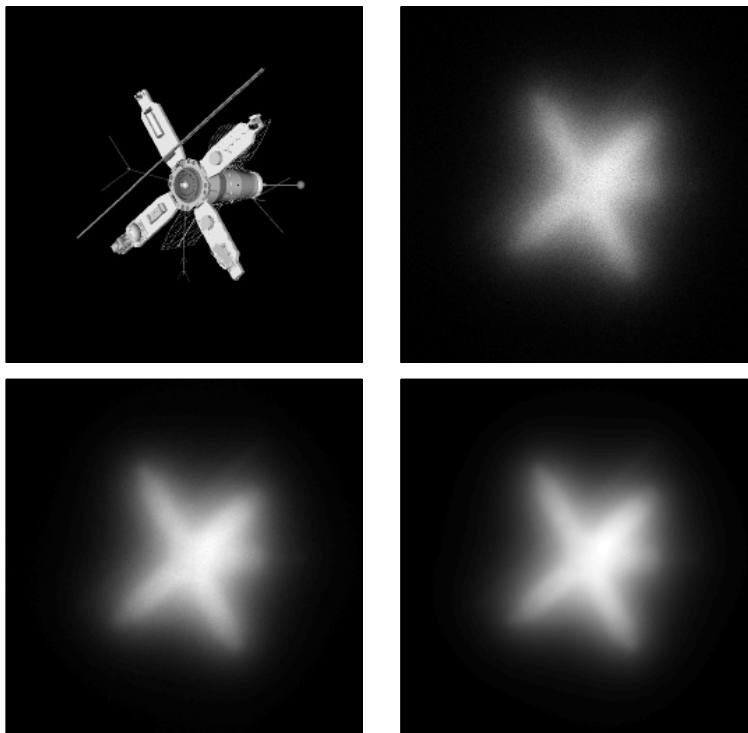

Figure 3. Satellite test problems: original object (top left), S2550 (top right), S25500 (bottom left) and S255000 (bottom right) blurred and noisy images.

The results obtained by the algorithms are shown in table 1, where we reported the number of steps k of the secant-based method required to satisfy either the relation

$$|\mathcal{D}_A(\boldsymbol{x}_{\nu_k}) - \eta| \leq \varepsilon_1$$

or both the inequalities

$$|\nu_k - \nu_{k-1}| \le \varepsilon_2 \nu_k \qquad ; \qquad |\mathcal{D}_A(\boldsymbol{x}_{\nu_k}) - \eta| \le 10\varepsilon_1,$$

being $\varepsilon_1 = 5 \times 10^{-4}$ and $\varepsilon_2 = 5 \times 10^{-3}$, the total number of iterations k_{tot} performed by each method in the k steps, the final regularization parameter ν_k, the relative reconstruction error between $\boldsymbol{x}_{\nu_k}$ and $\boldsymbol{x}^*$ and the execution time in seconds. These numerical experiments has been carried out on a Dual CPU Intel(R) Xeon(R) X5690 at 3.47GHz with 188 GB RAM (see also fermi.unife.it) in a Matlab2013a environment.

Table 1. Results obtained with GP, SGP*, SGP and PidSlit+ on the three Satellite datasets. Here k denotes the number of steps of the secant-based method proposed in [37], k_{tot} the total number of iterations performed by each method in the k steps, ν_k the final regularization parameter, "err" the relative reconstruction error between $\boldsymbol{x}_{\nu_k}$ and $\boldsymbol{x}^*$ and "time" the execution time in seconds.

Test problem	Algorithm	k	k_{tot}	ν_k	err	time
S2550	PidSplit+	8	1801	5.42e-05	0.322	67.84
	GPRitz	16	3297	2.39e-04	0.308	56.55
	SGPRitz*	13	3091	6.29e-04	0.301	51.36
	SGPRitz	7	2223	5.68e-04	0.305	36.52
S25500	PidSplit+	17	3486	1.00e-06	0.280	131.8
	GPRitz	41	10388	1.00e-41	0.433	193.0
	SGPRitz*	15	7753	8.27e-05	0.257	130.3
	SGPRitz	13	6189	9.91e-05	0.260	113.9
S255000	PidSplit+	13	4108	7.97e-07	0.238	145.6
	GPRitz	40	31150	1.00e-11	0.777	602.5
	SGPRitz*	16	10851	8.09e-06	0.230	189.7
	SGPRitz	7	5341	1.00e-05	0.237	105.6

The performances summarized in table 1 confirm what already observed in the previous section, since SGP equipped with the scaling matrices with variable bounds succeeds in reducing the overall number of iterations required to provide the regularization parameter and the corresponding reconstruction if compared with SGP with fixed bounds for the scaling matrices or GP (which, in two of the three tests, often fails in satisfying the stopping criterium (39) within the maximum number of iterations allowed). As concerns the comparison with PidSplit+, we can observe that the number of iterations performed by this latter strategy is lower than that of SGP, but the higher cost per iteration which characterizes PidSplit+ makes the procedure more expensive in terms of total CPU time with respect to the SGP method.

5. Conclusions

In this paper we revisited the SGP method, originally published in 2009 and exploited in the successive years in several inverse problems as image denoising/deblurring, Fourier-based image reconstruction, blind deconvolution, system identification and non-negative matrix factorization, with several applications in astronomy, microscopy and engineering. Despite all the good numerical results provided in solving these problems, the only theoretical convergence result proved so far is the stationarity of any limit point of the sequence generated by SGP. In this paper we showed that stronger results can be proved in the convex case, if the sequence of scaling matrices characterizing the SGP iterations is chosen as convergent to the identity matrix at a certain rate. Moreover, in the same setting we provided also a convergence rate estimate on the objective function values, as provided in the literature for several other optimization methods. Some numerical tests showed that the specific rule introduced on the scaling matrices to prove the theoretical convergence results helps also to improve the performances of the method, making SGP competitive also with methods for which the $\mathcal{O}(1/k^2)$ convergence rate has been demonstrated.

Acknowledgments

References

[1] J. Bardsley and J. Nagy. Covariance-preconditioned iterative methods for nonnegatively constrained astronomical imaging. *SIAM J. Matrix Anal. A.*, 27(4):1184–1198, 2006.

[2] M. Bertero, H. Lantéri, and L. Zanni. Iterative image reconstruction: a point of view. In Y. Censor, M. Jiang, and A. K. Louis, editors, *Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT)*, pages 37–63. Birkhauser-Verlag, Pisa, Italy, 2008.

[3] M. Figueiredo, R. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal Process.*, 1(4):586–597, December 2007.

[4] C. J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.*, 19(10):2756–2779, October 2007.

[5] T. Serafini, G. Zanghirati, and L. Zanni. Gradient projection methods for quadratic programs and applications in training support vector machines. *Optim. Methods Softw.*, 20(2–3):353–378, January 2005.

[6] T. Serafini and L. Zanni. On the working set selection in gradient projection-based decomposition techniques for support vector machines. *Optim. Methods Softw.*, 20(4–5):583–596, August 2005.

[7] L. Zanni. An improved gradient projection-based decomposition technique for support vector machines. *Comput. Manag. Sci.*, 3(2):131–145, April 2006.

[8] D. Bertsekas. *Convex optimization theory. Supplementary Chapter 6 on convex optimization algorithms*. Athena Scientific, Belmont, MA, 2 december 2013 edition, 2009.

[9] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[10] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, May 2005.

[11] 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.

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

[13] P. L. Combettes and B. C. Vũ. Variable metric forward-backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):17–31, September 2014.

[14] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forwardbackward algorithm for minimizing the sum of a differentiable function and a convex function. *J. Optim. Theory Appl.*, 162(1):107–132, July 2014.

[15] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *J. Opt. Theory Appl.*, 165(3):874–900, June 2015.

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

[17] Y. H. Dai, W. W. Hager, K. Schittkowski, and H. Zhang. The cyclic Barzilai-Borwein method for unconstrained optimization. *IMA J. Numer. Anal.*, 26(3):604–627, July 2006.

[18] R. De Asmundis, D. di Serafino, F. Riccio, and G. Toraldo. On spectral properties of steepest descent methods. *IMA J. Numer. Anal.*, 33(4):1416–1435, October 2013.

[19] R. Fletcher. A limited memory steepest descent method. *Math. Program.*, 135(1–2):413–436, October 2012.

[20] 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.

[21] S. Bonettini and M. Prato. Nonnegative image reconstruction from sparse Fourier data: a new deconvolution algorithm. *Inverse Probl.*, 26(9):095001, September 2010.

[22] M. Prato, R. Cavicchioli, L. Zanni, P. Boccacci, and M. Bertero. Efficient deconvolution methods for astronomical imaging: algorithms and IDL-GPU codes. *Astron. Astrophys.*, 539:A133, March 2012.

[23] S. Bonettini. Inexact block coordinate descent methods with application to the nonnegative matrix factorization. *IMA J. Numer. Anal.*, 31(4):1431–1452, October 2011.

[24] S. Bonettini, A. Chiuso, and M. Prato. A scaled gradient projection method for Bayesian learning in dynamical systems. *SIAM J. Sci. Comput.*, 37(3):A1297–A1318, 2015.

[25] S. Bonettini, A. Cornelio, and M. Prato. A new semiblind deconvolution approach for Fourier-based image restoration: an application in astronomy. *SIAM J. Imaging Sci.*, 6(3):1736–1757, 2013.

[26] M. Prato, A. La Camera, S. Bonettini, and M. Bertero. A convergent blind deconvolution method for post-adaptive-optics astronomical imaging. *Inverse Probl.*, 29(6):065017, June 2013.

[27] D. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, 1999.

[28] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Oper. Res. Lett.*, 26(3):127–136, April 2000.

[29] A. Auslender, P. J. S. Silva, and M. Teboulle. Nonmonotone projected gradient methods based on barrier and Euclidean distances. *Comput. Optim. Appl.*, 38(3):305–327, December 2007.

[30] A. N. Iusem. On the convergence properties of the projected gradient method for convex optimization. *Comput. Optim. Appl.*, 22(1):37–52, 2003.

[31] S. Bonettini, G. Landi, E. Loli Piccolomini, and L. Zanni. Scaling techniques for gradient projection-type methods in astronomical image deblurring. *Int. J. Comput. Math.*, 90(1):9–29, January 2013.

[32] P. L. Combettes and B. C. Vũ. Variable metric quasi-Féjer monotonicity. *Nonlinear Anal.-Theor.*, 78:17–31, February 2013.

[33] A. Nedić. *Subgradient methods for convex minimization.* PhD thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 2002.

[34] H. Lantéri, M. Roche, and C. Aime. Penalized maximum likelihood image restoration with positivity constraints: multiplicative algorithms. *Inverse Probl.*, 18(5):1397–1419, October 2002.

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

[36] 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.

[37] L. Zanni, A. Benfenati, M. Bertero, and V. Ruggiero. Numerical methods for parameter estimation in poisson data inversion. *J. Math. Imaging Vis.*, 52(3):397–413, July 2015.

[38] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984.

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

[40] M. Bertero, P. Boccacci, G. Talenti, R. Zanella, and L. Zanni. A discrepancy principle for Poisson data. *Inverse Probl.*, 26(10):105004, October 2010.

[41] S. Bonettini and V. Ruggiero. An alternating extragradient method for total variation based image restoration from Poisson data. *Inverse Probl.*, 27(9):095001, September 2011.

[42] S. Bonettini and V. Ruggiero. On the uniqueness of the solution of image reconstruction problems with Poisson data. In T. E. Simos, G. Psihoyios, and Ch. Tsitouras, editors, *International Conference of Numerical Analysis and Applied Mathematics 2010*, volume 1281 of *AIP Conf. Proc.*, pages 1803–1806, 2010.

[43] N. Pustelnik, C. Chaux, and J.-C. Pesquet. Parallel proximal algorithm for image restoration using hybrid regularization. *IEEE Trans. Image Process.*, 20(9):2450–2462, sep. 2011.

[44] F. Porta, R. Zanella, G. Zanghirati, and L. Zanni. Limited-memory scaled gradient projection methods for real-time image deconvolution in microscopy. *Commun. Nonlinear Sci. Numer. Simul.*, 21(1–3):112–127, April 2015.

[45] F. Porta, M. Prato, and L. Zanni. A new steplength selection for scaled gradient methods with application to image deblurring. *J. Sci. Comput.*, 65(3):895–919, December 2015.

[46] S. Setzer, G. Steidl, and T. Teuber. Deblurring Poissonian images by split Bregman techniques. *J. Vis. Commun. Image R.*, 21(3):193–199, April 2010.

[47] R. M. Willet and R. D. Nowak. Platelets: A multiscale approach for recovering edges and surfaces in photon limited medical imaging. *IEEE Trans. Med. Imaging*, 22(3):332–350, March 2003.

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

[49] J. M. Bardsley and J. Goldes. Regularization parameter selection methods for ill-posed poisson maximum likelihood estimation. *Inverse Probl.*, 25(9):095005, September 2009.

[50] M. Carlavan and L. Blanc-Féraud. Regularizing parameter estimation for Poisson noisy image restoration. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, pages 597–601, 2011.

[51] M. Carlavan and L. Blanc-Féraud. Sparse Poisson noisy image deblurring. *IEEE Trans. Image Process.*, 21(4):1834–1846, April 2012.

[52] T. Teuber, G. Steidl, and R. H. Chan. Minimization and parameter estimation for seminorm regularization models with I-divergence constraints. *Inverse Probl.*, 29(3):035007, March 2013.

[53] S. Bonettini and M. Prato. Accelerated gradient methods for the X-ray imaging of solar flares. *Inverse Probl.*, 30(5):055004, May 2014.