SIAM J. OPTIM.
Vol. 26, No. 3, pp. 1741–1772

SCALING TECHNIQUES FOR ϵ-SUBGRADIENT METHODS*

S. BONETTINI†, A. BENFENATI†, AND V. RUGGIERO†

Abstract. The recent literature on first order methods for smooth optimization shows that significant improvements on the practical convergence behavior can be achieved with variable step size and scaling for the gradient, making this class of algorithms attractive for a variety of relevant applications. In this paper we introduce a variable metric in the context of the ϵ-subgradient methods for nonsmooth, convex problems, in combination with two different step size selection strategies. We develop the theoretical convergence analysis of the proposed approach in the general framework of forward-backward ϵ-subgradient splitting methods and we also discuss practical implementation issues. In order to illustrate the effectiveness of the method, we consider a specific problem in the image restoration framework and we numerically evaluate the effects of a variable scaling and of the step length selection strategy on the convergence behavior.

Key words. forward-backward ϵ-subgradient method, variable metric, step size selection rules, scaled primal-dual hybrid gradient algorithm, TV restoration

AMS subject classifications. 65K05, 90C25

DOI. 10.1137/14097642X

1. Introduction.

Several models arising in relevant applications such as image and signal restoration, statistical inference, and data analysis lead to the following constrained optimization problem

$$\min_{x\in X} f(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex, proper, lower semicontinuous function and X is a nonempty, closed, convex subset of $\mathbb{R}^n$ contained in the domain of f. We denote by X^* the set of solutions of (1.1). We are interested in the case where f is nondifferentiable and a subgradient or an approximate subgradient of f can be easily computed. This arises for example in duality and minimax contexts. A well-known method to solve problem (1.1) is the ϵ-subgradient projection method

$$x^{(k+1)} = P_X\left(x^{(k)} - \alpha_k u^{(k)}\right), \tag{1.2}$$

where $u^{(k)} \in \partial_{\epsilon_k} f(x^{(k)})$ for some $\epsilon_k \geq 0$, α_k is a positive step size, and $P_X(\cdot)$ is the Euclidean projection operator onto the set X. The choice $\epsilon_k = 0$ for all k corresponds to the subgradient method, which has been extensively investigated (see, for example, the contributions collected in [34, 39, 56, 63]).

The more general case allowing $\epsilon_k > 0$ was introduced and developed in [34, 56], while more recent convergence results under different assumptions are given in [1, 27,

*Received by the editors July 7, 2014; accepted for publication (in revised form) June 7, 2016; published electronically September 1, 2016.

http://www.siam.org/journals/siopt/26-3/97642.html

Funding: The research of the authors was partially supported by MIUR (Italian Ministry for University and Research), under the projects FIRB–Futuro in Ricerca 2012, contract RBFR12M3AC, and by the Italian Spinner2013 PhD project *High-complexity inverse problems in biomedical applications and social systems*. The Italian GNCS–INdAM (Gruppo Nazionale per il Calcolo Scientifico–Istituto Nazionale di Alta Matematica) is also acknowledged.

†Dipartimento di Matematica e di Informatica, Università di Ferrara, Polo Scientifico Tecnologico, Blocco B, I-44122 Ferrara, Italy (silvia.bonettini@unife.it, alessandro.benfenati@unife.it, valeria.ruggiero@unife.it).

47, 50, 58, 64]. A typical assumption on the sequence $\{\epsilon_k\}$ is that

$$\lim_{k\to\infty} \epsilon_k = 0, \tag{1.3}$$

and, in this case, the subgradient and the ϵ-subgradient methods have very similar convergence properties. In the following discussion we assume that (1.3) holds.

The ϵ-subgradient method is interesting in itself since, when the projection onto X is easy to compute and an approximate subgradient is available, it can be easily implemented, but it is also a useful tool to analyze the theoretical convergence properties of a variety of algorithms [16, 35, 55, 58].

We can distinguish different variants of the method (1.2) according to the rule adopted to select the step size. We list below the most studied step size choices for subgradient methods which, with minor modifications, could also be applied to the case $\epsilon_k \geq 0$:

($\mathcal{R}_1$) the *constant step size* rule $\alpha_k = \alpha > 0$;

($\mathcal{R}_2$) the *Polyak rule*

$$\alpha_k = c_k \frac{f(x^{(k)}) - f^*}{\|u^{(k)}\|^2} \quad \text{or} \quad \alpha_k = c_k \frac{f(x^{(k)}) - f^*}{\max\{1, \|u^{(k)}\|^2\}}, \quad c_k \in (0, 2),$$

where $f^* = \inf_{x\in X^*} f(x)$;

($\mathcal{R}_3$) the *Ermoliev* or *diminishing, divergent series* step size rule, which includes any sequences $\{\alpha_k\}$ such that

$$\alpha_k > 0, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty; \tag{1.4}$$

($\mathcal{R}_4$) the *diminishing, divergent series, square summable* step size rule which, in addition to (1.4), also requires $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$;

($\mathcal{R}_5$) the *dynamic* or *adaptive* step size rule

$$\alpha_k = \frac{f(x^{(k)}) - f_k}{\|u^{(k)}\|^2} \quad \text{or} \quad \alpha_k = \frac{f(x^{(k)}) - f_k}{\max\{1, \|u^{(k)}\|^2\}}, \tag{1.5}$$

where f_k is an adaptively computed estimate of f^*; several further variants of this rule, which can be considered as an approximation of ($\mathcal{R}_2$) when f^* is not known, depend on how f_k is defined.

Keeping a constant step size as in ($\mathcal{R}_1$), only the convergence of a subsequence of $\{f(x^{(k)})\}$ to a possibly suboptimal value is established, i.e., $\liminf_k f(x^{(k)}) \leq f^* + C\alpha$, for some positive constant C [10, 53]; for rules ($\mathcal{R}_3$) and ($\mathcal{R}_5$) stronger results have been proved [39, 50, 53], showing that $\lim_{k\to\infty} f(x^{(k)}) = f^*$ and, if $X^* \neq \emptyset$, $\min_{x^*\in X^*} \|x^* - x^{(k)}\| \to 0$ (with the assumption $\epsilon_k = 0$ for the latter case). Finally, if $X^* \neq \emptyset$, the convergence of the sequence $\{x^{(k)}\}$ to a solution of (1.1) can be proved in the cases ($\mathcal{R}_2$) with $\epsilon_k = 0$, and ($\mathcal{R}_4$) [1, 50].

A number of variants to accelerate the subgradient iteration have been investigated (see, for instance, [45, 46, 55, 31]). Results about an optimal step size choice to obtain a suboptimal rate of convergence for the subgradient method are reported in [54]; in [10, section 6.3], a convergence analysis is performed for different step size choices while in [3] analogous results are obtained with respect to non-Euclidean metrics.

The key property that the step size parameter has to induce on the iterates (1.2) which is exploited in the standard convergence analysis for subgradient methods is the *quasi-Féjer monotonicity* with respect to the set X^*,

$$\|x^{(k+1)} - x^*\|^2 \leq \|x^{(k)} - x^*\|^2 + \eta_k \quad \forall x^* \in X^*$$

for some nonnegative sequence $\{\eta_k\}$ such that $\sum \eta_k < \infty$ (see [1, 7, 23, 28]).

It is worth noticing that the step size in subgradient methods plays quite a different role than in the smooth case, where an analogous parameter is employed to ensure the sufficient decrease of the objective function and, in some kind of schemes, also to accelerate the convergence, for example, by means of the well-known Barzilai–Borwein rules [6, 30, 38] (see also [2, 37] for recent developments in this field). Thus, these valid approaches to the step size selection for the smooth case are difficult to extend to the method (1.2).

On the other side, recent advances in the context of gradient-based methods show that introducing a variable scaling matrix for the gradient can lead to significant improvements in the practical performances [5, 18, 12, 41, 65], especially on large scale and ill-conditioned problems. Variable metric was also introduced in [25] in the context of monotone operators and in [52, Chapter 5] in the subgradient methods for unconstrained optimization.

Motivated by this, we propose to introduce a variable scaling matrix for the ϵ-subgradient vector in the iteration (1.2) and we analyze the convergence properties of the resulting method. For the sake of generality, we perform our analysis from the point of view of a forward-backward ϵ-subgradient method that includes (1.2) as a special case. Indeed, we consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) + \Phi(x), \tag{1.6}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex, proper, lower semicontinuous function whose domain is contained in the domain of f. Problem (1.6) reduces to (1.1) by setting $\Phi(x) = \iota_X(x)$, where $\iota_X(x)$ is the indicator function of the set X, i.e.,

$$\iota_X(x) = \begin{cases} 0 & \text{if } x \in X, \\ +\infty & \text{if } x \notin X. \end{cases}$$

In the frame of problem (1.6), the ϵ-subgradient projection method (1.2) can be viewed as a special case of the general forward-backward ϵ-subgradient scheme

$$x^{(k+1)} = \operatorname{prox}_{\alpha_k \Phi, I}\left(x^{(k)} - \alpha_k u^{(k)}\right) \equiv \operatorname*{argmin}_{x \in \mathbb{R}^n}\left(\Phi(x) + \frac{1}{2\alpha_k}\|x - (x^{(k)} - \alpha_k u^{(k)})\|^2\right), \tag{1.7}$$

where $u^{(k)} \in \partial_{\epsilon_k} f(x^{(k)})$ for some $\epsilon_k \geq 0$, α_k is a positive step size, and $\|\cdot\|$ is the Euclidean vector norm. We recall that, given a symmetric positive definite matrix D, $\operatorname{prox}_{\Phi, D^{-1}}(x)$ is the so-called proximity operator of $\Phi(x)$ relative to the metric introduced by D^{-1} (see [44, section XV.4]), defined as

$$\operatorname{prox}_{\Phi, D^{-1}}(x) = \operatorname*{argmin}_{z \in \mathbb{R}^n}\left(\Phi(z) + \frac{1}{2}\|z - x\|^2_{D^{-1}}\right),$$

where $\|y\|^2_{D^{-1}} = y^T D^{-1} y$.

The main contribution of this paper is to provide the convergence analysis, under standard assumptions, of the following scaled forward-backward ϵ-subgradient scheme

$$x^{(k+1)} = \text{prox}_{\alpha_k \Phi, D_k^{-1}} \left(x^{(k)} - \alpha_k D_k u^{(k)} \right), \tag{1.8}$$

where D_k is a symmetric positive definite matrix with bounded eigenvalues and α_k is chosen either as an a priori selected sequence obeying the diminishing, divergent series, summable squares step size rule ($\mathcal{R}_4$), or with an adaptive rule ($\mathcal{R}_5$) of Brännlund's type [19, 40, 53].

We point out that the convergence of a forward-backward subgradient method, named proximal subgradient splitting (PSS) method, is discussed in [28]. Here, we introduce the use of an ϵ-subgradient for f in the forward step and a variable metric in the backward or proximal step. In particular, when α_k is chosen as in ($\mathcal{R}_4$), assuming that the set X^* is nonempty, we prove the convergence of the sequence $\{x^{(k)}\}$ to a point $x^* \in X^*$, providing also a convergence rate estimate, while, for α_k chosen by an adaptive rule, we prove that $\liminf_{k\to\infty}\{f(x^{(k)})\} + \Phi(x^{(k)})$ is equal to the optimal value $f^* = \min_{x\in\mathbb{R}^n} f(x) + \Phi(x)$. When $X^* = \emptyset$, we prove that $\liminf_{k\to\infty} f(x^{(k)}) + \Phi(x^{(k)}) = \inf_{x\in\mathbb{R}^n} f(x) + \Phi(x)$.

A further contribution of the paper is to introduce, as a special case of (1.8) for problem (1.1) and also as a generalization of the method in [16], a scaled primal–dual hybrid gradient (SPDHG) method, which applies to the case

$$\min_{x\in\mathbb{R}^n} f_0(x) + f_1(Ax) + \Phi(x), \tag{1.9}$$

where f_0 and f_1 are convex, proper, lower semicontinuous functions and A is a linear operator.

When f_0 is continuously differentiable with Lipschitz continuous gradient, suitable splitting and forward-backward methods can be applied to (1.9) [20, 25, 26, 51]. Very recently, in [13] and [29], some variants of forward-backward methods with line-search techniques were proposed for differentiable f_0 without any Lipschitz continuity assumption on the gradient.

As we will show in section 5, problem (1.9) can be handled also by SPDHG, even when f_0 is nondifferentiable or its gradient is not Lipschitz continuous on $\text{dom}(f_0)$.

In particular, we provide an especially tailored implementation of SPDHG for the total variation (TV) restoration of images corrupted by Poisson noise. This problem is related to several applications such as astronomical imaging, electronic microscopy, single particle emission computed tomography, and positron emission tomography, and a variety of specialized methods have been proposed for its solution (see [4, 15, 32, 33, 36, 61] and references therein). For this special case of SPDHG, we devise an effective strategy to choose the scaling matrix, showing that significant improvements on the practical convergence speed can be obtained.

The paper is organized as follows. In section 2 we introduce some preliminary results that will be used in the subsequent sections and we present the complexity analysis of iteration (1.8). In section 3 we present a convergence analysis for the scaled forward-backward ϵ-subgradient method (1.8) when the step sizes are chosen according to a diminishing, divergent series, square summable step size rule. At the end of the section our results are also compared to the very recent works [24, 25], where variable metrics are studied from the point of view of more general operators.

Building on this material, in section 4 we propose a generalization to variable scaling and approximate subgradients of the *level algorithm* in [40, 53], based on a

dynamic step size selection rule, showing that in our more general settings the main properties still hold. Further, in section 5 we consider problem (1.9) and we present the SPDHG method, proving its convergence as a special case of an ϵ-subgradient scheme. In order to illustrate a practical implementation of SPDHG, in section 6 we describe the problem of deblurring an image corrupted by Poisson noise via the TV regularization. A suitable scaling for the SPDHG method is discussed and an algorithm for its computation is detailed. In section 7, we describe some numerical simulations concerning the considered application, with the aim to evaluate the effectiveness of the scaling technique in the ϵ-subgradient method in combination with the two step size selection strategies analyzed in the previous sections. The numerical experiments show that a suitable selection of the scaling matrix can also be a very effective tool to improve the convergence behavior in nonsmooth methods. Finally, some concluding remarks are given in section 8.

Notations and definitions. In the following, $\|\cdot\|$ denotes the Euclidean vector or matrix norm. Given $x \in \mathbb{R}^n$ and a symmetric and positive definite matrix D of order n, $\|x\|_D$ denotes the energy norm, i.e., $\|x\|_D = \sqrt{x^T D x}$. By $\mathrm{dom}(f)$ we indicate the domain of any function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, i.e., $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : f(x) < \infty\}$, while $\mathrm{diam}(X)$ denotes the diameter of the closed, convex set $X \subset \mathbb{R}^n$, $\mathrm{diam}(X) = \max_{x,z \in X} \|x - z\|$. The Fenchel dual or conjugate of f is defined as $f^*(y) = \sup_{x \in \mathbb{R}^n} x^T y - f(x)$. Furthermore, we recall that the ϵ-subdifferential of f at $x \in \mathrm{dom}(f)$ for some $\epsilon \in \mathbb{R}$, $\epsilon \geq 0$, is the set

$$\partial_\epsilon f(x) = \{p \in \mathbb{R}^n : \ f(z) \geq f(x) + p^T(z - x) - \epsilon, \ \ \forall z \in \mathbb{R}^n\}.$$

Any vector $p \in \partial_\epsilon f(x)$ is an ϵ-subgradient of f at x [59, section 23]. For $\epsilon = 0$ the standard subdifferential is recovered, i.e., $\partial_0 f(x) = \partial f(x)$.

If $f(x) = \sum_{i=1}^n \beta_i f_i(x)$, where $\beta_i \geq 0$, $u_i \in \partial_{\epsilon_i} f_i(x)$, and $x \in \bigcap_{i=1}^n \mathrm{dom}(f_i)$, then $\sum_{i=1}^n \beta_i u_i \in \partial_\epsilon f(x)$, where $\epsilon = \sum_{i=1}^n \epsilon_i$. The proof of this property can be found in [16, 59].

2. Assumptions and preliminary results. Given a symmetric, positive definite matrix D and the positive parameter γ, we consider the proximity operator of a convex, proper, and lower semicontinuous function $\gamma\Phi$ with respect to the metric induced by D^{-1},

$$\mathrm{prox}_{\gamma\Phi, D^{-1}}(x) = \operatorname*{argmin}_z \left(\Phi(z) + \frac{1}{2\gamma} \|z - x\|^2_{D^{-1}} \right), \tag{2.1}$$

and we recall some simple properties.

The operator $\mathrm{prox}_{\gamma\Phi, D^{-1}}$ is well-definite, since the solution of the strongly convex minimization problem in (2.1) exists and is unique. The optimality conditions of this problem can be written also as

$$D^{-1} \frac{(x - \mathrm{prox}_{\gamma\Phi, D^{-1}}(x))}{\gamma} \in \partial\Phi(\mathrm{prox}_{\gamma\Phi, D^{-1}}(x)). \tag{2.2}$$

From the previous inclusion combined with the convexity of Φ at $\mathrm{prox}_{\gamma\Phi, D^{-1}}(x)$, we obtain that the following inequality holds for any $x, z \in \mathbb{R}^n$,

$$(\mathrm{prox}_{\gamma\Phi, D^{-1}}(x) - x)^T D^{-1} (z - \mathrm{prox}_{\gamma\Phi, D^{-1}}(x)) \geq \gamma(\Phi(\mathrm{prox}_{\gamma\Phi, D^{-1}}(x)) - \Phi(z)).$$

Consequently, it is immediate to verify that the proximal operator is nonexpansive with respect to the energy norm, i.e., for any $x, z \in \mathbb{R}^n$ we have

$$\|\mathrm{prox}_{\gamma\Phi, D^{-1}}(x) - \mathrm{prox}_{\gamma\Phi, D^{-1}}(z)\|_{D^{-1}} \leq \|x - z\|_{D^{-1}}. \tag{2.3}$$

Finally, we observe that if L is a positive number such that $\|D\| = \lambda_{\max}(D) \leq L$ and $\|D^{-1}\| = \frac{1}{\lambda_{\min}(D)} \leq L$, then, for any $z, x \in \mathbb{R}^n$, we can write

$$\|\text{prox}_{\gamma\Phi,D^{-1}}(x) - \text{prox}_{\gamma\Phi,D^{-1}}(z)\| \leq L\|x - z\|. \tag{2.4}$$

Indeed, since for any vector $z \in \mathbb{R}^n$ we have

$$\frac{1}{L}\|z\|^2 \leq \lambda_{\min}(D^{-1})\|z\|^2 \leq \|x\|^2_{D^{-1}} \leq \lambda_{\max}(D^{-1})\|z\|^2 \leq L\|x\|^2, \tag{2.5}$$

the inequality (2.4) follows from (2.3).

We now report two technical lemmas: the first one concerns sequences of positive numbers, while the latter one states a crucial inequality about iteration (1.8) which will be extensively used in the analysis performed in the next sections.

LEMMA 2.1. *Let $\{L_k\}$ be a sequence of positive numbers such that $1 \leq L_k^2 \leq 1+\gamma_k$, $\gamma_k \geq 0$, and define*

$$\theta_j^k = \prod_{i=j}^{k} L_i^2 \quad \textit{and} \quad \tilde{\theta}_j^k = \theta_j^k / L_j. \tag{2.6}$$

If $\sum_{k=0}^{\infty} \gamma_k < \infty$, then there exist two constants, $L, M \geq 1$, such that, for all $k \geq 0$ we have

$$\begin{gathered} 1 \leq L_k \leq L, \\ \tilde{\theta}_j^k \leq \theta_j^k \leq \theta_0^k \leq M, \quad 0 \leq j \leq k. \end{gathered}$$

Proof. Since $\sum_{k=0}^{\infty} \gamma_k < \infty$, we have $\lim_{k\to\infty} \gamma_k = 0$, thus the sequence $\{L_k\}$ is bounded above by some constant $L \geq 1$. Moreover, we observe that, since $L_k \geq 1$ we have $\theta_{j-1}^k \leq \theta_j^k$ and $\tilde{\theta}_j^k \leq \theta_j^k$. Finally, the rightmost inequality follows by observing that $\theta_0^k \leq \prod_{j=0}^{k}(1+\gamma_j) = \exp(\log(\prod_{j=0}^{k}(1+\gamma_j))) \leq \exp(\sum_{j=0}^{k}\gamma_j) \leq \exp(\sum_{j=0}^{\infty}\gamma_j)$. ∎

LEMMA 2.2. *Let $\{x^{(k)}\}$ be the sequence generated by iteration (1.8), where $u^{(k)} \in \partial_{\epsilon_k} f(x^{(k)})$ for a given sequence $\{\epsilon_k\} \subset \mathbb{R}$, $\epsilon_k \geq 0$. Then, for any $x \in dom(\Phi)$ and for all $k \geq 0$, we have*

$$\begin{aligned} \|x^{(k+1)} - x\|^2_{D_k^{-1}} \leq \|x^{(k)} - x\|^2_{D_k^{-1}} + 2\alpha_k(f(x) + \Phi(x) \\ - (f(x^{(k)}) + \Phi(x^{(k)})) + \epsilon_k) + \alpha_k^2\|u^{(k)} + w^{(k)}\|^2_{D_k} \end{aligned} \tag{2.7}$$

for any $w^{(k)} \in \partial\Phi(x^{(k)})$.

Moreover, if there exists a sequence of positive numbers $\{L_k\}$ such that $\|D_k\| \leq L_k$, $\|D_k^{-1}\| \leq L_k$, it follows that

$$\begin{aligned} \|x^{(k+1)} - x\|^2 \leq L_k^2\|x^{(k)} - x\|^2 + 2\alpha_k L_k(f(x) + \Phi(x) \\ - (f(x^{(k)}) + \Phi(x^{(k)})) + \epsilon_k) + L_k\alpha_k^2\|u^{(k)} + w^{(k)}\|^2_{D_k}. \end{aligned} \tag{2.8}$$

Proof. From the optimality conditions (2.2) related to the iteration in (1.8), we have that there exists $\tilde{w}^{(k+1)} \in \partial\Phi(x^{(k+1)})$, such that

$$\tilde{w}^{(k+1)} = D_k^{-1}\frac{x^{(k)} - x^{(k+1)}}{\alpha_k} - u^{(k)}. \tag{2.9}$$

Consequently, we can write

$$\begin{aligned}\|\tilde{w}^{(k+1)}+u^{(k)}\|_{D_k}^2 &= \frac{1}{\alpha_k^2}(x^{(k)}-x^{(k+1)})D_k^{-1}D_kD_k^{-1}(x^{(k)}-x^{(k+1)})\\ &= \frac{1}{\alpha_k^2}\|x^{(k)}-x^{(k+1)}\|_{D_k^{-1}}^2.\end{aligned} \tag{2.10}$$

Then, for any $x\in\operatorname{dom}(\Phi)$, we have

$$\begin{aligned}
&\alpha_k^2\|u^{(k)}+\tilde{w}^{(k+1)}\|_{D_k}^2+\|x^{(k)}-x\|_{D_k^{-1}}^2-\|x^{(k+1)}-x\|_{D_k^{-1}}^2\\
&\quad=\|x^{(k)}-x^{(k+1)}\|_{D_k^{-1}}^2+\|x^{(k)}-x\|_{D_k^{-1}}^2-\|x^{(k+1)}-x\|_{D_k^{-1}}^2\\
&\quad=2(x^{(k)}-x)^TD_k^{-1}(x^{(k)}-x^{(k+1)})=2\alpha_k(x^{(k)}-x)^T(u^{(k)}+\tilde{w}^{(k+1)})\\
&\quad=2\alpha_k(x^{(k)}-x)^Tu^{(k)}+2\alpha_k(x^{(k)}-x^{(k+1)})^T\tilde{w}^{(k+1)}+2\alpha_k(x^{(k+1)}-x)^T\tilde{w}^{(k+1)}\\
&\quad=2\alpha_k(x^{(k)}-x)^Tu^{(k)}+2\alpha_k(x^{(k)}-x^{(k+1)})^T\left(D_k^{-1}\frac{x^{(k)}-x^{(k+1)}}{\alpha_k}-u^{(k)}\right)\\
&\qquad+2\alpha_k(x^{(k+1)}-x)^T\tilde{w}^{(k+1)}\\
&\quad=2\alpha_k(x^{(k)}-x)^Tu^{(k)}+2\|x^{(k)}-x^{(k+1)}\|_{D_k^{-1}}^2-2\alpha_k(x^{(k)}-x^{(k+1)})^Tu^{(k)}\\
&\qquad+2\alpha_k(x^{(k+1)}-x)^T\tilde{w}^{(k+1)}\\
&\quad\geq 2\alpha_k(f(x^{(k)})-f(x)-\epsilon_k)+2\|x^{(k)}-x^{(k+1)}\|_{D_k^{-1}}^2+2\alpha_k(x^{(k+1)}-x^{(k)})^Tu^{(k)}\\
&\qquad+2\alpha_k(\Phi(x^{(k+1)})-\Phi(x))\\
&\quad=2\alpha_k\Big(f(x^{(k)})+\Phi(x^{(k)})-(f(x)+\Phi(x))-\epsilon_k+\Phi(x^{(k+1)})-\Phi(x^{(k)})\\
&\qquad\qquad+(x^{(k+1)}-x^{(k)})^Tu^{(k)}\Big)+2\|x^{(k)}-x^{(k+1)}\|_{D_k^{-1}}^2\\
&\quad\geq 2\alpha_k\Big(f(x^{(k)})+\Phi(x^{(k)})-(f(x)+\Phi(x))-\epsilon_k+(x^{(k+1)}-x^{(k)})^T(w^{(k)}+u^{(k)})\Big)\\
&\qquad+2\|x^{(k)}-x^{(k+1)}\|_{D_k^{-1}}^2\\
&\quad=2\alpha_k\Big(f(x^{(k)})+\Phi(x^{(k)})-(f(x)+\Phi(x))-\epsilon_k+(x^{(k+1)}-x^{(k)})^T(w^{(k)}+u^{(k)})\Big)\\
&\qquad+2\alpha_k^2\|u^{(k)}+\tilde{w}^{(k+1)}\|_{D_k}^2,
\end{aligned} \tag{2.11}$$

where the first inequality follows from the definition of the ϵ-subgradient of f and from $\tilde{w}^{(k+1)}\in\partial\Phi(x^{(k+1)})$, while the second one holds for any $w^{(k)}\in\partial\Phi(x^{(k)})$. Finally, we use (2.10) in the last equality.

Thus, for $x\in\operatorname{dom}(\Phi)$, we can write

$$\begin{aligned}
\|x^{(k+1)}-x\|_{D_k^{-1}}^2 &\leq \|x^{(k)}-x\|_{D_k^{-1}}^2+2\alpha_k(f(x)+\Phi(x)-(f(x^{(k)})+\Phi(x^{(k)}))+\epsilon_k)\\
&\quad+2\alpha_k^2(u^{(k)}+\tilde{w}^{(k+1)})^TD_k(w^{(k)}+u^{(k)})-\alpha_k^2\|u^{(k)}+\tilde{w}^{(k+1)}\|_{D_k}^2\\
&=\|x^{(k)}-x\|_{D_k^{-1}}^2+2\alpha_k(f(x)+\Phi(x)-(f(x^{(k)})+\Phi(x^{(k)}))+\epsilon_k)\\
&\quad+\alpha_k^2\|u^{(k)}+w^{(k)}\|_{D_k}^2-\alpha_k^2\|w^{(k)}-\tilde{w}^{(k+1)}\|_{D_k}^2\\
&\leq\|x^{(k)}-x\|_{D_k^{-1}}^2+2\alpha_k(f(x)+\Phi(x)-(f(x^{(k)})+\Phi(x^{(k)}))+\epsilon_k)\\
&\quad+\alpha_k^2\|u^{(k)}+w^{(k)}\|_{D_k}^2,
\end{aligned}$$

concluding that inequality (2.7) holds for any $w^{(k)}\in\partial\Phi(x^{(k)})$. Finally, from (2.5), we can easily obtain (2.8). ∎

The previous lemma is analogous to [28, Lemma 2.2], which deals with the case $D_k = I$, $\epsilon_k = 0$. Similar inequalities are obtained also in [1, Lemma 1] for ϵ-subgradient projection methods, i.e., $D_k = I$, $\Phi = \iota_X$.

We now present and discuss the two basic assumptions that we make on the sequence $\{u^{(k)}\}$ and $\{D_k\}$ in iteration (1.8). In particular, except in Theorem 3.2, we will assume what follows.

A. There exist two positive constants ρ_u, ρ_w and a sequence $\{w^{(k)}\}$, $w^{(k)} \in \partial\Phi(x^{(k)})$ such that

$$\|u^{(k)}\| \le \rho_u \quad \text{and} \quad \|w^{(k)}\| \le \rho_w.$$

The assumption on $\{u^{(k)}\}$ is satisfied, for example, when $\operatorname{diam}(\operatorname{dom}(f^*))$ is finite. Indeed, it holds $\operatorname{dom}(f^*) = \bigcup_{x\in\mathbb{R}^n} \partial_\epsilon f(x)$ for every $\epsilon > 0$ (see [66, Remark 2]). Moreover, if $\operatorname{diam}(\operatorname{dom}(f^*)) = M$ for some $M > 0$, by definition of ϵ-subdifferential we obtain $|f(x) - f(z)| \le M\|x - z\| + \epsilon$. Since ϵ is arbitrary, it follows that f is Lipschitz continuous with constant M.

In order to guarantee the existence of a bounded sequence $\{w^{(k)}\}$, $w^{(k)} \in \partial\Phi(x^{(k)})$, in [28] it is assumed that $\partial\Phi(x)$ has bounded elements in $\operatorname{dom}(\Phi)$, i.e., for all $x \in \operatorname{dom}(\Phi)$ there exists $w \in \partial\Phi(x)$ such that $\|w\| \le \rho$ for some $\rho > 0$ independent of x. This is actually equivalent to assuming that Φ is Lipschitz continuous on its domain. Indeed, if Φ is Lipschitz continuous with constant C in $\operatorname{dom}(\Phi)$, then for all $x \in \operatorname{dom}(\Phi)$ and $w \in \partial\Phi(x)$ we have $\|w\| \le C$ [62, Lemma 2.6]. On the other side, if $\partial\Phi(x)$ has bounded elements in $\operatorname{dom}(\Phi)$, for any $x \in \operatorname{dom}(\Phi)$ we can pick a subgradient $w_x \in \partial\Phi(x)$ such that $\|w_x\| \le \rho$. From the definition of subgradient we obtain $\Phi(x) - \Phi(z) \le \rho\|x - z\|$ for all $x \in \operatorname{dom}(\Phi)$ and $z \in \mathbb{R}^n$. On the other side, for any $z \in \operatorname{dom}(\Phi)$ there exists $w_z \in \partial\Phi(z)$ such that $\|w_z\| \le \rho$, yielding $\Phi(z) - \Phi(x) \le \rho\|x - z\|$. Therefore, Φ is Lipschitz continuous with constant ρ in $\operatorname{dom}(\Phi)$.

Under assumption A, setting $\{(f+\Phi)^k_{\text{best}}\}$ as the smallest function value over the first k iterations,

$$(f+\Phi)^k_{\text{best}} = \min_{j=0,\dots,k} \{f(x^{(j)}) + \Phi(x^{(j)})\}, \tag{2.12}$$

we can prove the following inequalities, which are analogous to the estimates obtained in [28, Lemmas 2.2 and 2.3] for the case $D_k = I$, $\epsilon_k = 0$.

LEMMA 2.3. *Let $\{x^{(k)}\}$ be the sequence generated by iteration (1.8), where $u^{(k)} \in \partial_{\epsilon_k} f(x^{(k)})$ for a given sequence $\{\epsilon_k\} \subset \mathbb{R}$, $\epsilon_k \ge 0$. Let assumption A be satisfied and assume that there exists a sequence of positive numbers $\{L_k\}$ such that $\|D_k\| \le L_k$, $\|D_k^{-1}\| \le L_k$. If there exists a solution $\tilde{x}$ of (1.6), we have that*

$$(f+\Phi)^k_{\text{best}} - (f(\tilde{x}) + \Phi(\tilde{x})) \le \frac{\theta_0^k \|x^{(0)} - \tilde{x}\|^2 + 2\sum_{j=0}^k \tilde{\theta}_j^k \alpha_j \epsilon_j + \sigma \sum_{j=0}^k \theta_j^k \alpha_j^2}{2\sum_{j=0}^k \tilde{\theta}_j^k \alpha_j}, \tag{2.13}$$

where θ_j^k and $\tilde{\theta}_j^k$ are defined as in Lemma 2.1 and $\sigma = (\rho_u + \rho_w)^2$. Moreover, defining the ergodic sequence $\{\bar{x}^{(k)}\}$ of $\{x^{(k)}\}$ as

$$\bar{x}^{(k)} = \frac{\sum_{j=0}^k \tilde{\theta}_j^k \alpha_j x^{(j)}}{\sum_{j=0}^k \tilde{\theta}_j^k \alpha_j}, \tag{2.14}$$

we also have

(2.15)
$$(f+\Phi)(\bar{x}^{(k)}) - (f(\tilde{x})+\Phi(\tilde{x})) \le \frac{\theta_0^k \|x^{(0)}-\tilde{x}\|^2 + 2\sum_{j=0}^k \tilde{\theta}_j^k \alpha_j \epsilon_j + \sigma \sum_{j=0}^k \alpha_j^2 \theta_j^k}{2\sum_{j=0}^k \tilde{\theta}_j^k \alpha_j}.$$

Proof. Let $\{w^{(k)}\}$ be a sequence of subgradients of Φ as in assumption A. Then, we have

(2.16)
$$\|u^{(k)}+w^{(k)}\|_{D_k}^2 \le L_k \|u^{(k)}+w^{(k)}\|^2 \le L_k(\|u^{(k)}\|+\|w^{(k)}\|)^2 \le L_k(\rho_u+\rho_w)^2 \le L_k\sigma.$$

From (2.8) with $x=\tilde{x}$, we obtain

$$\begin{aligned}\|x^{(k+1)}-\tilde{x}\|^2 \le L_k^2\|x^{(k)}-\tilde{x}\|^2 &+ 2\alpha_k L_k(f(\tilde{x})+\Phi(\tilde{x}) \\ &- (f(x^{(k)})+\Phi(x^{(k)}))+\epsilon_k) + L_k^2\alpha_k^2\sigma.\end{aligned}$$

By repeatedly applying the previous inequality we further obtain

$$\begin{aligned}\|x^{(k+1)}-\tilde{x}\|^2 \le \theta_0^k\|x^{(0)}-\tilde{x}\|^2 &+ 2\sum_{j=0}^k \tilde{\theta}_j^k\alpha_j\epsilon_j + \sigma\sum_{j=0}^k \theta_j^k\alpha_j^2 \\ &+ 2\sum_{j=0}^k \tilde{\theta}_j^k\alpha_j(f(\tilde{x})+\Phi(\tilde{x})-(f(x^{(k)})+\Phi(x^{(k)}))).\end{aligned}$$

Rearranging terms and neglecting the negative ones on the right-hand side, the inequality above implies

(2.17)
$$2\sum_{j=0}^k \tilde{\theta}_j^k\alpha_j(f(x^{(j)})+\Phi(x^{(j)})-(f(\tilde{x})+\Phi(\tilde{x}))) \le \theta_0^k\|x^{(0)}-\tilde{x}\|^2 + 2\sum_{j=0}^k \tilde{\theta}_j^k\alpha_j\epsilon_j + \sigma\sum_{j=0}^k \alpha_j^2\theta_j^k.$$

From definition (2.12), we directly obtain (2.13). Finally, using the convexity of $f+\Phi$ and (2.14) in inequality (2.17), we immediately obtain (2.15). ∎

Here and in the following we will make the following assumption on the matrices D_k in iteration (1.8):

B. There exists a sequence of nonnegative numbers $\{\gamma_k\}$ such that

$$L_k^2 \le 1+\gamma_k, \quad \sum_{k=0}^{\infty}\gamma_k < \infty. \tag{2.18}$$

where $L_k = \max(\|D_k\|, \|D_k^{-1}\|)$.

Under assumption B, we have that Lemma 2.1 holds and this will be extensively exploited in the proof of our results.

This assumption is closely related to those in [24, 25] and implies that $D_k \to I$ when k diverges. Therefore, we can expect that the asymptotic behavior of the scaled method is similar to that of the nonscaled one. However, introducing D_k in iteration (1.8) allows us to design more flexible algorithms, having a further parameter to better exploit the problem features. Indeed, several papers in the recent literature propose suitable scaling techniques especially tailored for some relevant problems; see, for example, the split gradient technique [48, 68], the Barzilai–Borwein affine scaling

[41] (see also [21] and [42]), and the majorization-minimization approach [22]. All the mentioned references show that a careful choice of the metric can lead to significantly better practical performances of the algorithms. In general, the choice of the metric is problem dependent, i.e., is strongly related to the specific objective function and constraints structure. For this reason, here and in the following we investigate the properties of method (1.8) only under assumption B, without focusing on a specific scaling matrix choice. Indeed, our aim is to develop a theoretical framework for (1.8) such that different scaling techniques could be included as a special case of it.

We conclude this section with the complexity analysis of iteration (1.8).

2.1. Complexity analysis. We want to estimate the expected error after a finite number of iterations (1.8) with constant step size $\alpha_k = \alpha$ and $\epsilon_k = \epsilon$, $\alpha, \epsilon > 0$.

Setting $\alpha/\epsilon = \nu$ and borrowing the ideas in [28, Corollary 2.4], from (2.13) we obtain

$$(f+\Phi)^k_{\text{best}} - (f(\tilde{x}) + \Phi(\tilde{x})) \leq \frac{\theta_0^k \|x^{(0)} - \tilde{x}\|^2 + 2\nu \sum_{j=0}^k \tilde{\theta}_j^k \alpha^2 + \sigma \sum_{j=0}^k \theta_j^k \alpha^2}{2\sum_{j=0}^k \tilde{\theta}_j^k \alpha}$$
$$\leq \frac{M\|x^{(0)} - \tilde{x}\|^2 + (2\nu + \sigma L)\sum_{j=0}^k \tilde{\theta}_j^k \alpha^2}{2\sum_{j=0}^k \tilde{\theta}_j^k \alpha}.$$

Notice that for $\epsilon = 0$ we would obtain an analogous inequality with $\nu = 0$.

Choosing $\alpha = \frac{\|x^{(0)} - \tilde{x}\|\sqrt{M}}{(2\nu+\sigma L)^{\frac{1}{2}}(\sum_{j=0}^k \tilde{\theta}_j^k)^{\frac{1}{2}}}$ to minimize the right-hand side we obtain

$$(f+\Phi)^k_{\text{best}} - (f(\tilde{x})+\Phi(\tilde{x})) \leq \frac{\sqrt{M}(2\nu+\sigma L)^{\frac{1}{2}}\|x^{(0)} - \tilde{x}\|}{(\sum_{j=0}^k \tilde{\theta}_j^k)^{\frac{1}{2}}} \leq \frac{\sqrt{M}(2\nu+\sigma L)^{\frac{1}{2}}\|x^{(0)} - \tilde{x}\|}{(k+1)^{\frac{1}{2}}},$$

where the last inequality follows from the fact that $\tilde{\theta}_j^k \geq 1$. Thus, the scaled iteration (1.8) has the $\mathcal{O}(1/\sqrt{k+1})$ complexity.

A lower complexity bound for iteration (1.8) can be obtained using the same arguments as in [54, Theorem 3.2.1], when the scaling matrices D_k are diagonal; thus, in this case, the $\mathcal{O}(1/\sqrt{k+1})$ complexity obtained above is optimal. For the sake of completeness we give some details on how to derive the lower complexity bound. We consider the worst case example $\Phi = 0$, $f(x) = \gamma \max_{1\leq i\leq k+1} x_i + \frac{\mu}{2}\|x\|^2$, $\gamma, \mu > 0$, whose subdifferential is given by $\partial f(x) = \mu x + \gamma\{\sum_{i\in I(x)} a_i e_i, a_i \geq 0\}$, where e_j is the jth vector of the standard basis and $I(x) = \{1 \leq j \leq k+1 : x_i = \max_{1\leq i\leq k+1} x_i\}$.

The minimum of f is $f^* = -\gamma^2/(2\mu(k+1))$ which is attained at the point $\tilde{x}$ whose components are $\tilde{x}_i = -\gamma/(\mu(k+1))$, $i = 1, \ldots, k+1$, $\tilde{x}_i = 0$, $i = k+2, \ldots, n$.

Starting from $x^{(0)} = 0$ and selecting the (ϵ-)subgradient $u^{(j)} = \mu x^{(j)} + \gamma e_{\min(I(x^{(j)}))} \in \partial f(x^{(j)})$ in (1.8), it is easy to see that, for any diagonal scaling matrix and any choice of the step size, $x_i^{(j)} = 0$ for $i = j+1, \ldots, n$, $j = 0, \ldots, k$. This implies that $f(x^{(j)}) \geq 0$ for all $j = 0, \ldots, k$, i.e., there are no function improvements in the first k steps. Following the proof of [54, Theorem 3.2.1], setting $\gamma = C\sqrt{k+1}/(1+\sqrt{k+1})$, and $\mu = C/(1+\sqrt{k+1}) \cdot 1/\|x^{(0)} - \tilde{x}\|$, where C is the Lipschitz constant of f in the ball centered at $x^{(0)}$ with radius $\|x^{(0)} - \tilde{x}\|$, it can be shown that $f(x^{(k)}) - f^* \geq C\|x^{(0)} - \tilde{x}\|/(2+2\sqrt{k+1})$.

The above arguments do not apply when the matrices D_k are chosen with nonzero off diagonal entries. In this case, the $\mathcal{O}(1/\sqrt{k+1})$ complexity estimate could be non-optimal.

On the other side, we observe that choosing a nondiagonal scaling matrix in iteration (1.8) could be impractical for large scale problems, since the proximity operator is associated with its inverse.

3. Convergence analysis with square summable step size sequences. In this section we show that the method (1.8) with the rule $(\mathcal{R}_4)$ generates a sequence of points converging to a solution of (1.6), if any, under standard assumptions on the error sequence ϵ_k and the scaling matrices D_k.

THEOREM 3.1. *Let $\{x^{(k)}\}$ be the sequence generated by iteration* (1.8)*, where $u^{(k)} \in \partial_{\epsilon_k} f(x^{(k)})$, for a given sequence $\{\epsilon_k\} \subset \mathbb{R}$, $\epsilon_k \geq 0$. Assume that* A *and* B *hold and that*

$$\lim_{k\to\infty} \epsilon_k = 0, \tag{3.1}$$

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \tag{3.2}$$

$$\sum_{k=0}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=0}^{\infty} \epsilon_k \alpha_k < \infty. \tag{3.3}$$

Setting $f^* = \inf_{x\in\mathbb{R}^n}(f(x)+\Phi(x))$ (possibly $f^* = -\infty$) and defining the set X^* of the solutions of (1.6)*, we have*

(a) $\liminf_{k\to\infty}(f(x^{(k)}) + \Phi(x^{(k)})) = f^*$;
(b) *If $\{x^{(k)}\}$ is bounded, there exists a limit point of it belonging to X^*;*
(c) *If X^* is nonempty, the sequence $\{x^{(k)}\}$ converges to a solution of* (1.6) *and* $\lim_{k\to\infty}(f(x^{(k)}) + \Phi(x^{(k)})) = f^*$;
(d) *If X^* is empty, the sequence $\{x^{(k)}\}$ is unbounded.*

Proof. (a) Setting $\bar{f} = \liminf_{k\to\infty}(f(x^{(k)}) + \Phi(x^{(k)}))$, we now have to prove that $\bar{f} = f^*$.

Assume, to arrive at a contradiction, that there exists an $\epsilon > 0$ such that

$$\bar{f} - 2\epsilon > f^*.$$

Then, there exists $\hat{x} \in \mathrm{dom}(\Phi)$ such that

$$\bar{f} - 2\epsilon > f(\hat{x}) + \Phi(\hat{x}).$$

Since $\bar{f} = \liminf_{k\to\infty} f(x^{(k)}) + \Phi(x^{(k)})$, there exists k_0 such that, for all $k \geq k_0$, we have

$$f(x^{(k)}) + \Phi(x^{(k)}) \geq \bar{f} - \epsilon.$$

Summing up the above two relations, for all $k \geq k_0$ we obtain

$$(f(x^{(k)}) + \Phi(x^{(k)})) - (f(\hat{x}) + \Phi(\hat{x})) > \epsilon. \tag{3.4}$$

Consider now inequality (2.8) with $x = \hat{x}$:

$$\begin{aligned}
\|x^{(k+1)} - \hat{x}\|^2 &\leq L_k^2 \|x^{(k)} - \hat{x}\|^2 + 2L_k\alpha_k(f(\hat{x}) + \Phi(\hat{x}) - (f(x^{(k)}) + \Phi(x^{(k)})) + \epsilon_k) \\
&\quad + L_k\alpha_k^2 \|u^{(k)} + w^{(k)}\|_{D_k}^2 \\
&\leq L_k^2 \|x^{(k)} - \hat{x}\|^2 + 2L_k\alpha_k(\epsilon_k - \epsilon) + \sigma L_k^2 \alpha_k^2 \\
&\leq L_k^2 \|x^{(k)} - \hat{x}\|^2 + 2L_k^2\alpha_k\epsilon_k + \sigma L_k^2\alpha_k^2 - 2\alpha_k\epsilon,
\end{aligned}$$

where the second inequality follows from (2.16) and (3.4) and the third one from

$L_k \geq 1$. By repeatedly applying the same arguments for $j = k_0, \ldots, k$ we obtain

$$\|x^{(k+1)} - \hat{x}\|^2 \leq \theta_{k_0}^k \|x^{(k)} - \hat{x}\|^2 + 2\sum_{j=k_0}^{k} \theta_j^k \alpha_j \epsilon_j + \sigma \sum_{j=k_0}^{k} \theta_j^k \alpha_j^2 - 2\epsilon \sum_{j=k_0}^{k} \alpha_j \tag{3.5}$$

$$\leq M \left(\|x^{(k)} - \hat{x}\|^2 + 2\sum_{j=k_0}^{k} \alpha_j \epsilon_j + \sigma \sum_{j=k_0}^{k} \alpha_j^2 \right) - 2\epsilon \sum_{j=k_0}^{k} \alpha_j, \tag{3.6}$$

where θ_j^k and M are defined as in Lemma 2.1. Thanks to the assumptions (3.2)–(3.3), for k sufficiently large we have a contradiction.

(b) Now we assume that $\{x^{(k)}\}$ is bounded and we show that there exists a limit point of it belonging to X^*. To this end, we consider a subsequence $\{x^{(k_i)}\}$ of $\{x^{(k)}\}$ such that $\lim_{i\to\infty} f(x^{(k_i)}) + \Phi(x^{(k_i)}) = f^*$ (the existence of this subsequence is guaranteed by part (a) of the theorem). Since $\{x_k\}$ is bounded, $\{x^{k_i}\}$ is also bounded and, without loss of generality, we can assume that $\{x^{(k_i)}\}$ converges to a point x^∞ for $i \to \infty$. Then, recalling that $f + \Phi$ is lower semicontinuous and that f^* is its minimum, we have

$$f^* \leq f(x^\infty) + \Phi(x^\infty) \leq \lim_{i\to\infty} f(x^{(k_i)}) + \Phi(x^{(k_i)}) = f^*.$$

Thus, $f(x^\infty) + \Phi(x^\infty) = f^*$ and $x^\infty \in X^*$.

(c) Now we assume $X^* \neq \emptyset$ and we consider $\tilde{x} \in X^*$. We first prove that the sequence $\{x^{(k)}\}$ is bounded. Invoking again inequality (2.8) in Lemma 2.2 with $x = \tilde{x}$, observing that $f(\tilde{x}) + \Phi(\tilde{x}) \leq f(x^{(k)}) + \Phi(x^{(k)})$ and recalling (2.16) and $L_k \geq 1$, reasoning exactly as in the proof of (3.6), we obtain

$$\|x^{(k+1)} - \tilde{x}\|^2 \leq M \left(\|x^{(0)} - \tilde{x}\|^2 + \sigma \sum_{j=0}^{k} \alpha_j^2 + 2\sum_{j=0}^{k} \alpha_j \epsilon_j \right). \tag{3.7}$$

Thus, by conditions (3.3), the sequence $\{x^{(k)}\}$ is bounded.

By part (b) there exists a limit point x^∞ of $\{x^{(k)}\}$ belonging to X^*. We now show that the whole sequence converges to x^∞. Let δ be any positive number; since x^∞ is an accumulation point of $\{x^{(k)}\}$ and from (3.3), there exists a positive integer m_δ such that $\|x^\infty - x^{(m_\delta)}\|^2 \leq \delta/(3M)$, $\sum_{j=m_\delta}^{\infty} \alpha_k^2 \leq \delta/(3\sigma M)$, and $\sum_{j=m_\delta}^{\infty} \alpha_k \epsilon_k \leq \delta/(6M)$. Then, for any $k > m_\delta$, using the same arguments as in (3.6)–(3.7), we obtain

$$\begin{aligned} \|x^{(k)} - x^\infty\|^2 &\leq M \left(\|x^{(m_\delta)} - x^\infty\|^2 + \sigma \sum_{j=m_\delta}^{k-1} \alpha_j^2 + 2\sum_{j=m_\delta}^{k-1} \alpha_j \epsilon_j \right) \\ &\leq M \left(\|x^{(m_\delta)} - x^\infty\|^2 + \sigma \sum_{j=m_\delta}^{\infty} \alpha_j^2 + 2\sum_{j=m_\delta}^{\infty} \alpha_j \epsilon_j \right) \\ &\leq \delta. \end{aligned}$$

Since δ can be chosen arbitrarily small, then $\{x^{(k)}\}$ converges to x^∞.

It remains to show that $\lim_{k\to\infty} f(x^{(k)}) + \Phi(x^{(k)}) = f^*$. We observe that, for any $v^{(k)} \in \partial(f + \Phi)(x^{(k)})$, we have

$$f(x^{(k)}) + \Phi(x^{(k)}) - f^* \leq (x^{(k)} - x^\infty)^T v^{(k)} \leq \|x^{(k)} - x^\infty\| \cdot \|v^{(k)}\|.$$

Since $\{x^{(k)}\}$ is bounded and $\mathbb{R}^n$ is a finite dimensional space, $\bigcup_{k\geq 0} \partial(f + \Phi)(x^{(k)})$ is

bounded [1, p. 25] (see also Remark 2 to follow). Thus, the right-hand side of the previous inequality goes to zero as k diverges and this yields the result.

(d) The last claim is a direct consequence of part (b). ∎

We now discuss some further issues about Theorem 3.1, relating our results to the recent literature, in particular with the papers [24, 25] and [52].

Remark 1. From the inequality (2.7) in Lemma 2.2 with $x = \tilde{x}$, and from $\|w^{(k)} + u^{(k)}\|^2_{D_k^{-1}} \le L_k\sigma \le L\sigma$, observing that

$$\begin{aligned}\|x^{(k+1)} - \tilde{x}\|^2_{D_k^{-1}} &\ge \lambda_{\min}(D_k^{-1})\|x^{(k+1)} - \tilde{x}\|^2 \\ &= \frac{\lambda_{\min}(D_k^{-1})}{\lambda_{\max}(D_{k+1}^{-1})}\lambda_{\max}(D_{k+1}^{-1})\|x^{(k+1)} - \tilde{x}\|^2 \\ &\ge \lambda_{\min}(D_k^{-1})\lambda_{\min}(D_{k+1})\|x^{(k+1)} - \tilde{x}\|^2_{D_{k+1}^{-1}} \\ &\ge \frac{1}{L_kL_{k+1}}\|x^{(k+1)} - \tilde{x}\|^2_{D_{k+1}^{-1}},\end{aligned}$$

we obtain

$$\|x^{(k+1)} - \tilde{x}\|^2_{D_{k+1}^{-1}} \le \zeta_k\|x^{(k)} - \tilde{x}\|^2_{D_k^{-1}} + \xi\zeta_k\alpha_k^2 + 2\zeta_k\alpha_k\epsilon_k,$$

where $\zeta_k = \sqrt{(1+\gamma_k)(1+\gamma_{k+1})}$ and $\xi = L\sigma$. By the assumptions made on $\{\gamma_k\}$, the sequence $\{\zeta_k\}$ is bounded. We can also set $\zeta_k = 1+\eta_k$ with $\eta_k = \sqrt{(1+\gamma_k)(1+\gamma_{k+1})}-1$, and observe that the series $\sum\eta_k$ and $\sum\gamma_k$ have the same behavior, thanks to the limit $\lim_{z\to 0}(\sqrt{1+z}-1)/z = 1/2$. Then, from the assumption (2.18), we can conclude that $\sum\eta_k$ is a convergent series.

Thus, the sequence $\{x^{(k)}\}$ is *quasi-Fejér monotone* with respect to X^* relative to $\{D_k^{-1}\}$, in the sense of [24, Definition 3.1] and we could apply [24, Proposition 3.2] (see also [25]) to obtain that $\{\|x^{(k)} - \tilde{x}\|_{D_k^{-1}}\}$ converges and, thus, $\{x^{(k)}\}$ is bounded.

Variable metric was introduced also in [52, Chapter 5] in the context of subgradient methods for unconstrained problems (i.e., $X = \mathbb{R}^n$). In this case, setting $D_k = B_kB_k^T$, the scaling matrices are assumed to satisfy $\|B_{k+1}^{-1}B_k\| \ge 1$ and $\prod_{k=0}^{\infty}\|B_{k+1}^{-1}B_k\|^2 < \infty$. Even if the second condition is verified under the assumptions of Theorem 3.1, we observe that the requirement $\|B_{k+1}^{-1}B_k\| \ge 1$ restricts the choice of the scaling matrix, strictly connecting the metrics adopted in two successive iterates.

Finally, we note that similar results are obtained also in [1, Lemma 1] for ϵ-subgradient projection methods, i.e., $D_k = I$, $\Phi = \iota_X$, and in [28, Theorem 2.6] for the case $\epsilon_k = 0$ and $D_k = I$.

3.1. Convergence rate estimate. When the solution set is nonempty, we are able to provide a convergence rate estimate for method (1.8) with the step size rule (3.2)–(3.3). From (2.17), recalling that $\tilde{\theta}_j^k \ge 1$ and $\tilde{\theta}_j^k \le \theta_j^k \le M$, we also obtain

$$\sum_{j=0}^{k}\alpha_j(f(x^{(j)})+\Phi(x^{(j)})-(f(\tilde{x})+\Phi(\tilde{x}))) \le \frac{M}{2}\left(\|x^{(0)} - \tilde{x}\|^2 + 2\sum_{j=0}^{k}\alpha_j\epsilon_j + \sigma\sum_{j=0}^{k}\alpha_j^2\right).$$

By the assumptions (3.2)–(3.3) this implies

$$\sum_{j=0}^{\infty}\alpha_j(f(x^{(j)}) + \Phi(x^{(j)}) - (f(\tilde{x}) + \Phi(\tilde{x}))) < \infty.$$

Then, proceeding as in the proof of [1, Theorem 2] we can also conclude that there exists a subsequence $\{x^{(\ell_k)}\}$ of $\{x^{(k)}\}$ such that

$$f(x^{(\ell_k)}) + \Phi(x^{(\ell_k)}) - (f(\tilde{x}) + \Phi(\tilde{x})) \leq \left(\sum_{k=0}^{\ell_k} \alpha_k \right)^{-1}.$$

The previous inequality gives a quite pessimistic convergence rate estimate: indeed, when the step length is chosen as $\alpha_k = \mathcal{O}(\frac{1}{k})$, we can only conclude that there exists a subsequence $\{f(x^{(\ell_k)}) + \Phi(x^{(\ell_k)})\}$ of $\{f(x^{(k)}) + \Phi(x^{(k)})\}$ such that

$$f(x^{(\ell_k)}) + \Phi(x^{(\ell_k)}) - (f(\tilde{x}) + \Phi(\tilde{x})) \leq \frac{1}{\log(\ell_k)}.$$

However, in spite of this poor theoretical estimate, the practical performances of the method (1.8) can be significantly improved by introducing a variable metric, as shown in the toy example presented in the following subsection.

3.2. Scaling matrix impact: An illustrative example. Consider, for simplicity, problem (1.1), i.e., choose $\Phi = \iota_X$ in (1.6). Iteration (1.8) with $D_k = I$ then coincides with (1.2) and corresponds to the solution of the following quadratic subproblem, related to matrix $\frac{1}{\alpha_k} I$:

$$\min_{x \in X} f(x^{(k)}) + (x - x^{(k)})^T u^{(k)} + \frac{1}{2\alpha_k} \|x - x^{(k)}\|^2 \tag{3.8}$$

with $u^{(k)} \in \partial_{\epsilon_k} f(x^{(k)})$. In practice, a simple quadratic model approximates the function f at $x^{(k)}$. This model is motivated by the fact that $-(x^{(k)} - \tilde{x})^T u^{(k)} \leq \epsilon_k$ with ϵ_k decreasing to 0 as k increases. Then, $-u^{(k)}$ is an approximate descent direction for the function $\|x - \tilde{x}\|^2$ at $x^{(k)}$ [54]. Introducing a variable metric as in (1.8), any iteration is a solution of

$$\min_{x \in X} f(x^{(k)}) + (x - x^{(k)})^T u^{(k)} + \frac{1}{2\alpha_k} \|x - x^{(k)}\|^2_{D_k^{-1}}. \tag{3.9}$$

As pointed out in section 6, there are a number of applications where the structure of the feasible set and of the objective function suggests simple rules to devise a scaling matrix D_k [8, 11, 14, 49, 57, 65, 68].

In order to give a visual example of the different behaviors of the ϵ-subgradient method in the scaled and nonscaled case, we consider this simple problem

$$\min_{x_1 \geq 0, x_2 \geq -1} f(x) \equiv 0.5 \log\left(\frac{0.5}{x_1}\right) + x_1 + 2x_2 - 0.5 + |x_1| + |x_2|. \tag{3.10}$$

An ϵ-subgradient of f at $x^{(k)}$ can be obtained as

$$u^{(k)} = (1 - \frac{0.5}{x_1^{(k)}} + y_1^{(k+1)}, 2 + y_2^{(k+1)})^T,$$

where $y^{(k+1)}$ is an ϵ–subgradient of $|x_1| + |x_2|$ at $x^{(k)}$ computed as explained in section 5 with the rule (5.2).

Following the ideas proposed in [9, 11, 48], the presence of bound constraints suggests setting the scaling matrix D_k as the diagonal matrix whose entries are given by $(D_k)_{ii} = \max(\min(d_i^{(k)}, \sqrt{1+\gamma_k}), 1/\sqrt{1+\gamma_k})$, where

$$d^{(k)} = \begin{pmatrix} x_1^{(k)}/(1 + \max(0, y_1^{(k+1)})) \\ (x_2^{(k)} + 1)/(2 + \max(0, y_2^{(k+1)})) \end{pmatrix}$$

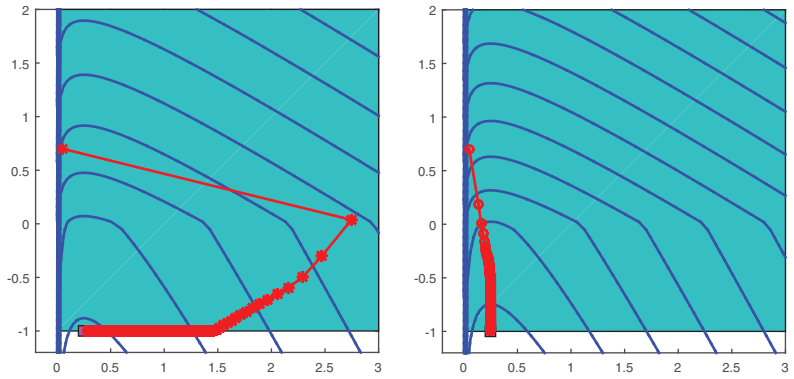

FIG. 1. *Path generated by methods* (1.2) *(left) and* (1.8) *(right) on the example* (3.10). *The starting point is* $(0.05, 0.7)^T$ *for both methods; blue lines are contours of the objective function.*

and γ_k are such that condition (2.18) is met. This kind of choice is motivated by the structure of the constraints since, as a simple computation shows, when $\alpha_k \leq 1$ and $x^{(k)}$ is feasible, then the point whose components are $x_i^{(k)} - \alpha_k d_i^{(k)} u_i^{(k)}$, $i = 1, 2$, is still feasible.

Figure 1 shows the paths followed by the methods (1.2) (left) and (1.8) (right), starting from $x^{(0)} = (0.05, 0.7)^T$ and with $\alpha_k = 0.3/(k+1)$.

We observe that the trajectory corresponding to the scaled method, as a consequence of the scaling matrix choice, approaches the solution, which is located at $(0.25, -1)$, from the interior of the feasible region, while the points generated by (1.2) lay on its boundary.

Moreover, in this example, 221 iterations of the scaled method are enough to obtain an approximation of the solution with an error equal to 0.009, while 500 iterations (1.2) provide an error equal to 0.041.

This improvement of the convergence behavior observed in the scaled case can be due to its better capability to capture the local features of the problem. Figure 2 is a close up of the first iteration of methods (1.2) (left) and (1.8) (right). The contour lines of the quadratic models (3.8) and (3.9) are drawn in black. It can be observed that (3.9) better approximates locally the contour lines of the objective function.

3.3. Convergence analysis with normalization of the direction. In this subsection we consider the following variant of iteration (1.8):

$$x^{(k+1)} = \operatorname{prox}_{\alpha_k \Phi, D_k^{-1}} \left(x^{(k)} - \frac{\alpha_k}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})} D_k u^{(k)} \right), \tag{3.11}$$

where $w^{(k)}$ is an arbitrary element of $\partial\Phi(x^{(k)})$.

The convergence of the method (3.11) can be analyzed as follows, without any assumption on the boundedness of the ϵ-subgradients of f and of the subgradients of Φ.

THEOREM 3.2. *Let* $\{x^{(k)}\}_{k\in\mathbb{N}}$ *be a sequence satisfying* (3.11). *Let assumption* B *and conditions* (3.1)–(3.3) *be satisfied. Then,* $\liminf_{k\to\infty}(f(x^{(k)}) + \Phi(x^{(k)})) = \inf_{x\in\mathbb{R}^n}(f(x) + \Phi(x)) = f^*$. *Moreover, if the set of the solutions of* (1.6), X^*, *is nonempty, the sequence* $\{x^{(k)}\}$ *converges to a solution of* (1.6) *and* $\lim_{k\to\infty}(f(x^{(k)}) + \Phi(x^{(k)})) = f^*$, *while if* X^* *is empty, the sequence* $\{x^{(k)}\}$ *is unbounded.*

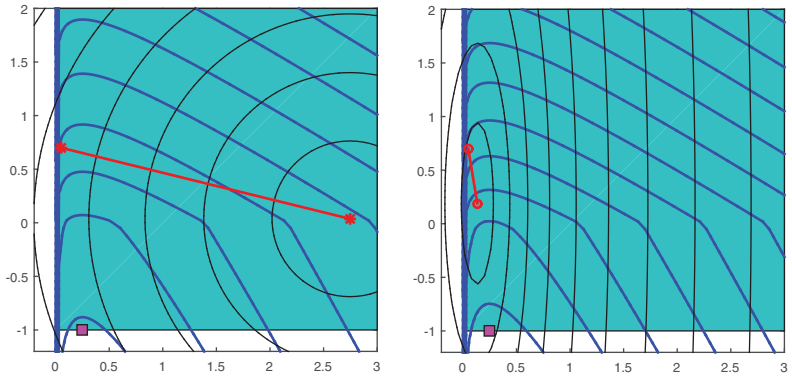

FIG. 2. *First iteration of methods* (1.2) *(left) and* (1.8) *(right) on the example* (3.10)*, starting from* $x^{(0)} = (0.05, 0.7)^T$ *and* $u^{(0)} = (-8.985, 2.21)^T \in \partial_{0.602} f(x^{(0)})$*. The blue lines are contours of the objective function; the black lines are contours of the two quadratic models* (3.8) *(left) and* (3.9) *(right). The magenta square is the minimum point.*

Proof. Let us define $\bar{\alpha}_k = \frac{\alpha_k}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})}$. We have

$$\bar{\alpha}_k \leq \alpha_k \quad \text{and} \quad \bar{\alpha}_k^2 \|u^{(k)} + w_k\|^2 \leq \alpha_k^2. \tag{3.12}$$

Assume that there exists a point $\hat{x}$ such that (3.4) holds for all $k \geq k_0$ for some $\epsilon \geq 0$. By applying (2.8) to iteration (3.11) with $x = \hat{x}$ we obtain

$$\begin{aligned} \|x^{(k+1)} - \hat{x}\|^2 &\leq L_k^2 \|x^{(k)} - \hat{x}\|^2 + 2L_k \bar{\alpha}_k (f(\hat{x}) + \Phi(\hat{x}) - (f(x^{(k)}) + \Phi(x^{(k)})) + \epsilon_k) \\ &\quad + L_k \bar{\alpha}_k^2 \|u^{(k)} + w^{(k)}\|_{D_k}^2 \\ &\leq L_k^2 \|x^{(k)} - \hat{x}\|^2 + 2L_k \bar{\alpha}_k (\epsilon_k - \epsilon) + L_k^2 \alpha_k^2 \\ &\leq L_k^2 \|x^{(k)} - \hat{x}\|^2 + 2L_k^2 \alpha_k \epsilon_k + L_k^2 \alpha_k^2. \end{aligned} \tag{3.13}$$

By repeatedly applying the same arguments for $j = k_0, \ldots, k$ we obtain

$$\begin{aligned} \|x^{(k+1)} - \hat{x}\|^2 &\leq \theta_{k_0}^k \|x^{(k)} - \hat{x}\|^2 + 2 \sum_{j=k_0}^{k} \theta_j^k \alpha_j \epsilon_j + \sum_{j=k_0}^{k} \theta_j^k \alpha_j^2 \\ &\leq M \left(\|x^{(k)} - \hat{x}\|^2 + 2 \sum_{j=k_0}^{k} \alpha_j \epsilon_j + \sum_{j=k_0}^{k} \alpha_j^2 \right), \end{aligned} \tag{3.14}$$

where θ_j^k and M are defined as in Lemma 2.1. Thus, by assumption (3.3), we can conclude that $\{x^{(k)}\}$ is bounded. As a consequence of this and of assumption (3.1), since we are in a finite dimensional space, we have that there exist two positive constants ρ_u and ρ_w such that $\bigcup_{k \geq 0} \partial_{\epsilon_k} f(x^{(k)}) \subseteq \{u \in \mathbb{R}^n : \|u\| \leq \rho_u\}$ and $\bigcup_{k \geq 0} \partial \Phi(x^{(k)}) \subseteq \{u \in \mathbb{R}^n : \|u\| \leq \rho_w\}$ [1, p. 25]. This implies $\|u^{(k)} + w^{(k)}\|_{D_k}^2 \leq L\sigma$, where $\sigma = (\rho_u + \rho_w)^2$ and, therefore,

$$\sum_{k=0}^{\infty} \bar{\alpha}_k \geq \frac{1}{\max(1, L\sigma)} \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Now, from (3.13) it follows that

$$\|x^{(k+1)} - \hat{x}\|^2 \leq M \left(\|x^{(k)} - \hat{x}\|^2 + 2 \sum_{j=k_0}^{k} \alpha_j \epsilon_j + \sigma \sum_{j=k_0}^{k} \alpha_j^2 \right) - 2 \frac{\epsilon}{\max(1, L\sigma)} \sum_{j=k_0}^{k} \alpha_j.$$

The previous inequality is analogous to (3.6), while (3.14) is analogous to (3.7). Based on these remarks, the rest of the proof can be obtained by the same arguments employed for Theorem 3.1. ∎

We remark that the convergence rate estimate in section 3.1 also holds for iteration (3.11), with very minor modifications.

Remark 2. Similar results can be found also in [1, Theorem 1] and [50, Theorem 10] for the ϵ-subgradient projection method and for the forward-backward subgradient algorithm in [28, section 2.1]. When dealing with infinite dimensional Hilbert spaces, to get the same results stated in Theorems 3.1 and 3.2, it could be assumed that $\partial_\epsilon f$ and $\partial\Phi$ are bounded on bounded sets on the domain of Φ, i.e., $\cup_{x\in B}\partial_\epsilon f(x)$ and $\cup_{x\in B}\partial\Phi(x)$ are bounded for any $\epsilon > 0$ and any bounded closed subset B of $\mathrm{dom}(\Phi)$ [1, 28].

The results in this section can be exploited for the practical implementation of the methods (1.8) and (3.11), since they indicates how to choose the sequence $\{\alpha_k\}$, and they are employed also in the convergence analysis of (3.11) equipped with an adaptive step size rule, as the one described in the following.

4. Convergence analysis with dynamic step size rule. A critical point for the implementation of the methods (1.8) and (3.11) is how to select the sequence $\{\alpha_k\}$; a practical strategy to obtain good performances is still an open problem, since they are, in general, quite sensitive to this choice [16]. Borrowing the ideas of [19] and [40], in this section we describe a *level algorithm* that allows us to adaptively compute a dynamic step size α_k in the iteration (3.11).

The resulting algorithm is detailed in Algorithm 1, whose underlying assumption is that, for any given $\epsilon_k \geq 0$, we are able to provide an element $u^{(k)}$ of the set $\partial_{\epsilon_k} f(x^{(k)})$ and a subgradient $w^{(k)} \in \partial\Phi(x^{(k)})$ such that assumption B holds.

In Algorithm 1, we have $f_k^{rec} = \min_{i=0,\ldots,k}(f(x^{(i)})+\Phi(x^{(i)}))$, while l is the number of times that the value f^{lev} has been updated and $k(l)$ is the iteration where the lth updating occurred. Finally, σ_k is the cumulative path length between two successive updates of f^{lev}.

Steps 2–5 aim to provide in f_k^{lev} an estimate of the optimal function value at the iterate k, which is used as the target level for the successive iterates until the objective function value is sufficiently close to it or the iterates move through a long path without approaching it. In the first case, i.e., when the inequality at Step 3 is satisfied, f_k^{lev} is reduced at Step 5 by subtracting the positive quantity δ_l to the best value obtained so far, f^{rec}. In the other case, when the inequality at Step 4 is satisfied, the estimated difference from the optimal value δ_l is reduced and, as a consequence of Step 5, the target level f_k^{lev} is increased.

One of the main differences between the step size computed by Algorithm 1 and the square summable sequence considered in the previous section is that the former one does not necessarily converge to zero.

In the rest of this section we prove that $\liminf_{k\to\infty}\{f(x^{(k)})+\Phi(x^{(k)})\}$, where $x^{(k)}$ is computed by Algorithm 1, is equal to the infimum of $f+\Phi$, using similar techniques as in [53].

Algorithm 1 Scaled forward–backward ϵ-Subgradient Level Algorithm (SSL)

Choose $B > 0$, $\nu_1, \nu_2 \in (0,1)$, $f_{-1}^{rec} = \infty$; $k = 0$, $l = 0$, $k(l) = 0$, $\delta_0 > 0$; choose $x^{(0)} \in X$.

For $k = 0, 1, 2, \ldots$

Step 1. Computation of $f(x^{(k)}) + \Phi(x^{(k)})$.

Step 2. If $f(x^{(k)}) + \Phi(x^{(k)}) < f_{k-1}^{rec}$, then $f_k^{rec} = f(x^{(k)}) + \Phi(x^{(k)})$ else $f_k^{rec} = f_{k-1}^{rec}$.

Step 3. If $f(x^{(k)}) + \Phi(x^{(k)}) < f_{k(l)}^{rec} - \nu_1 \delta_l$, then $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \delta_l$, $l = l + 1$ and go to Step 5.

Step 4. If $\sigma_k > B$, then $k(l+1) = k$, $\sigma_k = 0$, $\delta_{l+1} = \nu_2 \delta_l$, $l = l + 1$.

Step 5. Set $f_k^{lev} = f_{k(l)}^{rec} - \delta_l$.

Step 6. Update the step size and compute the new iterate

$$\alpha_k = \frac{f(x^{(k)}) + \Phi(x^{(k)}) - f_k^{lev}}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})},$$

$$x^{(k+1)} = \operatorname{prox}_{\alpha_k \Phi, D_k^{-1}}\left(x^{(k)} - \alpha_k D_k \frac{u^{(k)}}{\max(1, \|w^{(k)} + u^{(k)}\|_{D_k})}\right), \tag{4.1}$$

where $w^{(k)} \in \partial\Phi(x^{(k)})$.

Step 7. $\sigma_{k+1} = \sigma_k + \alpha_k$ and go to Step 1.

End

Before giving the main result, whose proof also exploits the results of the previous section, we recall the following technical lemma. We omit the proof, since it runs as that of [53, Lemma 2.2].

Lemma 4.1. *Let assumption* A *be satisfied and assume that there exists a sequence of positive numbers* $\{L_k\}$ *such that* $\|D_k\| \le L_k$, $\|D_k^{-1}\| \le L_k$, *with* $1 \le L_k \le L$ *for some positive constant* L *for all* $k \ge 0$. *Given* $B > 0$ *and* $\{\epsilon_k\}$ *such that* $\epsilon_k \to 0$ *as* $k \to \infty$ *in Algorithm* 1, *we have* $l \to \infty$ *and* $\liminf_{k\ge 0}(f(x^{(k)}) + \Phi(x^{(k)})) = -\infty$ *or* $\delta_l \to 0$ *as* $l \to \infty$.

The following theorem can be considered as a generalization of [53, Proposition 2.7], which only deals with the case $D_k = I$, $\epsilon_k = 0$ for all k.

Theorem 4.1. *Let assumptions* A *and* B *be satisfied. Then, for SSL we have* $\bar{f} = \liminf_{k\ge 0}(f(x^{(k)}) + \Phi(x^{(k)})) = \inf_{k\ge 0}(f(x) + \Phi(x)) = f^*$. *If*

$$X^* \ne \emptyset, \liminf_{k\ge 0}(f(x^{(k)}) + \Phi(x^{(k)})) = f(x^*) + \Phi(x^*), x^* \in X^*.$$

Proof. The first part of the proof aims to show that $\sum_j \alpha_j = \infty$ and runs as [53, Proposition 2.7]. For the sake of completeness, we report below the detailed derivation of the result.

From the previous lemma, if $\lim_{l\to\infty} \delta_l > 0$, we have $\liminf_{k\ge 0}(f(x^{(k)}) + \Phi(x^{(k)})) = -\infty$.

Now, we assume that $\delta_l \to 0$ as $l \to \infty$. Let S be given by

$$S = \{l \in \{1, 2, \ldots\}, \delta_l = \nu_2 \delta_{l-1}\}.$$

Then, from Steps 4 and 6 of Algorithm 1, we obtain

$$\sigma_k = \sigma_{k-1} + \alpha_{k-1} = \sum_{j=k(l)}^{k-1} \alpha_j,$$

so that $k(l+1) = k$ and $l+1 \in S$ whenever $\sum_{j=k(l)}^{k-1} \alpha_j > B$ at Step 4. Hence

$$\sum_{j=k(l-1)}^{k(l)-1} \alpha_j > B \qquad \forall l \in S$$

and since the cardinality of S is infinite, we have

$$\sum_{k=k(l)}^{\infty} \alpha_k \geq \sum_{l \geq \overline{l}, l \in S} \sum_{j=k(l-1)}^{k(l)-1} \alpha_j > \sum_{l \geq \overline{l}, l \in S} B = \infty. \tag{4.2}$$

Now in order to obtain a contradiction, assume that $\bar{f} > f^*$, so that for some $\tilde{y} \in \mathbb{R}^n$ and some $\eta > 0$

$$\bar{f} - \eta \geq f(\tilde{y}) + \Phi(\tilde{y}). \tag{4.3}$$

Since $\delta_l \to 0$ and $\epsilon_k \to 0$, there are large enough $\overline{l}$ and $\overline{k}$ such that, for all $l \geq \overline{l}$ and $k \geq \overline{k}$, we have $\delta_l < \eta/2$ and $\epsilon_k < \eta/2$; then for all $k \geq \tilde{k} = \max(k(\overline{l}), \overline{k})$,

$$f_k^{lev} - \epsilon_k = f_{k(l)}^{rec} - \epsilon_k - \delta_l > \bar{f} - \eta \geq f(\tilde{y}) + \Phi(\tilde{y}).$$

From this inequality, by (2.7) in Lemma 2.2 with $\alpha_k \equiv \frac{\alpha_k}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})}$, the definition of ϵ-subgradient, and the definition of α_k, we obtain

$$\begin{aligned}
&\|x^{(k+1)} - \tilde{y}\|^2_{D_k^{-1}} \\
&\quad \leq \|x^{(k)} - \tilde{y}\|^2_{D_k^{-1}} - 2\frac{\alpha_k}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})}(f(x^{(k)}) + \Phi(x^{(k)}) - f(\tilde{y}) - \Phi(\tilde{y}) - \epsilon_k) \\
&\qquad + \alpha_k^2 \\
&\quad \leq \|x^{(k)} - \tilde{y}\|^2_{D_k^{-1}} - 2\frac{\alpha_k}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})}(f(x^{(k)}) + \Phi(x^{(k)}) - f_k^{lev}) + \alpha_k^2 \\
&\quad \leq \|x^{(k)} - \tilde{y}\|^2_{D_k^{-1}} - \alpha_k^2.
\end{aligned}$$

In view of (2.5) and $L_k \geq 1$, we can write

$$\|x^{(k+1)} - \tilde{y}\|^2 \leq L_k^2 \|x^{(k)} - \tilde{y}\|^2 - \alpha_k^2. \tag{4.4}$$

By repeatedly applying the previous inequality we obtain

$$\|x^{(k+1)} - \tilde{y}\|^2 \leq \theta_{\tilde{k}}^k \|x^{(\tilde{k})} - \tilde{y}\|^2 - \sum_{j=\tilde{k}}^{k} \theta_{j+1}^k \alpha_j^2,$$

where $\theta_j^k = L_j^2 \cdot \cdots \cdot L_k^2$, $\theta_{k+1}^k = 1$; since $\theta_j^k \leq \theta_0^k \leq M$, where M is a positive constant (see Lemma 2.1) and $\theta_j^k \geq 1$ we have

$$\sum_{\tilde{k}}^{\infty} \alpha_j^2 \leq M \|x^{(\tilde{k})} - \tilde{y}\|^2$$

and consequently $\sum_{k=\tilde{k}}^{\infty} \alpha_k^2 < \infty$. Then $\alpha_k \to 0$ as $k \to \infty$ and, from (4.2), $\sum_{k=\tilde{k}}^{\infty} \alpha_k = \infty$.

Now we show that $\sum \alpha_k \epsilon_k < \infty$. Indeed, since $\epsilon_k \to 0$ as $k \to \infty$, there exists $\bar{k}$ such that $2\epsilon_k < \eta$ for $k \geq \bar{k}$, where η is such that (4.3) holds. We consider the inequality

(4.5)
$$\|x^{(k+1)} - \tilde{y}\|^2_{D_k^{-1}} \leq \|x^{(k)} - \tilde{y}\|^2_{D_k^{-1}} + \alpha_k^2 - 2\frac{\alpha_k}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})}(u^{(k)} + w^{(k)})^T(x^{(k)} - \tilde{y}).$$

For the convexity of $f + \Phi$, the inequality (4.3), and $2\epsilon_k < \eta$, we have

$$f(x^{(k)}) + \Phi(x^{(k)}) + (u^{(k)} + w^{(k)})^T(\tilde{y} - x^{(k)}) - \epsilon_k \leq f(\tilde{y}) + \Phi(\tilde{y}) \leq \bar{f} - \eta \leq f(x^{(k)}) + \Phi(x^{(k)}) - 2\epsilon_k.$$

Then we have

$$(u^{(k)} + w^{(k)})^T(\tilde{y} - x^{(k)}) \leq -\epsilon_k.$$

Using this inequality in (4.5), we obtain

$$\|x^{(k+1)} - \tilde{y}\|^2_{D_k^{-1}} \leq \|x^{(k)} - \tilde{y}\|^2_{D_k^{-1}} + \alpha_k^2 - 2\frac{\alpha_k \epsilon_k}{\max(1, \|u^{(k)} + w^{(k)}\|_{D_k})}.$$

Using the same arguments as above, we obtain

$$\|x^{(k+1)} - \tilde{y}\|^2 \leq L_k^2 \|x^{(k)} - \tilde{y}\|^2 + L_k^2 \alpha_k^2 - 2\frac{\alpha_k \epsilon_k}{\max(1, L\sqrt{\sigma})},$$

where $\sigma = (\rho_u + \rho_w)^2$. By repeatedly applying the previous inequality we have

$$\begin{aligned}\|x^{(k+1)} - \tilde{y}\|^2 &\leq \theta_{\bar{k}}^k \|x^{(\bar{k})} - \tilde{y}\|^2 + \theta_{\bar{k}}^k \sum_{j=\bar{k}}^{k} \alpha_j^2 - \frac{2}{\max(1, L\sqrt{\sigma})} \sum_{j=\bar{k}}^{k} \alpha_j \epsilon_j \\ &\leq M \left(\|x^{(\bar{k})} - \tilde{y}\|^2 + \sum_{j=\bar{k}}^{k} \alpha_j^2 \right) - \frac{2}{\max(1, L\sqrt{\sigma})} \sum_{j=\bar{k}}^{k} \alpha_j \epsilon_j.\end{aligned}$$

Then we have

$$\sum_{j=\bar{k}}^{\infty} \alpha_j \epsilon_j \leq \frac{M}{2} \max(1, L\sqrt{\sigma}) \left(\|x^{(\bar{k})} - \tilde{y}\|^2 + \sum_{j=\bar{k}}^{\infty} \alpha_j^2 \right) < \infty.$$

According to Theorem 3.2 we have $\bar{f} = f^*$ which contradicts (4.3). ∎

Clearly, since $f_k^{rec} = (f + \Phi)_{\text{best}}^k$ is monotone nonincreasing, Theorem 4.1 guarantees that its limit is f^*.

Further generalizations of Algorithm 1 could be included in the analysis of the previous theorem following [53, p. 122], where the authors suggest some modifications of Steps 2, 3, and 4 allowing a variable path bound B and different strategies to update the parameter δ_l. For the sake of simplicity we omit these details here.

5. SPDHG. The aim of this section is to present a concrete example of the method (1.8) for the problem

$$\min_{x\in\mathbb{R}^n} f_0(x) + f_1(Ax) + \Phi(x), \tag{5.1}$$

where $A \in \mathbb{R}^{m\times n}$, $f_0(x), f_1(x), \Phi(x)$ are convex, proper, lower semicontinuous functions such that $\operatorname{diam}(\operatorname{dom}(f_1^*))$ is finite, and $f_1^*(y)$ is the Fenchel dual of f_1. Clearly, (5.1) is a special case of (1.6). We propose the following SPDHG method for the solution of (5.1),

$$y^{(k+1)} = \operatorname{prox}_{\tau_k f_1^*, I}(y^{(k)} + \tau_k A x^{(k)}), \tag{5.2}$$

$$u^{(k)} = d^{(k)} + A^T y^{(k+1)}, \tag{5.3}$$

$$x^{(k+1)} = \operatorname{prox}_{\alpha_k \Phi, D_k^{-1}}(x^{(k)} - \alpha_k D_k u^{(k)}), \tag{5.4}$$

where $d^{(k)} \in \partial_{\mu_k} f_0(x^{(k)})$ for some $\mu_k \geq 0$, and $\{\tau_k\}$, $\{\alpha_k\}$ are the dual and primal step length sequences, respectively. Method (5.2)–(5.4) is a special case of the scaled forward-backward ϵ-subgradient method (1.8), where $f = f_0 + f_1 \circ A$. The key point of this interpretation is that $A^T y^{(k+1)}$ is an ϵ-subgradient of $f_1 \circ A$ at $x^{(k)}$ as stated in the following lemma.

LEMMA 5.1 (see [16, Lemma 1]). *Let $y^{(k+1)}$ defined as in (5.2). Then, $y^{(k+1)} \in \operatorname{dom}(f_1^*)$ and, thus, $A^T y^{(k+1)} \in \partial_{\psi_k}(f_1 \circ A)(x^{(k)})$, where $\psi_k = f_1(Ax^{(k)}) + f_1^*(y^{(k+1)}) - {y^{(k+1)}}^T A x^{(k)}$. Moreover, if there exists a positive number D such that $\operatorname{diam}(\operatorname{dom}(f_1^*)) \leq D$, then $\psi_k \leq (2\tau_k)^{-1} D^2$.*

Thus, recalling the additivity of the ϵ-subgradient, we can conclude that

$$u^{(k)} = d^{(k)} + A^T y^{(k+1)} \in \partial_{\epsilon_k} f(x^{(k)}), \quad \epsilon_k = \mu_k + \psi_k. \tag{5.5}$$

Motivated by the previous observation, building on the material developed in sections 3 and 4, we discuss two step size selection strategies for the method (5.2)–(5.4), providing two different SPDHG implementations. In the first case, we assume that $\{\tau_k\}$, $\{\alpha_k\}$, $\{\gamma_k\}$ are prefixed sequences.

The following corollary shows a proper setting of these parameters and states the convergence properties of the corresponding algorithm.

COROLLARY 5.1. *Let $\{x^{(k)}\}$ be the sequence generated by iteration (5.2)–(5.4). Assume that $d^{(k)} \in \partial_{\mu_k} f_0(x^{(k)})$ and that there exists $\rho > 0$ such that $\|d^{(k)}\| \leq \rho$ for all k. Assume also that there exists a sequence $\{w^{(k)}\}$, $w^{(k)} \in \partial\Phi(x^{(k)})$, such that $\|w^{(k)}\| \leq \rho_w$ for some $\rho_w > 0$. Define $L_k = \max(\|D_k\|, \|D_k^{-1}\|)$ and assume that $L_k \leq \sqrt{1+\gamma_k}$ for some nonnegative sequence of parameters $\{\gamma_k\}$. Let the step length sequences $\{\tau_k\}$, $\{\alpha_k\}$, and $\{\gamma_k\}$ satisfy*

$$\alpha_k = \mathcal{O}\left(\frac{1}{k^p}\right), \quad \tau_k = \mathcal{O}(k^p), \quad \gamma_k = \mathcal{O}\left(\frac{1}{k^q}\right), \quad \frac{1}{2} < p \leq 1, \quad q > 1. \tag{5.6}$$

Moreover, assume that μ_k converges to zero at least as $\frac{1}{\tau_k}$. If $\operatorname{diam}(\operatorname{dom}(f_1^*))$ is finite,

$\liminf_{k\to 0} f(x^{(k)}) + \Phi(x^{(k)}) = f^*$; *if, in addition, the set of the solutions of* (1.1) *is nonempty, the sequence* $\{x^{(k)}\}$ *converges to a solution of* (5.1) *and* $\lim_{k\to\infty} f(x^{(k)}) + \Phi(x^{(k)}) = f^*$.

Proof. Since $\mathrm{diam}(\mathrm{dom}(f_1^*))$ is finite, we can apply Lemma 5.1 obtaining $\psi_k \leq (2\tau_k)^{-1}D^2$ in (5.5). By the assumption (5.6) on τ_k and μ_k we obtain that $\epsilon_k = \mathcal{O}\left(\frac{1}{k^p}\right)$ and, as a consequence, $\alpha_k\epsilon_k = \mathcal{O}\left(\frac{1}{k^{2p}}\right)$. Since $\frac{1}{2} < p \leq 1$ and $q > 1$, all assumptions of Theorem 3.1 are satisfied and we obtain the result. ∎

On the other side, the SSL procedure for dynamically computing the primal step size α_k can also be implemented. For the sake of simplicity we assume $\mu_k = 0$. In this case, we have to provide a sequence of the scaling matrix bounds $\{\gamma_k\}$ and of the dual step size τ_k, while α_k is computed by Steps 2–5 of Algorithm 1.

The following corollary establishes the convergence properties of this implementation of method (5.2)–(5.4).

COROLLARY 5.2. *Let* $\{x^{(k)}\}$ *be the sequence generated by Algorithm* 1, *where* $u^{(k)}$ *in* (4.1) *is given by* (5.3), $d^{(k)} \in \partial f_0(x^{(k)})$. *Assume that that there exists* $\rho > 0$ *such that* $\|d^{(k)}\| \leq \rho$ *for all* k. *Assume also that there exists a sequence* $\{w^{(k)}\}$, $w^{(k)} \in \partial\Phi(x^{(k)})$, *such that* $\|w^{(k)}\| \leq \rho_w$ *for some* $\rho_w > 0$. *Define* $L_k = \max(\|D_k\|, \|D_k^{-1}\|)$ *and assume that* $\lim_{k\to\infty}\tau_k = \infty$, $L_k \leq \sqrt{1+\gamma_k}$, $\gamma_k = \mathcal{O}\left(\frac{1}{k^q}\right)$ *with* $q > 1$, *and that there exists* $\rho > 0$ *such that* $\|d^{(k)}\| \leq \rho$. *If* $\mathrm{diam}(\mathrm{dom}(f_1^*))$ *is finite, then we have* $\liminf_{k\to\infty} f(x^{(k)}) + \Phi(x^{(k)}) = f^*$.

Proof. Since $d^{(k)} \in \partial f_0(x^{(k)})$, by Lemma 5.1 we have $u^{(k)} \in \partial_{\epsilon_k} f(x^{(k)})$, where $\epsilon_k = f_1(Ax^{(k)}) + f_1^*(y^{(k+1)}) - {y^{(k+1)}}^T Ax^{(k)}$. Since $\mathrm{diam}(\mathrm{dom}(f_1^*))$ is finite, we can apply the second part of Lemma 5.1 obtaining $\epsilon_k \leq (2\tau_k)^{-1}D^2$ for a positive constant D such that $\mathrm{diam}(\mathrm{dom}(f_1^*)) \leq D$. Since $\lim_{k\to\infty}\tau_k = \infty$, we have $\lim_{k\to\infty}\epsilon_k = 0$ and by Theorem 4.1 we obtain the result. ∎

We conclude this section by observing that the complexity analysis in section 2 applies also to SPDHG. We also point out that in [43] the authors prove a $\mathcal{O}(1/k)$ complexity result for the ergodic sequence generated by iteration (5.2)–(5.4) with $D_k = I$ and $f_0 = 0$ under the assumption that Φ is strongly convex.

6. Application: Edge preserving deblurring of Poisson images. In this section we further specialize the SPDHG method, by focusing on a specific application in the image restoration context. Our aim is to suggest a strategy to compute a suitable scaling matrix D_k, fully defining the algorithm; as observed by several authors, this choice should be driven according to the specific problem features, such as the structure of the constraints and objective function [18, 48, 68].

For these reasons, we describe first some details of the image reconstruction problems which, in the Bayesian framework, can be formulated as constrained convex minimization problems of the form (5.1). For these problems, the function $f_0(x)$ measures the data discrepancy and should be chosen according to the noise statistics: in particular, when the data suffer from Poisson noise, the maximum likelihood principle leads to the generalized Kullback–Leibler divergence

$$f_0(x) = \sum_{i=1}^{n} g_i \log\frac{g_i}{(Hx)_i + b} + (Hx)_i + b - g_i, \tag{6.1}$$

where $g \in \mathbb{R}^n$ is the observed image, $H \in \mathbb{R}^{n\times n}$ represents the blurring operator, while $b \in \mathbb{R}$ is a nonnegative background term. Standard assumptions on H are that

it has nonnegative entries and $H^T e > 0$, where $e \in \mathbb{R}^n$ is the vector of all ones. Function (6.1) is convex (see [36] and [17] for the explicit expression of its Hessian and gradient) and when $b = 0$ its gradient is not Lipschitz continuous.

Since the entries of the unknown vector x represent the image pixels, a meaningful solution is obtained by defining the constraint set as the nonnegative orthant, i.e., $X = \{x \in \mathbb{R}^n : x_i \geq 0\}$.

On the other side, $f_1(Ax)$ plays the role of a regularization term enforcing suitable properties on the solution of (5.1). Typically, to preserve the edges in the solutions of (5.1), $f_1(Ax)$ can be chosen as

$$f_1(Ax) = \beta TV(x), \quad TV(x) = \sum_{i=1}^{n} \|A_i x\|, \quad A_i \in \mathbb{R}^{2\times n}, \tag{6.2}$$

where $TV(x)$ is the discrete, nonsmooth, TV functional, β is a positive regularization parameter, and $A_i \in \mathbb{R}^{2\times n}$ is defined such that $A_i x$ represents the discrete gradient of the image x at the pixel i. In these settings, the matrix A is defined by blocks as $A = \begin{pmatrix} A_1^T & A_2^T & \cdots & A_n^T \end{pmatrix}^T \in \mathbb{R}^{2n\times n}$.

In order to simplify the notation, we assume that $x \in \mathbb{R}^n$ is an $N \times N$ image, i.e., $n = N^2$ and we will indicate the component x_ℓ, $\ell = 1, \ldots, n$, also as $x_{i,j}$, $i, j = 1, \ldots, N$, with the correspondence $j = \lfloor (\ell - 1)/N \rfloor + 1$, $i = \ell - \lfloor (\ell - 1)/N \rfloor \cdot N$, where $\lfloor \cdot \rfloor$ denotes the integer quotient. With this notation, the ℓth discrete gradient of the image x can be written as

$$A_\ell x = \begin{pmatrix} x_{i+1,j} - x_{i,j} \\ x_{i,j+1} - x_{i,j} \end{pmatrix},$$

where some boundary conditions are assumed.

The minimization of the nonsmooth TV functional is especially relevant since it allows us to preserve the sharpness of the edges in the reconstructed image. It is well known that the same effects cannot be obtained by means of a Tikhonov regularization, i.e., by squaring the gradient norm (see, for example, [60]), then this image reconstruction problem has to be handled with nonsmooth optimization tools.

The problem to be solved has the structure (5.1), where $\Phi(x) = \iota_X(x)$. In this case, since f_0 is differentiable, we define $d^{(k)} = \nabla f_0(x^{(k)})$ in (5.4), so that $\mu_k = 0$ for all k in Corollary 5.1. Moreover, the evaluation of the proximity operators in (5.2) and (5.4) consists of a simple Euclidean projection onto the set $B \times B \times \cdots \times B \subset \mathbb{R}^{2n}$, where $B = \{z \in \mathbb{R}^2 : \|z\| \leq 1\}$, and in a projection onto the nonnegative orthant with respect to the norm induced by D_k^{-1}, respectively.

In order to devise a suitable scaling matrix D_k for SPDHG, we adapt to our case the split gradient strategy proposed in [9, 48] for nonnegatively constrained differentiable problems, which demonstrated to be very effective in several applications [8, 14, 57, 65, 68].

The key point of this approach consists in finding a subgradient decomposition of the form $u^{(k)} = V(x^{(k)}) - U(x^{(k)})$ with $V(x^{(k)}) > 0$ and $U(x^{(k)}) \geq 0$ for all k and then defining D_k in (5.4) as a diagonal scaling matrix whose entries are the projection of $x_i^{(k)}/V_i(x^{(k)})$ onto the set $[1/\sqrt{1+\gamma_k}, \sqrt{1+\gamma_k}]$.

This strategy has the advantage to agree with the nonnegativity constraints and strongly depends on the form of the subgradient $u^{(k)}$.

For a practical implementation of this strategy, we have to find a decomposition of the vector $u^{(k)} = \nabla f_0(x^{(k)}) + \beta A^T y^{(k+1)}$ as the difference of two nonnegative terms.

As concerns the first term, the gradient of f_0 has the natural decomposition $\nabla f_0(x) = H^T e - H^T v(x)$, where $v(x)$ denotes the vector with entries $v_i(x) = g_i/(Hx+b)_i$; by the assumptions on H, we have $H^T e > 0$ and $H^T v(x) \geq 0$ for all $x \geq 0$.

Thus, it remains to find a decomposition of the vector $A^T y^{(k+1)}$ in (5.4). To this end, we compute the explicit expression of it as a function of $x^{(j)}$, $j = 0, \ldots, k$. We first observe that, if the dual variable is partitioned as $y = \begin{pmatrix} y_1^T & y_2^T & \cdots & y_n^T \end{pmatrix}^T$, $y_i \in \mathbb{R}^2$, the updating rule (5.2) can be written as

$$\begin{aligned} \tilde{y}^{(k)} &= y^{(k)} + \tau_k \beta A x^{(k)}, \\ y^{(k+1)} &= S_k \tilde{y}^{(k)}, \end{aligned}$$

where S_k is a diagonal $2n \times 2n$ matrix with the following diagonal entries

$$(S_k)_{2i-1,2i-1} = (S_k)_{2i,2i} = \frac{1}{\max\{1, \|\tilde{y}_i^{(k)}\|\}}, \quad i = 1, \ldots, n. \tag{6.3}$$

If the method is initialized with $y^{(0)} = 0$, the dual variable can be written as

$$\begin{aligned} y^{(0)} &= 0, \\ y^{(1)} &= \beta \tau_0 S_0 A x^{(0)}, \\ y^{(2)} &= \beta S_1 (\tau_0 S_0 A x^{(0)} + \tau_1 A x^{(1)}), \\ y^{(3)} &= \beta S_2 (\tau_0 S_1 S_0 A x^{(0)} + \tau_1 S_1 A x^{(1)} + \tau_2 A x^{(2)}), \\ &\vdots \\ y^{(k+1)} &= \beta \sum_{j=0}^{k} \tau_j \tilde{S}_j^k A x^{(j)}, \end{aligned}$$

where we set

$$\tilde{S}_j^k = S_k S_{k-1} \ldots S_j = \prod_{i=j}^{k} S_i.$$

As a consequence, the ϵ-subgradient of $f_1 \circ A$ employed in (5.4) can be expressed as

$$\beta A^T y^{(k+1)} = \beta^2 \sum_{j=0}^{k} \tau_j A^T \tilde{S}_j^k A x^{(j)}. \tag{6.4}$$

The following simple lemma, which directly follows from the definition of A, indicates a possible decomposition of each term in the summation at the right-hand side of (6.4) as the difference between a positive and a nonnegative term.

LEMMA 6.1. *Every matrix–vector product of the form $A^T S A x$, where S is a $2n \times 2n$ diagonal matrix with positive entries such that $S_{2\ell,2\ell} = S_{2\ell-1,2\ell-1} = s_\ell$, $\ell = 1, \ldots, n$, $x \geq 0$, can be decomposed as*

$$A^T S A x = V_S x - U_S x,$$

where

$$\begin{aligned} (V_S x)_{i,j} &= (2 s_{i,j} + s_{i,j-1} + s_{i-1,j}) x_{i,j} \geq 0, \\ (U_S x)_{i,j} &= s_{i,j}(x_{i+1,j} + x_{i,j+1}) + s_{i,j-1} x_{i,j-1} + s_{i-1,j} x_{i-1,j} \geq 0 \end{aligned}$$

with the correspondence $s_\ell \equiv s_{i,j}$, $j = \lfloor (\ell-1)/N \rfloor + 1$, $i = \ell - \lfloor (\ell-1)/N \rfloor \cdot N$.

Algorithm 2 SPDHG

Choose the starting point $x^{(0)} \in X$ and set $y^{(0)} = 0$, $p^{(-1)} = q^{(-1)} = r^{(-1)} = 0$. Choose the sequences $\{\alpha_k\}$, $\{\tau_k\}$, $\{\gamma_k\}$.

FOR $k = 0, 1, 2, \ldots$ DO THE FOLLOWING STEPS:

STEP 1. Compute $\tilde{y}^{(k)} = y^{(k)} + \beta\tau_k A x^{(k)}$;

STEP 2. Compute $s_\ell^{(k)} = \frac{1}{\max\{1, \|\tilde{y}_\ell^{(k)}\|\}}$, $\ell = 1, n$; $(S_k)_{2i-1,2i-1} = (S_k)_{2i,2i} = \frac{1}{\max\{1, \|\tilde{y}_i^{(k)}\|\}}$.

STEP 3. Dual update: $y^{(k+1)} = S_k \tilde{y}^{(k)}$.

STEP 4. Auxiliary vectors update for the decomposition:

$$p_{i,j}^{(k)} = (p_{i,j}^{(k-1)} + \beta^2 \tau_k x_{i,j}^{(k)}) s_{i,j}^{(k)}, \tag{6.5}$$

$$q_{i,j}^{(k)} = (q_{i,j}^{(k-1)} + \beta^2 \tau_k x_{i,j}^{(k)}) s_{i-1,j}^{(k)}, \tag{6.6}$$

$$r_{i,j}^{(k)} = (r_{i,j}^{(k-1)} + \beta^2 \tau_k x_{i,j}^{(k)}) s_{i,j-1}^{(k)} \quad \text{for } i, j = 1, \ldots, N. \tag{6.7}$$

STEP 5. Compute the positive part of the decomposition:

$$V(x^{(k)}) = H^T e + (2p^{(k)} + q^{(k)} + r^{(k)}). \tag{6.8}$$

STEP 6. Compute the scaling matrix:

$$(D_k)_{\ell,\ell} = \min\left\{(1+\gamma_k)^{\frac{1}{2}}, \max\left\{(1+\gamma_k)^{-\frac{1}{2}}, \frac{x_\ell^{(k)}}{V(x^{(k)})_\ell}\right\}\right\}.$$

STEP 7. Primal update: $x^{(k+1)} = P_{\geq 0}(x^{(k)} - \alpha_k D_k(\nabla f_0(x^{(k)}) + \beta A^T y^{(k+1)}))$.

END

The ϵ-subgradient of f in (5.4) can be decomposed as

$$u^{(k)} = \nabla f_0(x^{(k)}) + \beta A^T y^{(k+1)} = V(x^{(k)}) - U(x^{(k)}),$$

where

$$V(x^{(k)}) = H^T e + \beta^2 \sum_{j=0}^{k} \tau_j V_{\tilde{S}_j^k} x^{(j)}. \tag{6.9}$$

Even if it seems quite complicated, the term

$$V^R(x^{(k)}) = \beta^2 \sum_{j=0}^{k} \tau_j V_{\tilde{S}_j^k} x^{(j)}$$

can be easily computed in a recursive way, by introducing three auxiliary vectors, as described in Algorithm 2. We also notice that, since the scaling matrix D_k is diagonal and the constraint set X is the nonnegative orthant, the projection with respect to the norm induced by D_k^{-1} reduces to the usual Euclidean projection $P_{\geq 0}(\cdot)$ (see Step 7).

By induction it can be shown that the computation of $V(x^{(k)})$ in (6.8) actually gives (6.9). For the sake of simplicity, we limit ourselves to show that this is true for $k = 0, 1$. Indeed, from Lemma 6.1 and from (6.5)–(6.7) we have

$$\begin{aligned}
V^R(x^{(0)})_{i,j} &= \beta^2\tau_0(V_{S_0}x^{(0)})_{i,j} \\
&= \beta^2\tau_0(2s^{(0)}_{i,j} + s^{(0)}_{i-1,j} + s^{(0)}_{i,j-1})x^{(0)}_{i,j} \\
&= 2p^{(0)}_{i,j} + q^{(0)}_{i,j} + r^{(0)}_{i,j}, \\
V^R(x^{(1)})_{i,j} &= \beta^2\tau_0(V_{S_0S_1}x^{(0)})_{i,j} + \beta^2\tau_1(V_{S_1}x^{(1)})_{i,j} \\
&= \beta^2\tau_0(2s^{(0)}_{i,j}s^{(1)}_{i,j} + s^{(0)}_{i-1,j}s^{(1)}_{i-1,j} + s^{(0)}_{i,j-1}s^{(1)}_{i,j-1})x^{(0)}_{i,j} \\
&\quad + \beta^2\tau_1(2s^{(1)}_{i,j} + s^{(1)}_{i-1,j} + s^{(1)}_{i,j-1})x^{(1)}_{i,j} \\
&= 2(\beta^2\tau_0 s^{(0)}_{i,j}x^{(0)}_{i,j} + \beta^2\tau_1 x^{(1)}_{i,j})s^{(1)}_{i,j} \\
&\quad + (\beta^2\tau_0 s^{(0)}_{i-1,j}x^{(0)}_{i,j} + \beta^2\tau_1 x^{(1)}_{i,j})s^{(1)}_{i-1,j} \\
&\quad + (\beta^2\tau_0 s^{(0)}_{i,j-1}x^{(0)}_{i,j} + \beta^2\tau_1 x^{(1)}_{i,j})s^{(1)}_{i,j-1} \\
&= 2(p^{(0)}_{i,j} + \beta^2\tau_1 x^{(1)}_{i,j})s^{(1)}_{i,j} + (q^{(0)}_{i,j} + \beta^2\tau_1 x^{(1)}_{i,j})s^{(1)}_{i-1,j} + (r^{(0)}_{i,j} + \beta^2\tau_1 x^{(1)}_{i,j})s^{(1)}_{i,j-1} \\
&= 2p^{(1)}_{i,j} + q^{(1)}_{i,j} + r^{(1)}_{i,j}.
\end{aligned}$$

Algorithm 2 can be adapted for both the step size selection strategies described in sections 3 and 4. In the first case, three prefixed sequences $\{\alpha_k\}$, $\{\tau_k\}$, and $\{\gamma_k\}$ satisfying the assumptions of Corollary 5.1 have to be provided.

In the other case, the SSL procedure for dynamically computing the primal step size α_k can be included in Algorithm 2. Here, only the sequences $\{\tau_k\}$, and $\{\gamma_k\}$ should be given such that $\lim_{k\to\infty}\tau_k = \infty$ and $\sum\gamma_k < \infty$.

7. Numerical experience. The aim of our numerical experience is twofold: first, we are interested in evaluating the effect of the scaling on the convergence behavior of the ϵ-subgradient method. Second, we compare the two step length selection strategies presented in sections 3–4.

To this end we consider four different versions of the method (5.2)–(5.4):

PDHG corresponds to the choices $D_k = I$ and α_k chosen as an a priori diminishing, divergent series, square summable sequence in (5.4). It actually consists in the method in [16];

SPDHG is Algorithm 2 with α_k chosen as an a priori diminishing, divergent series, square summable sequence;

SL is the ϵ-subgradient level method given in Algorithm 1 with $D_k = I$ and $u^{(k)} = \nabla f_0(x^{(k)}) + \beta A^T y^{(k+1)}$, where $y^{(k+1)}$ is updated as in (5.2);

SSL is the same as above but with the scaling matrix D_k defined as at Step 6 of Algorithm 2.

The numerical experiments described in this section have been performed in the MATLAB environment (R2015a) on a PC equipped with an Intel Core i7-3517U processor 1.9 GHz, 8 GB RAM.

In our experiments, we consider problem (5.1), where f_0, f_1 are defined in (6.1)–(6.2) and H represents the convolution operator with a given point spread function (psf). Thus, assuming periodic boundary conditions, the matrix–vector products

involving H can be computed by the fast Fourier transform and, by a simple normalization of the psf, we also have $H^T e = He = e$; moreover, in our experiments, H is nonsingular, although very ill-conditioned, and $g > 0$, so that problem (5.1) has a unique solution [36]. This problem can be solved also by the PSS method in [28] with positive exogenous sequences of step sizes, whose iteration for this application can be expressed as

$$x^{(k+1)} = P_{x\geq 0}\left(x^{(k)} - \frac{\alpha_k}{\max(1, \|\nabla f_0(x^{(k)}) + u^{(k)}\|)}(\nabla f_0(x^{(k)}) + u^{(k)})\right)$$

with $u^{(k)} \in \partial(f_1 \circ A)(x^{(k)})$.

We consider a set of three test problems generated as in [61], where the data g are obtained with the following procedure: the selected original image $\bar{x}$ is rescaled so that the maximum pixel intensity is a specified value $I_{\max}$. Then, the rescaled image is convolved with the psf and the background b is added. Finally, Poisson noise is introduced by the MATLAB `imnoise` function and the simulated data g are obtained after scaling back again by $I_{\max}$.

For each test problem, the regularization parameter β has been empirically selected by computing the solution of (5.1) for different values of β and choosing that for which we observed the minimum l_2 relative distance with respect to $\bar{x}$.

The features of each test problem are specified below.

cameraman: the 256×256 original image is the "cameraman" available in the MATLAB package, while the psf is a Gaussian function, with standard deviation 1.3, truncated at the 9×9 central pixels. The other parameters are $I_{\max} = 1000$, $b = 0$, $\beta = 0.005$; the l_2 relative distance between $\bar{x}$ and g is 0.1209, while $g_i \in [4, 250]$.

micro: the original image is the confocal microscopy phantom of size 128×128 described in [67], scaled by 10; the psf is the one in [67] truncated at the 9×9 central pixels. Here we set $I_{\max} = 1$, $b = 0$, $\beta = 0.0477$; the original image pixels are in the range $[10, 690]$, the l_2 relative distance between $\bar{x}$ and g is 0.1442, while $g_i \in [1, 778]$.

phantom: the original image $\bar{x}$ is the 256×256 Shepp–Logan phantom, generated by the MATLAB function `phantom`, scaled by a factor 1000, while the psf is a Gaussian function, with standard deviation 3, truncated at the 9×9 central pixels. In this case we set $I_{\max} = 1$, $b = 10$, and $\beta = 0.00526$. The values of the original image are in the range $[0, 1000]$, the l_2 relative distance between $\bar{x}$ and g is 0.4643, while $g_i \in [1, 934]$.

For all test problems we compute the solution x^* of the minimization problem (5.1) by running 50000 iterations of the PIDSplit method [61]. Then, we evaluate the progress toward this solution at each iteration in terms of the l_2 relative error from the minimum point and the relative difference from the optimal value,

$$e^k = \frac{\|x^{(k)} - x^*\|}{\|x^*\|}, \quad f^k = \frac{f(x^{(k)}) - f(x^*)}{f(x^*)}.$$

Following the assumptions of Corollaries 5.1 and 5.2 and those in [28] for PSS, we choose the sequences of parameters as follows:

$$\tau_k = t_1 + t_2 k, \quad \alpha_k = \frac{1}{t_3 + t_4 k}, \quad \gamma_k = \frac{t_5}{k^{1+t_6}}.$$

In order to illustrate the effectiveness of the methods, the values t_i have been manually optimized for each test problem to obtain a faster decrease of e^k (see Table 1).

Moreover, for the initialization of both SL and SSL, we adopt the rule $\delta_0 = 0.9\, f(x^{(0)})$, while the other parameters are $\nu_1 = \nu_2 = 0.5$, $B = 0.9\|u^{(0)}\|\|D_0\|_\infty^{\frac{1}{2}}$. The

TABLE 1
Parameter settings.

	PDHG		SPDHG		
	τ_k	α_k	τ_k	α_k	γ_k
Cameraman	$0.9 + 10^{-2}k$	$(0.05 + 10^{-5}k)^{-1}$	$0.5 + 5 \cdot 10^{-3}k$	$(0.5 + 10^{-6}k)^{-1}$	$10^{13}k^{-2}$
Micro	$0.9 + 10^{-3}k$	$(0.05 + 10^{-5}k)^{-1}$	$0.4 + 10^{-3}k$	$(0.5 + 10^{-7}k)^{-1}$	$10^{13}k^{-2}$
Phantom	$0.9 + 10^{-3}k$	$(0.01 + 10^{-5}k)^{-1}$	$0.5 + 10^{-4}k$	$(0.5 + 10^{-6}k)^{-1}$	$10^{13}k^{-2}$

	PSS	SL	SSL	
	α_k	τ_k	τ_k	γ_k
Cameraman	$(0.004 + 10^{-6}k)^{-1}$	$0.5 + 5 \cdot 10^{-2}k$	$0.7 + 5 \cdot 10^{-2}k$	$10^{13}k^{-2}$
Micro	$(0.002 + 10^{-6}k)^{-1}$	$0.9 + 10^{-1}k$	$0.9 + 10^{-2}k$	$10^{13}k^{-2}$
Phantom	$(0.002 + 10^{-6}k)^{-1}$	$0.9 + 10^{-2}k$	$0.9 + 10^{-2}k$	$10^{13}k^{-2}$

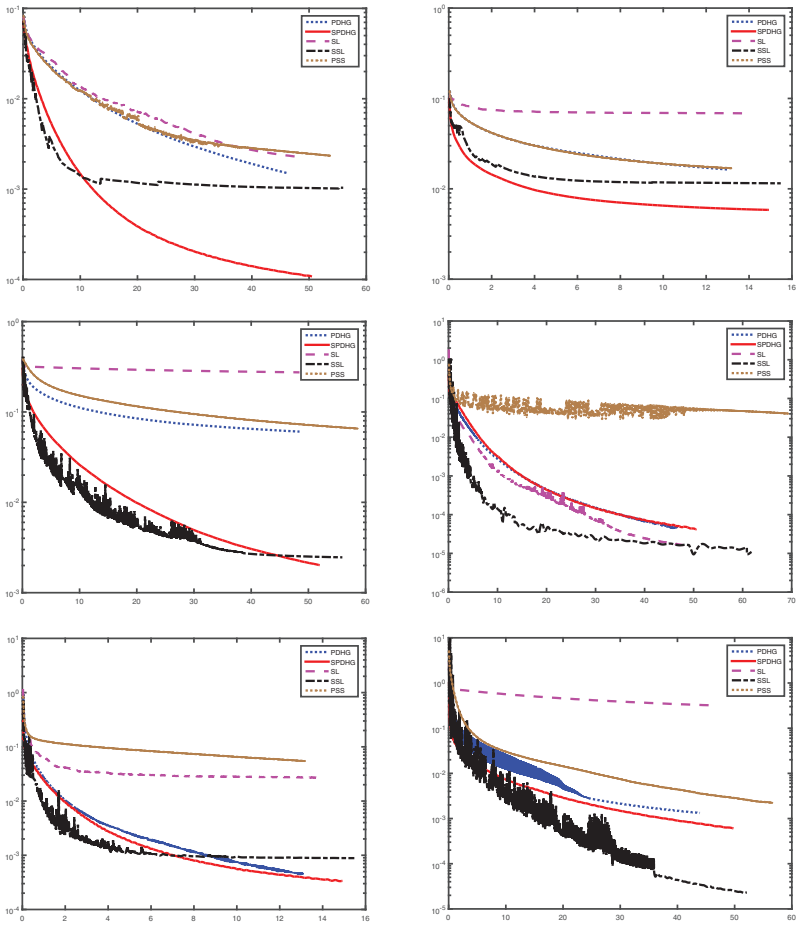

FIG. 3. *Image deblurring results. Upper row: plots of the relative minimization error e^k. Lower row: plots of the relative difference from the optimal function value f^k. Left column:* cameraman. *Middle column:* micro. *Right column:* phantom. *All plots are with respect to the computational time in seconds and use a logarithmic scale on the vertical axis.*

plots in Figure 3 have been obtained by running 3000 iterations of the algorithms, reporting the errors e^k, f^k with respect to the computational time in seconds.

From the numerical experience we observe that the presence of the scaling can help to accelerate the progress towards the solution, with both step size selection

strategies. As concerns the scaling matrix bounds, the best results are obtained by selecting large initial values for γ_k (see Table 1) and, thus, for L_k, allowing more freedom to choose the scaling matrix especially at the first iterations.

It is also interesting to observe that the adaptive computation of α_k combined with the proposed scaling technique in SSL seems to work quite well, leading to performances that are, in some cases, close to the "best" ones obtainable by manually tuning the step size sequences in PDHG, SPDHG, and PSS. Indeed, the performances of algorithms depending on exogeneous sequences of step sizes are sensitive to the choice of these parameters, making it difficult to devise a general rule to select them.

Comparing the first and the second rows in Figure 3, we also observe that a faster approach to the solution x^* does not always correspond to a faster decrease of the objective function: we suppose that this phenomenon is due to the ill-conditioning of problem (5.1). For completeness, we experimentally observed that too large initial values of the primal step size α_k in PDHG and SPDHG may produce an unbounded sequence $\{u^{(k)}\}$ and, as a consequence, the algorithms fail to converges: this indicates that the assumption on the ϵ-subgradient boundedness is crucial.

8. Conclusions. In this paper we proposed a generalization of the ϵ-subgradient method with variable scaling matrix for nonsmooth, convex optimization, developing the related convergence analysis when the step size parameters are either provided as a priori selected sequences or dynamically computed by an adaptive procedure. Exploiting the duality principle, we described a special case of the proposed method which applies to the minimization of the sum of two convex functions with a composite term. For a specific problem of this form in the image restoration framework, we fully detailed the algorithm, also suggesting a strategy to compute the scaling matrix. The numerical experience shows that the presence of a suitable variable scaling matrix can accelerate the progress of the iterates towards the solution. Moreover, the results obtained combining the variable scaling with the adaptive procedure for the computation of the step size parameter are encouraging.

Future work will be addressed to further investigate dynamic choices of the step size and of the scaling matrix, with the aim to devise effective "black–box" algorithms which are able to handle practical applications with a minimum of user supplied parameters.

Acknowledgment. We thank the anonymous reviewers for their careful reading and their many insightful comments and suggestions which stimulated us to improve our paper.

REFERENCES

[1] Y. I. Alber, A. N. Iusem, and M. V. Solodov, *On the projected subgradient method for nonsmooth convex optimization in a Hilbert space*, Math. Program., 81 (1998), pp. 23–35.

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

[3] A. Auslender and M. Teboulle, *Projected subgradient methods with non-Euclidean distances for non–differentiable convex minimization and variational equalities*, Math. Program. Ser. B, 120 (2009), pp. 27–48.

[4] J. M. Bardsley and A. Luttman, *Total variation-penalized Poisson likelihood estimation for ill-posed problems*, Adv. Comput. Math., 31 (2009), pp. 35–59.

[5] J. M. Bardsley and J. G. Nagy, *Covariance-preconditioned iterative methods for nonnegatively constrained astronomical imaging*, SIAM J. Matrix Anal. A., 27 (2006), pp. 1184–1197.

[6] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.
[7] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books Math. Ouvrages Math. SMC, Springer, New York, 2011.
[8] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini, *Image deblurring with Poisson data: From cells to galaxies*, Inverse Problems, 25 (2009), 123006.
[9] M. Bertero, H. Lantéri, and L. Zanni, *Iterative image reconstruction: A point of view*, in Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT), Y. Censor, M. Jiang, and A. K. Louis, eds., Birkhäuser-Verlag, Pisa, Italy, 2008, pp. 37–63.
[10] D. Bertsekas, *Convex optimization algorithms*, supplementary online chapter of Convex Optimization Theory, Athena Scientific, Nashua, NH, 2009, pp. 251–489; available online at http://www.athenasc.com/convexdualitychapter.pdf (accessed November 5, 2012).
[11] S. Bonettini, A. Chiuso, and M. Prato, *A scaled gradient projection method for Bayesian learning in dynamical systems*, SIAM J. Sci. Comput., 37 (2015), pp. A1297–A1318.
[12] S. Bonettini, G. Landi, E. L. Piccolomini, and L. Zanni, *Scaling techniques for gradient projection-type methods in astronomical image deblurring*, Int. J. Comput. Math., 90 (2013), pp. 9–29.
[13] S. Bonettini, I. Loris, F. Porta, and M. Prato, *Variable metric inexact line–search-based methods for nonsmooth optimization*, SIAM J. Optim., 26 (2016), pp. 891–921.
[14] S. Bonettini and M. Prato, *Nonnegative image reconstruction from sparse Fourier data: A new deconvolution algorithm*, Inverse Problems, 26 (2010), 095001.
[15] S. Bonettini and V. Ruggiero, *An alternating extragradient method for total variation based image restoration from Poisson data*, Inverse Problems, 27 (2011), 095001.
[16] S. Bonettini and V. Ruggiero, *On the convergence of primal-dual hybrid gradient algorithms for total variation image restoration*, J. Math. Imaging Vision, 44 (2012), pp. 236–253.
[17] S. Bonettini and T. Serafini, *Non-negatively constrained image deblurring with an inexact interior point method*, J. Comput. Appl. Math., 231 (2009), pp. 236–248.
[18] S. Bonettini, R. Zanella, and L. Zanni, *A scaled gradient projection method for constrained image deblurring*, Inverse Problems, 25 (2009), 015002.
[19] U. Brännlund, K. C. Kiwiel, and P. O. Lindberg, *A descent proximal level bundle method for convex nondifferentiable optimization*, Oper. Res. Lett., 17 (1995), pp. 121–126.
[20] K. Bredies and D. A. Lorentz, *Linear convergence of iterative soft-thresholding*, J. Fourier Anal. Appl., 14 (2008), pp. 813–837.
[21] Y. Chen, W. W. Hager, M. Yashtini, X. Ye, and H. Zhang, *Bregman operator splitting with variable stepsize for total variation image reconstruction*, Comput. Optim. Appl., 54 (2012), pp. 317–342.
[22] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, *Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function*, J. Optim. Theory Appl., 162 (2014), pp. 107–132.
[23] P. L. Combettes, *Quasi-Féjerian analysis of some optimization algorithms*, Stud. Comput. Math., 8 (2001), pp. 115–152.
[24] P. L. Combettes and B. C. Vũ, *Variable metric quasi-Féjer monotonicity*, Nonlinear Anal., 78 (2013), pp. 17–31.
[25] P. L. Combettes and B. C. Vũ, *Variable metric forward-backward splitting with applications to monotone inclusions in duality*, Optimization, 63 (2014), pp. 1289–1318.
[26] L. Condat, *A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, J. Optim. Theory Appl., 158 (2013), pp. 460–479.
[27] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Program., 62 (1993), pp. 261–275.
[28] J. Y. B. Cruz, *On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions*, Set-Valued Var. Anal., in press, doi:10.1007/s11228-016-0376-5.
[29] J. Y. B. Cruz and T. T. A. Nghia, *On the convergence of the forward–backward splitting method with linesearches*, Optim. Method Softw., in press, doi:10.1080/10556788.2016.1214959.
[30] Y. H. Dai, W. W. Hager, K. Schittkowski, and H. Zhang, *The cyclic Barzilai-Borwein method for unconstrained optimization*, IMA J. Numer. Anal., 26 (2006), pp. 604–627.
[31] G. d'Antonio and A. Frangioni, *Convergence analysis of detected conditional approximate subgradient methods*, SIAM J. Optim., 20 (2009), pp. 357–386.
[32] F.-X. Dupé, M. Fadili, and J.-L. Starck, *Inverse problems with Poisson noise: Primal and primal-dual splitting*, in 18th IEEE International Conference on Image Processing (ICIP), Brussels, 2011, IEEE, Piscataway, NJ, 2011, pp. 1901–1904.

[33] F.-X. Dupé, M. Fadili, and J.-L. Starck, *Linear inverse problems with various noise models and mixed regularizations*, in Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools, Brussels, 2011, Springer, Berlin, 2011, pp. 621–626.
[34] Y. M. Ermoliev, *Stochastic quasigradient methods and their application to system optimization*, Stochastics, 9 (1983), pp. 1–36.
[35] E. Esser, X. Zhang, and T. F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM J. Imaging Sci., 3 (2010), pp. 1015–1046.
[36] M. A. T. Figueiredo and J. M. Bioucas-Dias, *Restoration of Poissonian images using alternating direction optimization*, IEEE Trans. Image Process., 19 (2010), pp. 3133–3145.
[37] R. Fletcher, *A limited memory steepest descent method*, Math. Program., 135 (2012), pp. 413–436.
[38] G. Frassoldati, G. Zanghirati, and L. Zanni, *New adaptive stepsize selections in gradient methods*, J. Ind. Manag. Optim., 4 (2008), pp. 299–312.
[39] J. L. Goffin, *On convergence rates of subgradient optimization methods*, Math. Program., 13 (1977), pp. 329–347.
[40] J. L. Goffin and K. C. Kiwiel, *Convergence of a simple subgradient level method*, Math. Program., 85 (1999), pp. 207–211.
[41] W. W. Hager, B. A. Mair, and H. Zhang, *An affine-scaling interior-point CBB method for box-constrained optimization*, Math. Program., 119 (2009), pp. 1–32.
[42] W. W. Hager, M. Yashtini, and H. Zhang, *An $O(1/k)$ convergence rate for the variable stepsize Bregman operator splitting algorithm*, SIAM J. Numer. Anal., 54 (2016), pp. 1535–1556.
[43] B. He, Y. You, and X. Yuan, *On the convergence of primal-dual hybrid gradient algorithm*, SIAM J. Imaging Sci., 7 (2014), pp. 2526–2537.
[44] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, New York, 1993.
[45] K. C. Kiwiel, *The effciency of subgradient projection methods for convex optimization, Part* I: *General level methods*, SIAM J. Control Optim., 34 (1996), pp. 660–676.
[46] K. C. Kiwiel, *Effciency of subgradient projection methods for convex optimization, Part* II: *Implications and Extensions*, SIAM J. Control Optim., 34 (1996), pp. 677–697.
[47] K. C. Kiwiel, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM J. Optim., 14 (2004), pp. 807–840.
[48] H. Lantéri, M. Roche, and C. Aime, *Penalized maximum likelihood image restoration with positivity constraints: Multiplicative algorithms*, Inverse Problems, 18 (2002), pp. 1397–1419.
[49] 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 (2001), pp. 945–974.
[50] T. Larsson, M. Patriksson, and A.-B. Strömberg, *On the convergence of conditional ϵ-subgradient methods for convex programs and convex–concave saddle-point problems*, European J. Oper. Res., 151 (2003), pp. 461–473.
[51] D. A. Lorenz and T. Pock, *An inertial forward-backward algorithm for monotone inclusions*, J. Math. Imaging Vision, 51 (2015), pp. 311–325.
[52] A. Nedić, *Subgradient Methods for Convex Minimization*, Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA, 2002, http://hdl.handle.net/1721.1/16843.
[53] A. Nedic and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.
[54] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, Boston, 2004.
[55] E. S. H. Neto and A. R. D. Pierro, *Incremental subgradients for constrained convex optimization: A unified framework and new methods*, Math. Program., 103 (2005), pp. 127–152.
[56] B. Polyak, *Introduction to Optimization*, Optimization Software, NY, 1987.
[57] M. Prato, A. L. Camera, S. Bonettini, and M. Bertero, *A convergent blind deconvolution method for post-adaptive-optics astronomical imaging*, Inverse Problems, 29 (2013), 065017.
[58] S. M. Robinson, *Linear convergence of epsilon-subgradient descent methods for a class of convex functions*, Math. Program., Ser. A, 86 (1999), pp. 41–50.
[59] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[60] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, J. Phys. D., 60 (1992), pp. 259–268.

[61] S. Setzer, G. Steidl, and T. Teuber, *Deblurring Poissonian images by split Bregman techniques*, J. Vis. Commun. Image Represent., 21 (2010), pp. 193–199.
[62] S. Shalev-Shwartz, *Online learning and online convex optimization*, Found. Trends Mach. Learn., 4 (2011), pp. 107–194.
[63] N. Z. Shor, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, 1985.
[64] M. V. Solodov and S. K. Zaviev, *Error stability properties of generalized gradient-type algorithms*, J. Optim. Theory Appl., 98 (1998), pp. 663–680.
[65] C. Theys, N. Dobigeon, J.-Y. Tourneret, and H. Lantéri, *Linear unmixing of hyperspectral images using a scaled gradient method*, in Proceedings of the IEEE/SP 15th Workshop on Statistical Signal Processing, IEEE, Piscataway, NJ, 2009, pp. 729–732.
[66] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, *Accelerated and inexact forward-backward algorithms*, SIAM J. Optim., 23 (2013), pp. 1607–1633.
[67] R. M. Willett and R. D. Nowak, *Platelets: A multiscale approach for recovering edges and surfaces in photon limited medical imaging*, IEEE Trans. Med. Imaging, 22 (2003), pp. 332–350.
[68] R. Zanella, P. Boccacci, L. Zanni, and M. Bertero, *Efficient gradient projection methods for edge-preserving removal of Poisson noise*, Inverse Problems, 25 (2009), 045010.