SIAM J. OPTIM.
Vol. 26, No. 2, pp. 891–921

VARIABLE METRIC INEXACT LINE-SEARCH-BASED METHODS FOR NONSMOOTH OPTIMIZATION*

S. BONETTINI†, I. LORIS‡, F. PORTA†, AND M. PRATO§

Abstract. We develop a new proximal-gradient method for minimizing the sum of a differentiable, possibly nonconvex, function plus a convex, possibly nondifferentiable, function. The key features of the proposed method are the definition of a suitable descent direction, based on the proximal operator associated to the convex part of the objective function, and an Armijo-like rule to determine the stepsize along this direction ensuring the sufficient decrease of the objective function. In this frame, we especially address the possibility of adopting a metric which may change at each iteration and an inexact computation of the proximal point defining the descent direction. For the more general nonconvex case, we prove that all limit points of the iterates sequence are stationary, while for convex objective functions we prove the convergence of the whole sequence to a minimizer, under the assumption that a minimizer exists. In the latter case, assuming also that the gradient of the smooth part of the objective function is Lipschitz, we also give a convergence rate estimate, showing the $\mathcal{O}(\frac{1}{k})$ complexity with respect to the function values. We also discuss verifiable sufficient conditions for the inexact proximal point and present the results of two numerical tests on total-variation-based image restoration problems, showing that the proposed approach is competitive with other state-of-the-art methods.

Key words. proximal algorithms, nonsmooth optimization, generalized projection, nonconvex optimization

AMS subject classifications. 65K05, 90C30

DOI. 10.1137/15M1019325

1. Introduction.

In this paper we consider the problem

$$\min_{x\in\mathbb{R}^n} f(x) \equiv f_0(x) + f_1(x), \tag{1}$$

where f_1 is a proper, convex, lower semicontinuous function and f_0 is smooth, i.e., continuously differentiable, on an open subset Ω_0 of $\mathbb{R}^n$ containing $\mathrm{dom}(f_1) = \{x \in \mathbb{R}^n : f_1(x) < +\infty\}$. We also assume that f_1 is bounded from below and that $\mathrm{dom}(f_1)$ is nonempty and closed. Formulation (1) includes constrained problems over convex sets, which can be introduced by adding to f_1 the indicator function of the feasible set.

When in particular f_1 reduces to the indicator function of a convex set Ω, i.e., $f_1 = \iota_\Omega$, with

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

*Received by the editors April 30, 2015; accepted for publication (in revised form) January 4, 2016; published electronically April 5, 2016. This work was partially supported by MIUR under the two projects FIRB - Futuro in Ricerca 2012 (contract RBFR12M3AC) and PRIN 2012 (contract 2012MTE38N). The Italian GNCS - INdAM is also acknowledged.

http://www.siam.org/journals/siopt/26-2/M101932.html

†Dipartimento di Matematica e Informatica, Università di Ferrara, Via Saragat 1, 44122 Ferrara, Italy (silvia.bonettini@unife.it, federica.porta@unife.it).

‡Département de Mathématique, Université Libre de Bruxelles, Boulevard du Triomphe, 1050 Bruxelles, Belgium (igloris@ulb.ac.be). This author is a Research Associate of the Fonds de la Recherche Scientifique - FNRS.

§Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, Via Campi 213/b, 41125 Modena, Italy (marco.prato@unimore.it).

a simple and well-studied algorithm for the solution of (1) is the gradient projection method, which is particularly appealing for large-scale problems. Several variants of such methods have been proposed [7, 11, 20, 23], with the aim to accelerate the convergence, which, for the basic implementation, can be very slow. In particular, reliable acceleration techniques have been proposed for the so-called gradient projection method with line-search along the feasible direction [6, Chapter 2], whose iteration consists of

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

where $y^{(k)}$ is the Euclidean projection of the point $x^{(k)} - \nabla f_0(x^{(k)})$ onto the feasible set Ω, and $\lambda^{(k)} \in [0, 1]$ is a steplength parameter ensuring the sufficient decrease of the objective function. Typically, $\lambda^{(k)}$ is determined by a backtracking loop until an Armijo-type inequality is satisfied. Variants of the basic scheme are obtained by introducing a further variable stepsize parameter α_k, which controls the step along the gradient, in combination with a variable choice of the underlying metric. In practice, the point $y^{(k)}$ can be defined as

$$y^{(k)} = \arg\min_{y\in\Omega} \nabla f_0(x^{(k)})^T(y - x^{(k)}) + \frac{1}{2\alpha_k}(y - x^{(k)})^T D_k(y - x^{(k)}), \tag{3}$$

where α_k is a positive parameter and $D_k \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. The stepsizes α_k and the matrices D_k have to be considered as "free" parameters of the method, and a clever choice of them can lead to significant improvements in the practical convergence behavior [7, 9, 11].

In this paper we generalize the gradient projection scheme (2)–(3) by introducing the concept of descent direction for the case where f_1 is a general convex function and we propose a suitable variant of the Armijo rule for the nonsmooth problem (1). In particular, we focus on the case when the descent direction has the form $y^{(k)} - x^{(k)}$, with

$$y^{(k)} = \arg\min_{y\in\mathbb{R}^n} \nabla f_0(x^{(k)})^T(y - x^{(k)}) + d_{\sigma^{(k)}}(y, x^{(k)}) + f_1(y) - f_1(x^{(k)}), \tag{4}$$

where $d_{\sigma^{(k)}}(\cdot,\cdot)$ plays the role of a distance function, depending on the parameter $\sigma^{(k)} \in \mathbb{R}^q$. Clearly, (4) is a generalization of (3), which is recovered when $f_1 = \iota_\Omega$, by setting $d_\sigma(y, x) = \frac{1}{\alpha}(y - x)^T D(y - x)$, with $\sigma = (\alpha, D)$.

Formally, the scheme (2)–(4) is a forward–backward (or proximal-gradient) method [17, 18] depending on the parameters $\lambda^{(k)}$, $\sigma^{(k)}$. In particular, we deeply investigate the variant of the scheme (2)–(4) where the minimization problem in (4) is solved inexactly, and we devise two types of admissible approximations. We show that both approximation types can be practically computed when $f_1(x) = g(Ax)$, where $A \in \mathbb{R}^{m\times n}$ and $g : \mathbb{R}^m \to \mathbb{R}$ is a proper, convex, lower semicontinuous function with an easy-to-compute resolvent operator. In this case, our scheme consists in a double-loop method, where the inner loop is provided by an implementable stopping criterion. For general f_0, we are able to prove that any limit point of the sequence generated by our inexact scheme is stationary for problem (1). The proof of this fact is essentially based on the properties of the Armijo-type rule adopted for computing $\lambda^{(k)}$, and it does not require any Lipschitz property of the gradient of f_0. When f_0 is convex, we prove a stronger result, showing that the iterates converge to a minimizer of (1), if it exists. In the latter case, under the further assumption that ∇f_0 is Lipschitz continuous, we give an $\mathcal{O}(\frac{1}{k})$ convergence rate estimate for the objective function values. Our analysis includes as special cases several state-of-the-art methods, as those in [7, 10, 11, 28, 35].

Forward–backward algorithms based on a variable metric have been recently studied also in [16] for the convex case and in [15] for the nonconvex case under the Kurdyka–Łojasiewicz assumption (see also [22]). Even if our scheme is formally very similar to those in [15, 16], the involved parameters have a substantially different meaning. In our case, the theoretical convergence is ensured by the Armijo parameter $\lambda^{(k)}$ in combination with the descent direction properties; this results in an almost complete freedom to choose the other algorithm parameters (e.g., α_k and D_k), without necessarily relating them to the Lipschitz constant of ∇f_0 (actually, our analysis, except for the convergence rate estimate, is performed without this assumption). We believe that this is also one of the main strengths of our method, since acceleration techniques based on suitable choices of α_k and D_k, originally proposed for smooth optimization, can be adopted, leading to an improvement of the practical performances. The other crucial ingredient of our method is the inexact computation of the minimizer in (4): this issue has been considered in several papers in the context of proximal and proximal-gradient methods (see, for example, [1, 15, 34, 36] and references therein). The approach we follow in this paper is more similar to the one proposed in [36] and has the advantage of providing an implementable condition for the approximate computation of the proximal point. Moreover, we also generalize the ideas proposed in [7] for the inexact computation of the projection onto a convex set. Finally, we also mention the papers [2, 3, 4, 21] for the use of non-Euclidean distances in the context of forward–backward and proximal methods.

The paper is organized as follows: In section 2 the concept of descent direction for problem (1) is presented and developed, while in section 3 the modified Armijo rule is discussed. Then a general convergence result for line-search descent algorithms based on this rule is proved in the nonconvex case. Two different inexactness criteria, called ϵ-type and η-type, are proposed in sections 3.2 and 3.3, and the related implementation is discussed in sections 4.1 and 4.4. Section 3.5 deals with the convex case, where the convergence of an ϵ-approximation-based algorithm is proved and the related convergence rate is analyzed. The results of two numerical tests on total-variation-based image restoration problems are presented in section 5, and our conclusions are given in section 6.

Notation. We denote the extended real numbers set as $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and by $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{>0}$ the set of nonnegative and positive real numbers, respectively. The scaled Euclidean norm of an n-vector x, associated to a symmetric positive definite matrix D, is $\|x\|_D = \sqrt{x^T Dx}$. Given $\mu \geq 1$, we denote by $\mathcal{M}_\mu$ the set of all symmetric positive definite matrices with all eigenvalues contained in the interval $[\frac{1}{\mu}, \mu]$. For any $D \in \mathcal{M}_\mu$ we have that D^{-1} also belongs to $\mathcal{M}_\mu$ and

$$\frac{1}{\mu}\|x\|^2 \leq \|x\|_D^2 \leq \mu\|x\|^2 \tag{5}$$

for any $x \in \mathbb{R}^n$.

2. A family of descent directions.

When f is smooth, a vector $d \in \mathbb{R}^n$ is said to be a descent direction for f at x when $\nabla f(x)^T d < 0$. In the nonsmooth case (1), a vector $d \in \mathbb{R}^n$ is a *descent direction* for f at $x \in \mathrm{dom}(f)$ if $f'(x; d) < 0$, where $f'(x; d)$ is the *one-sided directional derivative* of f at x with respect to a vector d defined as (see [32, p. 213])

$$f'(x;d) = \lim_{\lambda\downarrow 0} \frac{f(x+\lambda d) - f(x)}{\lambda} \tag{6}$$

if the limit on the right-hand side exists in $\bar{\mathbb{R}}$. Thanks to [32, Theorem 23.1], the previous definition is well posed. In this section we define a family of descent directions for problem (1). To this end, we define the following set of nonnegative functions.

Given a convex set $\Omega \subseteq \mathbb{R}^n$ and a set of parameters $S \subseteq \mathbb{R}^q$, we denote by $\mathcal{D}(\Omega, S)$ the set of any *distance-like* function $d_\sigma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ continuously depending on $\sigma \in S$ such that for all $z, x \in \Omega$

($\mathcal{D}_1$) $d_\sigma(z,x)$ is continuous in (σ, z, x);

($\mathcal{D}_2$) $d_\sigma(z,x)$ is smooth with respect to $z \in \Omega$;

($\mathcal{D}_3$) $d_\sigma(z,x)$ is strongly convex with respect to z:

$$d_\sigma(z_2,x) \geq d_\sigma(z_1,x) + \nabla_1 d_\sigma(z_1,x)^T(z_2 - z_1) + \frac{m}{2}\|z_2 - z_1\|^2 \qquad \forall z_1, z_2 \in \Omega,$$

where $m > 0$ does not depend on σ or x (here ∇_1 denotes the gradient with respect to the first argument of a function);

($\mathcal{D}_4$) $d_\sigma(z,x) = 0$ if and only if $z = x$ (which implies that $\nabla_1 d_\sigma(x,x) = 0$ for all $x \in \Omega$).

The scaled Euclidean distance

$$d_\sigma(x,y) = \frac{1}{2\alpha}\|x - y\|_D^2 \tag{7}$$

with $\sigma = (\alpha, D)$, where $\alpha > 0$ and $D \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, is an interesting example of a function in $\mathcal{D}(\mathbb{R}^n, S)$. Other examples of distance-like functions can be obtained by considering Bregman distances associated to a strongly convex function.

It is well known [6, p. 667] that when ∇f_0 is Lipschitz continuous, and when α is sufficiently small, the following upper bound exists for f:

$$f(z) \leq f(x) + \nabla f_0(x)^T(z - x) + \frac{1}{2\alpha}\|z - x\|^2 + f_1(z) - f_1(x)$$

(equality when $z = x$). In other words, a negative sign of

$$\nabla f_0(x)^T(z - x) + \frac{1}{2\alpha}\|z - x\|^2 + f_1(z) - f_1(x) \tag{8}$$

corresponds to a descent of the function f. Our aim now is to drop the Lipschitz assumptions on f_0 and to generalize the expression (8) for an arbitrary distance function d_σ replacing the Euclidean distance squared.

For a given array of parameters $\sigma \in S \subseteq \mathbb{R}^q$, let us introduce the function $h_\sigma : \mathbb{R}^n \times \mathbb{R}^n \to \bar{\mathbb{R}}$ defined as

$$h_\sigma(z,x) = \nabla f_0(x)^T(z - x) + d_\sigma(z,x) + f_1(z) - f_1(x) \quad \forall z, x \in \mathbb{R}^n, \tag{9}$$

where $d_\sigma \in \mathcal{D}(\Omega, S)$ and $\Omega = \mathrm{dom}(f_1)$.

We remark that h_σ depends continuously on σ, as does d_σ. Moreover, since $d_\sigma(\cdot,x)$ and f_1 are convex, proper, and lower semicontinuous, $h_\sigma(\cdot,x)$ is also convex, proper, and lower semicontinuous for all $x \in \Omega_0$. Finally, for any point $x \in \Omega$ and for any $d \in \mathbb{R}^n$ we have

$$h'_\sigma(x, x; d) = f'(x; d), \tag{10}$$

where $h'_\sigma(z, x; d)$ denotes the directional derivative of $h_\sigma(\,\cdot\,, x)$ at the point z with respect to d.

From assumption $(\mathcal{D}_3)$, it follows that $h_\sigma(\cdot, x)$ is strongly convex and admits a unique minimum point for any $x \in \Omega$.

Now we introduce the following operator $p : \Omega_0 \to \Omega$ associated to any function h_σ of the form (9):

$$p(x; h_\sigma) = \arg\min_{z \in \mathbb{R}^n} h_\sigma(z, x). \tag{11}$$

When d_σ is chosen as in (7), the operator (11) becomes

$$p(x; h_\sigma) = \text{prox}^D_{\alpha f_1}(x - \alpha D^{-1} \nabla f_0(x)),$$

where prox^D_f is the *proximity* or *resolvent* operator associated to a convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ in the metric induced by a symmetric positive definite matrix D, defined as [22, section 2.3]

$$\text{prox}^D_f(x) = \arg\min_{z \in \mathbb{R}^n} f(z) + \frac{1}{2}\|z - x\|^2_D \quad \forall x \in \mathbb{R}^n.$$

Under assumption $(\mathcal{D}_3)$, one can show that $p(x; h_\sigma)$ depends continuously on (x, σ).

PROPOSITION 2.1. *Let $d_\sigma \in \mathcal{D}(\Omega, S)$ and h_σ be defined as in* (9). *Then $p(x; h_\sigma)$ depends continuously on (x, σ).*

Proof. Let $y = \arg\min_{z \in \mathbb{R}^n} h_\sigma(z, x)$. Then y is characterized by the equation $\nabla f_0(x) + \nabla_1 d_\sigma(y, x) + w = 0$, where $w \in \partial f_1(y)$. It follows that $f_1(u) \ge f_1(y) + w^T(u - y)$ for all $u \in \mathbb{R}^n$ or

$$f_1(u) \ge f_1(y) - (\nabla f_0(x) + \nabla_1 d_\sigma(y, x))^T (u - y) \qquad \forall u \in \mathbb{R}^n.$$

Assumption $(\mathcal{D}_3)$ expressed in y and u gives

$$d_\sigma(u, x) \ge d_\sigma(y, x) + \nabla_1 d_\sigma(y, x)^T (u - y) + \frac{m}{2}\|y - u\|^2 \qquad \forall u \in \mathbb{R}^n.$$

Together, these two inequalities yield

$$\frac{m}{2}\|y - u\|^2 \le f_1(u) - f_1(y) + d_\sigma(u, x) - d_\sigma(y, x) + \nabla f_0(x)^T (u - y) \qquad \forall u \in \mathbb{R}^n.$$

Let $y_1 = p(x_1; h_{\sigma_1})$ and $y_2 = p(x_2; h_{\sigma_2})$. Adding the previous inequality for $y = y_1$ (resp., $y = y_2$) and choosing $u = y_2$ (resp., $u = y_1$), one finds

$$\begin{aligned} m\|y_1 - y_2\|^2 \le\ & d_{\sigma_1}(y_2, x_1) - d_{\sigma_1}(y_1, x_1) \\ & + d_{\sigma_2}(y_1, x_2) - d_{\sigma_2}(y_2, x_2) + (\nabla f_0(x_1) - \nabla f_0(x_2))^T (y_2 - y_1), \end{aligned}$$

and hence

$$\begin{aligned} m\|y_1 - y_2\|^2 \le\ & d_{\sigma_2}(y_1, x_2) - d_{\sigma_1}(y_1, x_1) \\ & + d_{\sigma_1}(y_2, x_1) - d_{\sigma_2}(y_2, x_2) + \|\nabla f_0(x_1) - \nabla f_0(x_2)\|\,\|y_2 - y_1\|. \qquad \square \end{aligned}$$

It follows that $0 \leq \|y_1 - y_2\| \leq (b + \sqrt{b^2 + 4cm})/2m$, where $b = \|\nabla f_0(x_1) - \nabla f_0(x_2)\|$ and $c = d_{\sigma_2}(y_1, x_2) - d_{\sigma_1}(y_1, x_1) + d_{\sigma_1}(y_2, x_1) - d_{\sigma_2}(y_2, x_2)$. As f_0 is C^1, one has $\lim_{x_2 \to x_1} b = 0$. As $d_\sigma(z, x)$ is continuous in (σ, z, x), one also has that $\lim_{x_2 \to x_1} c = 0$. This shows then that $\lim_{x_2 \to x_1} \|y_2 - y_1\| = 0$; in other words $p(x_1; h_{\sigma_1})$ is continuous in (σ_1, x_1).

Given a function $d_\sigma \in \mathcal{D}(\Omega, S)$, we introduce also the function $\tilde{h}_{\sigma,\gamma} : \mathbb{R}^n \times \mathbb{R}^n \to \bar{\mathbb{R}}$ defined as

$$\tilde{h}_{\sigma,\gamma}(z, x) = \nabla f_0(x)^T (z - x) + \gamma d_\sigma(z, x) + f_1(z) - f_1(x) \quad \forall z, x \in \mathbb{R}^n \tag{12}$$

for some $\gamma \in [0, 1]$. We have

$$\tilde{h}_{\sigma,\gamma}(y, x) \leq h_\sigma(y, x) \quad \forall x, y \in \mathbb{R}^n \tag{13}$$

and $\tilde{h}_{\sigma,\gamma} = h_\sigma$ when $\gamma = 1$. In the following we show that

- the stationarity condition $f'(x; d) \geq 0$ for all $d \in \mathbb{R}^n$ [35, p. 394] can be reformulated in terms of fixed points of the operator $p(\,\cdot\,; h_\sigma)$;
- the negative sign of $\tilde{h}_{\sigma,\gamma}$ detects a descent direction.

To this purpose, we collect in the following proposition some properties of the function h_σ and the associated operator $p(\,\cdot\,; h_\sigma)$.

Proposition 2.2. *Let $\sigma \in S \subseteq \mathbb{R}^q$, $\gamma \in [0, 1]$, and h_σ, $\tilde{h}_{\sigma,\gamma}$ be defined as in* (9), (12)*, where $d_\sigma \in \mathcal{D}(\Omega, S)$. If $x \in \Omega$ and $y = p(x; h_\sigma)$, then*

(a) *$\tilde{h}_{\sigma,\gamma}(x, x) = 0$;*
(b) *if $z \in \mathbb{R}^n$ and $\tilde{h}_{\sigma,\gamma}(z, x) < 0$, then $f'(x; z - x) < 0$;*
(c) *$\tilde{h}_{\sigma,\gamma}(y, x) \leq 0$ ($\tilde{h}_{\sigma,\gamma}(y, x) = 0 \Leftrightarrow y = x$);*
(d) *$f'(x; y - x) \leq 0$ and the equality holds if and only if $\tilde{h}_{\sigma,\gamma}(y, x) = 0$ (if and only if $x = y$).*

Proof. (a) This is a direct consequence of definition (12) and condition $(\mathcal{D}_3)$ on d_σ.

(b) If $\tilde{h}_{\sigma,\gamma}(z, x) < 0$, we have

$$\begin{aligned} 0 &\geq -\gamma d_\sigma(z, x) > \nabla f_0(x)^T (z - x) + f_1(z) - f_1(x) \\ &\geq \nabla f_0(x)^T (z - x) + f_1'(x; z - x) = f'(x; z - x), \end{aligned}$$

where the second inequality follows from definition (12) of $\tilde{h}_{\sigma,\gamma}$ and the third one from [32, Theorem 23.1].

(c) Since y is the minimum point of $h_\sigma(\,\cdot\,, x)$, part (a) with $\gamma = 1$ yields $h_\sigma(y, x) \leq 0$, which, in view of (13), gives $\tilde{h}_{\sigma,\gamma}(y, x) \leq 0$. If $y = x$, part (a) implies $\tilde{h}_{\sigma,\gamma}(y, x) = 0$. Conversely, assume $\tilde{h}_{\sigma,\gamma}(y, x) = 0$. From inequality (13) we have $h_\sigma(y, x) \geq 0$. On the other side, since y is the minimum point of $h_\sigma(\cdot, x)$, part (a) with $\gamma = 1$ implies $h_\sigma(y, x) \leq 0$. Thus $h_\sigma(y, x) = 0$, and since y is the unique minimizer of $h_\sigma(\cdot, x)$, we can conclude that $x = y$.

(d) From (c) we have $\tilde{h}_{\sigma,\gamma}(y, x) \leq 0$. When $\tilde{h}_{\sigma,\gamma}(y, x) < 0$, then part (b) implies $f'(x; y - x) < 0$. When $\tilde{h}_{\sigma,\gamma}(y, x) = 0$, from (c) we obtain $y = x$ and, therefore, $f'(x; y - x) = 0$. Conversely, assume $f'(x; y - x) = 0$. This implies

$$0 = \nabla f_0(x)^T (y - x) + f_1'(x; y - x) \leq \nabla f_0(x)^T (y - x) + f_1(y) - f_1(x) \leq \tilde{h}_{\sigma,\gamma}(y, x).$$

Since $\tilde{h}_{\sigma,\gamma}(y, x) \leq 0$, we necessarily have $\tilde{h}_{\sigma,\gamma}(y, x) = 0$. ∎

The following proposition completely characterizes the stationary points of (1) in two equivalent ways: as fixed points of the operator $p(\cdot; h_\sigma)$, i.e., the solutions of the equation $x = p(x; h_\sigma)$, or as roots of the composite function $r_{\sigma,\gamma}(x) = \tilde{h}_{\sigma,\gamma}(p(x; h_\sigma), x)$.

PROPOSITION 2.3. *Let $S \subseteq \mathbb{R}^q$, $\sigma \in S$, h_σ, $\tilde{h}_{\sigma,\gamma}$ be defined as in* (9) *and* (12), $\gamma \in [0, 1]$, $x \in \Omega$, *and* $y = p(x; h_\sigma)$. *The following statements are equivalent:*
(a) *x is stationary for problem* (1)*;*
(b) *$x = y$;*
(c) *$\tilde{h}_{\sigma,\gamma}(y, x) = 0$.*

Proof. (a) $\Longleftrightarrow$ (b) Assume that $x = y$. Then $h_\sigma(\cdot, x)$ achieves its minimum at x and the stationarity condition applied to it yields $h'_\sigma(x, x; z - x) \geq 0$ for all $z \in \mathbb{R}^n$. Recalling (10), we have $h'_\sigma(x, x; z - x) = f'(x; z - x)$; hence x is a stationary point for problem (1).

Conversely, let $x \in \Omega$ be a stationary point of (1) and assume by contradiction that $x \neq y$. Then, by Proposition 2.2 (d), we obtain $f'(x, y-x) < 0$, which contradicts the stationarity assumption on x.

(b) $\Longleftrightarrow$ (c) See Proposition 2.2 (c). ∎

In the following sections, we will study iterative optimization algorithms based on the knowledge that the negative sign of $\tilde{h}_{\sigma,\gamma}(y, x)$ indicates a descent direction at x. At each iterate $x^{(k)}$, we will use the symbol $y^{(k)}$ to indicate the minimizer of $h_{\sigma^{(k)}}(\cdot, x^{(k)})$: $y^{(k)} = p(x^{(k)}; h_{\sigma^{(k)}})$. This minimizer may be difficult to compute. We therefore introduce the symbol $\tilde{y}^{(k)}$ to indicate an approximation of $y^{(k)}$ of which, initially, we only ask $\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) < 0$.

3. A line-search algorithm based on a modified Armijo rule.

In this section we consider the modified Armijo rule described in Algorithm LS, which is a generalization of the one in [35]. Indeed the rule proposed in [35] is recovered when d_σ is chosen as in (7) and $\gamma \in [0, 1)$. In the following we will prove that Algorithm LS is well defined and classical properties of the Armijo condition still hold for this modified case.

The modified Armijo line-search procedure represents an inner loop in the iterative optimization algorithm that will be presented in subsection 3.1. We prefer not to introduce explicitly an inner iteration counter, in order not to complicate notation.

Algorithm LS Modified Armijo line-search algorithm.

Let $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$ be two sequences of points in Ω, and let $\{\sigma^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of parameters in S. Choose some $\delta, \beta \in (0, 1)$, $\gamma \in [0, 1]$. For all $k \in \mathbb{N}$ compute $\lambda^{(k)}$ as follows:

1. Set $\lambda^{(k)} = 1$ and $d^{(k)} = \tilde{y}^{(k)} - x^{(k)}$.
2. IF

$$f(x^{(k)} + \lambda^{(k)} d^{(k)}) \leq f(x^{(k)}) + \beta\lambda^{(k)}\Delta^{(k)} \tag{14}$$

where

$$\Delta^{(k)} = \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \tag{15}$$

THEN go to step 3.

ELSE set $\lambda^{(k)} = \delta\lambda^{(k)}$ and go to step 2.

3. END

Here and in the following we define the function $h_\sigma(\cdot,\cdot)$ as in (9) and, for the sake of simplicity, make the following assumption:

(H0) $d_\sigma \in D(\Omega, S)$, where $\Omega = \mathrm{dom}(f_1)$ and $S \subseteq \mathbb{R}^q$ is a compact set.

PROPOSITION 3.1. *Let $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$ be two sequences of points in Ω, let $\{\sigma^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of parameters in $S \subseteq \mathbb{R}^q$, and let $\gamma \in [0,1]$. Assume that*

$$\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) < 0 \tag{16}$$

for all k. Then the line-search Algorithm LS is well defined; i.e., for each $k \in \mathbb{N}$ the loop at step 2 terminates in a finite number of steps. If, in addition, we assume that $\{x^{(k)}\}_{k\in\mathbb{N}}$ and $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$ are bounded sequences and $f(x^{(k+1)}) \le f(x^{(k)})$, then we have that $\Delta^{(k)} = \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)})$ is bounded. Assuming also that

$$\lim_{k\to\infty} f(x^{(k)}) - f(x^{(k)} + \lambda^{(k)} d^{(k)}) = 0, \tag{17}$$

where $\lambda^{(k)}$ and $d^{(k)}$ are computed with Algorithm LS, then we have

$$\lim_{k\to\infty} \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) = 0.$$

Proof. We prove first that the loop at step 2 of Algorithm LS terminates in a finite number of steps for any $k \in \mathbb{N}$. Assume by contradiction that there exists a $k \in \mathbb{N}$ such that Algorithm LS performs an infinite number of reductions; thus, for any $j \in \mathbb{N}$, we have

$$\begin{aligned}
\beta\Delta^{(k)} &< \frac{f(x^{(k)} + \delta^j d^{(k)}) - f(x^{(k)})}{\delta^j} \\
&= \frac{f_0(x^{(k)} + \delta^j d^{(k)}) - f_0(x^{(k)})}{\delta^j} + \frac{f_1(x^{(k)} + \delta^j d^{(k)}) - f_1(x^{(k)})}{\delta^j} \\
&\le \frac{f_0(x^{(k)} + \delta^j d^{(k)}) - f_0(x^{(k)})}{\delta^j} + \frac{\delta^j f_1(x^{(k)} + d^{(k)}) + (1-\delta^j) f_1(x^{(k)}) - f_1(x^{(k)})}{\delta^j} \\
&= \frac{f_0(x^{(k)} + \delta^j d^{(k)}) - f_0(x^{(k)})}{\delta^j} + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}),
\end{aligned}$$

where the second inequality is obtained by means of the Jensen inequality applied to the convex function f_1. Taking limits on the right-hand side for $j \to \infty$, we obtain

$$\begin{aligned}
\beta\Delta^{(k)} &\le \nabla f_0(x^{(k)})^T d^{(k)} + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) \\
&\le \nabla f_0(x^{(k)})^T d^{(k)} + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + \gamma d_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) \\
&= \Delta^{(k)} < 0,
\end{aligned}$$

where the second inequality follows from the nonnegativity of $d_\sigma \in \mathcal{D}(\Omega, S)$ and the last one from (16). Since $0 < \beta < 1$, this is an absurdum.

Assume now that $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$ are bounded sequences and that $f(x^{(k+1)}) \le f(x^{(k)})$. We show that $\Delta^{(k)} = \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)})$ is bounded. By assumption (16), $\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)})$ is bounded from above. We show that it is also bounded from below. Indeed we have

$$\begin{aligned}
\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) &= \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + \gamma d_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) \\
&\ge \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) \\
&= \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + f_1(\tilde{y}^{(k)}) - f(x^{(k)}) + f_0(x^{(k)}) \\
&\ge \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + f_1(\tilde{y}^{(k)}) - f(x^{(0)}) + f_0(x^{(k)}),
\end{aligned}$$

where the first inequality follows from the nonnegativity of d_σ, the next line is obtained by adding and subtracting $f_0(x^{(k)})$, and the last one is a consequence of $f(x^{(k+1)}) \leq f(x^{(k)})$.

Because f_1 is proper and convex, there exists a supporting hyperplane, i.e., there exist $a, b \in \mathbb{R}^n$ such that $f_1(u) \geq a^T u + b$ for all $u \in \mathbb{R}^n$. Thus,

$$\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \geq \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + a^T\tilde{y}^{(k)} + b - f(x^{(0)}) + f_0(x^{(k)}).$$

The right-hand side is a continuous function of $x^{(k)}$ and $\tilde{y}^{(k)}$. As these are assumed to lie on a closed and bounded set, the left-hand side is bounded (from below) as well.

We now show that the only limit point of $\Delta^{(k)}$ is zero. We observe that from (16) and (17) we obtain

$$0 = \lim_{k\to\infty} f(x^{(k)}) - f(x^{(k)} + \lambda^{(k)}d^{(k)}) = \beta \lim_{k\to\infty} \Delta^{(k)}\lambda^{(k)}. \tag{18}$$

Assume that there exists a subset of indices $K \subseteq \mathbb{N}$ such that $\lim_{k\in K,k\to\infty} \Delta^{(k)} = \bar{\Delta} \in \mathbb{R}$, with $\bar{\Delta} < 0$. By (18), this implies that

$$\lim_{k\in K,k\to\infty} \lambda^{(k)} = 0. \tag{19}$$

Denote by $\bar{K} \subseteq K$ a set of indices such that $\lim_{k\in\bar{K},k\to\infty} \sigma^{(k)} = \bar{\sigma}$, $\lim_{k\in\bar{K},k\to\infty} x^{(k)} = \bar{x}$, and $\lim_{k\in\bar{K},k\to\infty} \tilde{y}^{(k)} = \tilde{y}$ for some $\bar{\sigma} \in S$, $\bar{x}, \tilde{y} \in \Omega$. From (19) we have that for any sufficiently large index $k \in \bar{K}$, Algorithm LS makes at least a reduction; this means that

$$\beta(\lambda^{(k)}/\delta)\Delta^{(k)} < f(x^{(k)} + (\lambda^{(k)}/\delta)d^{(k)}) - f(x^{(k)})$$

for all sufficiently large $k \in \bar{K}$. Repeating the same arguments employed in the first part of the proof, we obtain

$$\begin{aligned}
\beta\Delta^{(k)} &< \frac{f_0(x^{(k)} + (\lambda^{(k)}/\delta)d^{(k)}) - f_0(x^{(k)})}{\lambda^{(k)}/\delta} + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) \\
&\leq \frac{f_0(x^{(k)} + (\lambda^{(k)}/\delta)d^{(k)}) - f_0(x^{(k)})}{\lambda^{(k)}/\delta} + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + \gamma d_\sigma(\tilde{y}^{(k)}, x^{(k)}).
\end{aligned}$$

Taking limits on both sides for $k \in \bar{K}, k \to \infty$, since $\{d^{(k)} = \tilde{y}^{(k)} - x^{(k)}\}_{k\in\mathbb{N}}$ is bounded, and by (19) we obtain $\beta\bar{\Delta} \leq \bar{\Delta} < 0$, which is an absurdum, being $0 < \beta < 1$. ◻

We prove also the following useful lemma.

Lemma 3.1. *Let $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$ be two sequences of points in Ω, let $\{\sigma^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of parameters in $S \subseteq \mathbb{R}^q$, and let $\gamma \in [0, 1]$. Assume that*

$$f(x^{(k+1)}) \leq f(x^{(k)} + \lambda^{(k)}d^{(k)}), \quad d^{(k)} = \tilde{y}^{(k)} - x^{(k)}, \tag{20}$$

where $\tilde{y}^{(k)}$ satisfies (16) *and $\lambda^{(k)}$ is computed by Algorithm LS for any $k \in \mathbb{N}$. Suppose that f is bounded from below. Then we have*

$$0 \leq -\sum_{k=0}^{\infty} \lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) < \infty. \tag{21}$$

Proof. Denote by $\ell \in \mathbb{R}$ a lower bound for f, i.e., $\ell \leq f(x)$ for all $x \in \mathbb{R}^n$. Inequalities (14) and (20) can be combined as

$$-\beta\lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \leq f(x^{(k)}) - f(x^{(k+1)}).$$

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

$$-\beta \sum_{k=0}^{j} \lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \leq \sum_{k=0}^{j}(f(x^{(k)})-f(x^{(k+1)})) = f(x^{(0)})-f(x^{(j+1)}) \leq f(x^{(0)})-\ell. \tag{22}$$

Thus, inequality (21) follows. ∎

The easiest way to satisfy the sufficient-decrease condition (20) is to simply set $x^{(k+1)} = x^{(k)} + \lambda^{(k)}d^{(k)}$. Our analysis applies to any general scheme satisfying (20); therefore we do not need to impose $x^{(k+1)} = x^{(k)} + \lambda^{(k)}d^{(k)}$.

3.1. A class of line-search-based algorithms. Proposition 3.1 allows the convergence analysis of a wide class of descent methods based on the Armijo condition (14). The crucial ingredients of these methods are

- a descent direction $d^{(k)} = \tilde{y}^{(k)} - x^{(k)}$, where $\tilde{y}^{(k)}$ is a suitable approximation of the point $p(x^{(k)}; h_\sigma)$;
- the sufficient decrease of the objective function between two successive iterations, which has to amount to at least $\lambda^{(k)}\tilde{h}_{\sigma,\gamma}(\tilde{y}^{(k)}, x^{(k)})$, where $\lambda^{(k)}$ is determined by the backtracking procedure given in Algorithm LS.

THEOREM 3.1. *Let $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$ be two sequences of points in Ω, and let $\{\sigma^{(k)}\}_{k\in\mathbb{N}} \subset S$ and $\gamma \in [0,1]$. Assume that there exists a limit point $\bar{x}$ of $\{x^{(k)}\}_{k\in\mathbb{N}}$, and let $K' \subseteq \mathbb{N}$ be a subset of indices such that $\lim_{k\in K',k\to\infty} x^{(k)} = \bar{x} \in \Omega$. Assume that for any $k \in \mathbb{N}$ we have*

$$f(x^{(k+1)}) \leq f(x^{(k)} + \lambda^{(k)}d^{(k)}), \quad d^{(k)} = \tilde{y}^{(k)} - x^{(k)},$$

where $\lambda^{(k)}$ is computed by Algorithm LS, $\tilde{y}^{(k)}$ satisfies (16)*, and there exists $K'' \subseteq K'$ such that*

$$\lim_{k\in K'',k\to\infty} h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) = 0, \quad \textit{with} \quad y^{(k)} = p(x^{(k)}; h_{\sigma^{(k)}}). \tag{23}$$

Then $\bar{x}$ is a stationary point for problem (1).

Proof. First, we notice that Algorithm LS is well defined, since (16) holds. We observe that, since $h_{\sigma^{(k)}}$ is strongly convex with modulus of convexity m and $y^{(k)}$ is its minimum point, we have

$$\frac{m}{2}\|z - y^{(k)}\|^2 \leq h_{\sigma^{(k)}}(z, x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \quad \forall z \in \mathbb{R}^n. \tag{24}$$

Setting $z = \tilde{y}^{(k)}$ in the previous inequality and using (23) gives

$$\lim_{k\in K'',k\to\infty} \|\tilde{y}^{(k)} - y^{(k)}\| = 0. \tag{25}$$

By continuity of the operator $p(x; h_\sigma)$, since $\{x^{(k)}\}_{k\in K'}$ is bounded, $\{y^{(k)}\}_{k\in K'}$ is bounded as well. Thus, (25) implies that $\{\tilde{y}^{(k)}\}_{k\in K''}$ is also bounded and there exists a limit point $\bar{y}$ of $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$. We define $K \subseteq K''$ such that $\lim_{k\in K,k\to\infty} \tilde{y}^{(k)} = \bar{y}$ and

$\lim_{k\in K,k\to\infty}\sigma^{(k)}=\bar\sigma$. By continuity of the operator $p(x;h_\sigma)$ with respect to all its arguments, (25) implies that $\bar y = p(\bar x; h_{\bar\sigma})$.

Consider now the sequence $\{f(x^{(k)})\}_{k\in\mathbb{N}}$. From assumption (20) it follows that

$$f(x^{(k+1)}) \le f(x^{(k)}+\lambda^{(k)}d^{(k)}) \le f(x^{(k)}). \tag{26}$$

Thus, the sequence $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ is monotone nonincreasing, and therefore it converges to some $\bar f \in \bar{\mathbb{R}}$. Since f is lower semicontinuous and $\bar x$ is a limit point of $\{x^{(k)}\}_{k\in\mathbb{N}}$, we have

$$\bar f = \lim_{k\to\infty} f(x^{(k)}) = \lim_{k\to\infty} f(x^{(k+1)}) \ge f(\bar x).$$

The previous inequality implies that $\bar f\in\mathbb{R}$, and this fact, together with inequality (26), gives

$$\lim_{k\to\infty} f(x^{(k)}) - f(x^{(k)}+\lambda^{(k)}d^{(k)}) = 0.$$

Thus we can apply Proposition 3.1 and obtain

$$\lim_{k\to\infty,k\in K} \tilde h_{\sigma^{(k)},\gamma}(\tilde y^{(k)},x^{(k)}) = 0.$$

Combining the previous equality with (13) and (23) yields

$$0 = \lim_{k\to\infty,k\in K} \tilde h_{\sigma^{(k)},\gamma}(\tilde y^{(k)},x^{(k)}) \le \lim_{k\to\infty,k\in K} h_{\sigma^{(k)}}(\tilde y^{(k)},x^{(k)}) = \lim_{k\to\infty,k\in K} h_{\sigma^{(k)}}(y^{(k)},x^{(k)}).$$

Since $h_{\sigma^{(k)}}(y^{(k)},x^{(k)})\le 0$, this implies $\lim_{k\to\infty,k\in K} h_{\sigma^{(k)}}(y^{(k)},x^{(k)})=0$. Expressing inequality (24) for $z=x^{(k)}$, we can write

$$\frac{m}{2}\|x^{(k)}-y^{(k)}\|^2 \le h_{\sigma^{(k)}}(x^{(k)},x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)},x^{(k)}) = -h_{\sigma^{(k)}}(y^{(k)},x^{(k)}) \overset{k\to\infty,k\in K}{\longrightarrow} 0.$$

Thus, we proved that $\bar y=\bar x$, and by Proposition 2.3 we have that $\bar x$ is stationary. ∎

Let us now discuss assumption (23) in Theorem 3.1, concerning the inexact solution of the minimum problem in (11). Assumption (16) guarantees that $d^{(k)} = \tilde y^{(k)} - x^{(k)}$ is a descent direction, which is needed for the line-search algorithm. However, it is not sufficient to ensure that the limit points are stationary; we also need to assume that (23) holds.

As a counterexample, consider the case $n=1$, $f_0(x)=x^2/2$, $f_1(x)=0$, $d_\sigma(x,y)=(x-y)^2/2$, $\beta=\delta=1/2$. The sequence $x^{(k+1)} = x^{(k)}+\lambda^{(k)}(\tilde y^{(k)}-x^{(k)})$ with $\lambda^{(k)}=1$, $\tilde y^{(k)} = x^{(k)}-(1/2)^{k+1}$ satisfies all the assumptions of Theorem 3.1 except (23). However, starting from $x^{(0)}=2$, the sequence writes as $x^{(k)}=1+(1/2)^k \overset{k\to\infty}{\to} 1$, while the only stationary point is 0.

Assumption (23) could be replaced by requiring that f_1 is continuous and (25) holds. Clearly, (23) cannot be checked directly, but it is very general. In the following sections, we consider two implementable conditions (the first one is limited to Euclidean distance functions; the second one is more general) that imply (23), and in sections 4.1–4.4 we show how $\tilde y^{(k)}$ can be computed in practice without knowing $p(x^{(k)};h_{\sigma^{(k)}})$.

3.2. Inexact proximal evaluation: ϵ-approximations. In this section we assume that d_σ has the form (7) and, in this case, we describe a sufficient condition for (23).

We observe that $y = p(x; h_\sigma) = \mathrm{prox}^D_{\alpha f_1}(x - \alpha D^{-1}\nabla f_0(x))$ if and only if $0 \in \partial h_\sigma(y, x)$, that is,

$$\frac{1}{\alpha} D(z - y) \in \partial f_1(y), \tag{27}$$

where $z = x - \alpha D^{-1}\nabla f_0(x)$. Borrowing the ideas in [34, 36], we consider a relaxed version of (27) and study the properties of any point $\tilde{y}$ satisfying the inclusion

$$\frac{1}{\alpha} D(z - \tilde{y}) \in \partial_\epsilon f_1(\tilde{y}), \tag{28}$$

where $\epsilon \in \mathbb{R}_{\geq 0}$ and $\partial_\epsilon f(z)$ is the ϵ-subdifferential of a convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ at a point $z \in \mathbb{R}^n$, defined as [38, p. 82]

$$\partial_\epsilon f(z) = \{w \in \mathbb{R}^n : f(x) \geq f(z) + (x - z)^T w - \epsilon \quad \forall x \in \mathbb{R}^n\}. \tag{29}$$

LEMMA 3.2. *Let d_σ be defined as in* (7) *and let $x \in \Omega$. Assume that $y = p(x; h_\sigma)$ and that $\tilde{y}$ satisfies* (28) *for some $\epsilon \in \mathbb{R}_{\geq 0}$. Then $\tilde{y} \in \Omega$ and we have*

(a) $h_\sigma(\tilde{y}, x) - h_\sigma(y, x) \leq \epsilon$;

(b) $\|\tilde{y} - y\|^2 \leq \alpha\mu\epsilon$ *for all $\mu \in \mathbb{R}_{>0}$ with $\frac{1}{\mu} \leq \lambda_{\min}(D)$, $\lambda_{\min}$ being the smallest eigenvalue of D.*

Proof. Since we have $\partial_\epsilon h_\sigma(\tilde{y}, x) \supseteq \{\frac{1}{\alpha} D(\tilde{y} - z) + w : w \in \partial_\epsilon f_1(\tilde{y})\}$ (see [38, Theorem 2.4.2 (viii)]), inclusion (28) implies $0 \in \partial_\epsilon h_\sigma(\tilde{y}, x)$, which, by definition (29) of ϵ-subdifferential, is equivalent to

$$h_\sigma(w, x) \geq h_\sigma(\tilde{y}, x) - \epsilon \quad \forall w \in \mathbb{R}^n. \tag{30}$$

Recall that $h_\sigma(\,\cdot\,, x)$ is strongly convex with modulus $m = 2/(\alpha\mu)$ and y is its minimizer. This yields

$$\frac{1}{\alpha\mu}\|\tilde{y} - y\|^2 \leq h_\sigma(\tilde{y}, x) - h_\sigma(y, x) \leq \epsilon,$$

where the rightmost inequality follows from (30) with $w = y$. ∎

The previous result combined with Theorem 3.1 directly implies the following corollary.

COROLLARY 3.1. *Let $0 < \alpha_{\min} \leq \alpha_{\max}$, $\gamma \in [0, 1]$, $\mu \geq 1$. Assume that $\{\alpha_k\}_{k\in\mathbb{N}} \subset [\alpha_{\min}, \alpha_{\max}]$, $\{D_k\}_{k\in\mathbb{N}} \subset \mathcal{M}_\mu$, $\{\epsilon_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}$, $\lim_{k\to\infty} \epsilon_k = 0$. Let $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{\tilde{y}^{(k)}\}_{k\in\mathbb{N}}$ be two sequences of points in Ω such that, for any $k \in \mathbb{N}$,* (20) *holds, where $\lambda^{(k)}$ is computed by Algorithm LS and $\tilde{y}^{(k)}$ satisfies* (16) *and*

$$\frac{1}{\alpha_k} D_k(z^{(k)} - \tilde{y}^{(k)}) \in \partial_{\epsilon_k} f_1(\tilde{y}^{(k)}), \tag{31}$$

with $z^{(k)} = x^{(k)} - \alpha_k D_k^{-1}\nabla f_0(x^{(k)})$. Then any limit point of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ is stationary for problem (1).

3.3. Inexact proximal evaluation: η-approximations. A different approach to defining a suitable approximation of the operator (11) is based on the following definition:

$$P_\eta(x; h_\sigma) = \{\tilde{y} \in \Omega : h_\sigma(\tilde{y}, x) \leq \eta h_\sigma(y, x), \text{ where } y = p(x; h_\sigma)\} \tag{32}$$

for some $\eta \in (0, 1]$. This idea of inexactness was introduced first in [7] to approximate the projection operator onto a convex set in the context of scaled gradient projection methods for smooth optimization. Clearly, if

$$\tilde{y} \in P_\eta(x; h_\sigma), \tag{33}$$

then $h_\sigma(\tilde{y}, x) \leq 0$ and $h_\sigma(\tilde{y}, x) = 0$ if and only if $h_\sigma(y, x) = 0$, which implies $\tilde{y} = y$.

The following theorem establishes a convergence result under the condition $\tilde{y}^{(k)} \in P_\eta(x^{(k)}; h_\sigma)$.

THEOREM 3.2. *Let $\eta \in (0, 1]$, $0 \leq \gamma \leq 1$, $\{\sigma^{(k)}\}_{k\in\mathbb{N}} \subset S$, and $\{x^{(k)}\}_{k\in\mathbb{N}} \subset \Omega$ satisfy* (20)*, where $\lambda^{(k)}$ is computed by Algorithm LS, with*

$$\tilde{y}^{(k)} \in P_\eta(x^{(k)}; h_{\sigma^{(k)}}). \tag{34}$$

Then either the iterate $x^{(k)}$ for some k is stationary for problem (1) *or any limit point $\bar{x}$ of $\{x^{(k)}\}_{k\in\mathbb{N}}$ is stationary for problem* (1).

Proof. Set $y^{(k)} = p(x^{(k)}; h_{\sigma^{(k)}})$ and first observe that $\gamma \leq 1$ and (34) imply

$$\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \leq h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) \leq \eta h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \leq 0. \tag{35}$$

If at some iterate $k \in \mathbb{N}$ we have $\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) = 0$ and, as a consequence, $h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) = 0$, then, by Proposition 2.3, $x^{(k)}$ is a stationary point for problem (1).

Otherwise $\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) < 0$ for all $k \in \mathbb{N}$, and thus (16) holds. Consider now a limit point $\bar{x} \in \Omega$ of $\{x^{(k)}\}_{k\in\mathbb{N}}$ (if one exists) such that $\lim_{k\to\infty, k\in K'} x^{(k)} = \bar{x}$ for some set of indices $K' \subseteq \mathbb{N}$.

We first prove that $\{\tilde{y}^{(k)}\}_{k\in K'}$ is bounded, using the strong convexity of $h_{\sigma^{(k)}}(\cdot, x^{(k)})$. From (34),

$$h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \leq (\eta - 1) h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}). \tag{36}$$

Since $h_{\sigma^{(k)}}(\cdot, x^{(k)})$ is strongly convex with modulus of convexity m, and $y^{(k)}$ is the minimizer of $h_{\sigma^{(k)}}(\cdot, x^{(k)})$,

$$\frac{m}{2}\|\tilde{y}^{(k)} - y^{(k)}\|^2 \leq h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \leq (\eta - 1) h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}).$$

Since $y^{(k)}$ depends continuously on $x^{(k)}$, when $\{x^{(k)}\}_{k\in K'}$ is bounded, and all lie in a closed set, then $\{y^{(k)}\}_{k\in K'}$ is also bounded. Recalling Proposition 3.1, we have that $\{\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)})\}_{k\in K'}$ is bounded from below; then, using inequalities (35), we can conclude that $h_{\sigma^{(k)}}(y^{(k)}, x^{(k)})$ is also bounded from below for $k \in K'$, and thus $\{\tilde{y}^{(k)}\}_{k\in K'}$ is bounded. We define $K \subseteq K'$ as the set of indices such that $\lim_{k\in K, k\to+\infty} \sigma^{(k)} = \bar{\sigma}$, $\lim_{k\in K, k\to+\infty} y^{(k)} = \bar{y}$ for some $\bar{\sigma} \in S$, $\bar{y} \in \Omega$. Thanks to the continuity of the operator (11), the set K is well defined, since the sequences $\{x^{(k)}\}_{k\in K'}$, $\{\sigma^{(k)}\}_{k\in\mathbb{N}}$ are bounded, and, moreover, we have $\bar{y} = p(\bar{x}; h_{\bar{\sigma}})$. Reasoning as in the proof of Theorem 3.1, the existence of a limit point guarantees that (17) is satisfied. Then, by Proposition 3.1, we obtain $\lim_{k\to\infty,k\in K} \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)},x^{(k)}) = 0$. Combining this with (34), we also have

$$0 = \lim_{k\to\infty,k\in K} \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)},x^{(k)}) \leq \lim_{k\to\infty,k\in K} h_{\sigma^{(k)}}(\tilde{y}^{(k)},x^{(k)}) \leq \eta \lim_{k\to\infty,k\in K} h_{\sigma^{(k)}}(y^{(k)},x^{(k)}),$$

which, since $h_{\sigma^{(k)}}(y^{(k)},x^{(k)}) \leq 0$, implies

$$\lim_{k\to\infty,k\in K} h_{\sigma^{(k)}}(y^{(k)},x^{(k)}) = 0. \tag{37}$$

Invoke again the strong convexity of $h_{\sigma^{(k)}}(\,\cdot\,,x^{(k)})$ to obtain

$$\frac{m}{2}\|x^{(k)} - y^{(k)}\|^2 \leq h_{\sigma^{(k)}}(x^{(k)},x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)},x^{(k)}) = -h_{\sigma^{(k)}}(y^{(k)},x^{(k)}),$$

which, together with (37), gives $\lim_{k\to\infty,k\in K}\|y^{(k)} - x^{(k)}\|^2 = 0$. Thus, $\bar{y} = \bar{x}$, and by Proposition 2.3, we conclude that $\bar{x}$ is stationary. ◻

3.4. Remarks. Different notions of inexactness have been proposed in the literature (see [34, 36] and references therein), especially in the context of proximal point methods, with the aim of approximating the resolvent operator, and some of them can be considered also in our framework. A synthetic description of possible inexactness notions and their relationships is given in Figure 1.

$$\begin{array}{lcl}
\frac{1}{\alpha}D(z-\tilde{y}) \in \partial_\epsilon f_1(\tilde{y}) & & \\
 & \searrow & \\
 & & 0 \in \partial_\epsilon h_\sigma(\tilde{y},x) \quad \Leftrightarrow \quad h_\sigma(\tilde{y},x) \leq h_\sigma(y,x)+\epsilon \quad \Rightarrow \quad \|\tilde{y}-y\|^2 \leq \kappa\epsilon \\
 & \nearrow & \\
\mathrm{dist}(0,\partial h_\sigma(\tilde{y},x)) \leq \epsilon & & \\
\text{(when } D = I\text{)} & &
\end{array}$$

Fig. 1. *Connection of different inexactness notions, under the assumption* (7). *The proof of the implications is given in Lemma* 3.2 *and in* [34, Proposition 1].

It is difficult to insert the inexactness criterion (33) in the scheme shown in Figure 1, since the shape of P_η in (33) depends on x, while the implications in Figure 1 are independent of x. In general, we observe that from inequality (36) and by definition of ϵ-subdifferential we have

$$0 \in \partial_{\epsilon_k} h_{\sigma^{(k)}}(\tilde{y}^{(k)},x^{(k)}), \text{ with } \ \epsilon_k = (\eta-1)h_{\sigma^{(k)}}(y^{(k)},x^{(k)}).$$

We give a pictorial example of the sets of admissible approximations $\tilde{y}$ of the exact minimizer y defined by conditions (33) and (28) in Figure 2. This example refers to the case where $f_1(x) = \iota_\Omega(x)$ is the indicator function of a convex closed set $\Omega \subseteq \mathbb{R}^n$. Choosing the Euclidean metric, i.e., (7) with $D = I$, $\alpha = 1$, as distance function, the operator $p(x;h_\sigma)$ reduces to the Euclidean projection of the point $z = x - \nabla f_0(x)$ onto Ω. Moreover, condition (28) becomes

$$\tilde{y} \in \Omega \text{ and } (w-\tilde{y})^T(z-\tilde{y}) \leq \epsilon \quad \forall w \in \Omega. \tag{38}$$

As well explained in [34, 36], from a geometrical point of view, a point $\tilde{y} \in \Omega$ satisfies (38) if and only if Ω is contained in the negative half-space determined by the hyperplane of equation $(w-\tilde{y})^T(z-\tilde{y})/\|z-\tilde{y}\| = \epsilon/\|z-\tilde{y}\|$, which is normal to $z-\tilde{y}$ at a distance $\epsilon/\|z-\tilde{y}\|$ from $\tilde{y}$.

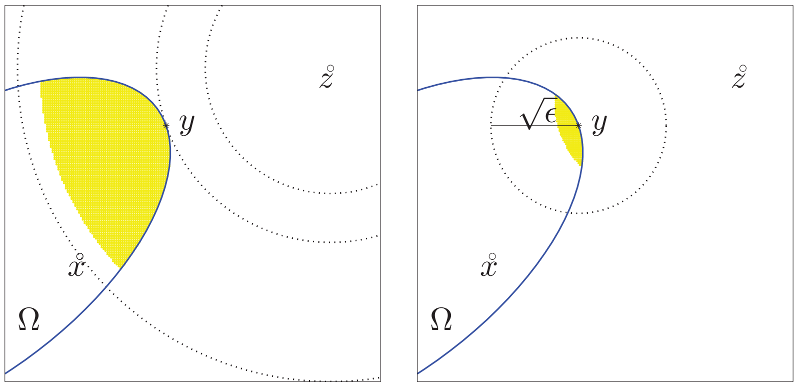

FIG. 2. *Example with* $f_1(x) = \iota_\Omega(x)$, d_σ *as in* (7) *with* $\alpha = 1$, $D = I$. *Shaded portions show the set* $P_\eta(x; h_\sigma)$ *defined in* (32) *(left panel) and the set of points* $\tilde{y}$ *satisfying* (28) *(right panel).*

On the other side, setting $\gamma = 1$ for simplicity, we have $\tilde{h}_{\sigma,\gamma}(\,\cdot\,, x) = h_\sigma(\,\cdot\,, x) = \frac{1}{2}\|\,\cdot\, - z\|^2 - \frac{1}{2}\|x - z\|^2 + \iota_\Omega(\,\cdot\,) - \iota_\Omega(x)$. Thus, the set $P_\eta(x; h_\sigma)$ is the intersection of the set Ω with the ball centered in z of radius $\sqrt{\eta\|y - z\|^2 + (1 - \eta)\|x - z\|^2}$.

In general, one of the main differences between definitions (34) and (31) consists in the fact that in the latter case the distance between the approximated and the exact minimum of $h_{\sigma^{(k)}}(\,\cdot\,, x^{(k)})$, i.e., $\|\tilde{y}^{(k)} - y^{(k)}\|$, can be controlled by the independent parameter ϵ_k, while in the other case this distance is algorithm and iteration dependent. This fact can be exploited to obtain a stronger convergence result, as shown in the next section.

3.5. Convergence analysis in the convex case with ϵ-approximations.

3.5.1. Convergence. In this section, we assume that f_0 is convex, and in this case we prove a stronger convergence result for a specific line-search algorithm where the descent direction is defined by means of an ϵ-approximation, provided that the sequence of parameters $\{\epsilon_k\}_{k\in\mathbb{N}}$ is summable and that the sequence of the matrices D_k satisfies suitable assumptions. The following theorem is a generalization of Theorem 3.1 in [10]. Further results on forward–backward variable metric algorithms which apply to problems of the form (1) when f_0 has a Lipschitz continuous gradient can be found in the recent papers [16, 19]. We stress that in all our analysis we do not need any Lipschitz continuity of the gradient of f_0 and, moreover, the sequence of errors $\|\tilde{y}^{(k)} - y^{(k)}\|$ needs to be square summable, while the convergence result stated in [16, Theorem 4.1] is given under the stronger assumption that $\|\tilde{y}^{(k)} - y^{(k)}\|$ is summable.

THEOREM 3.3. *Let* $0 < \alpha_{\min} \le \alpha_{\max}$, $\gamma \in [0, 1]$, $\{\alpha_k\}_{k\in\mathbb{N}} \subset [\alpha_{\min}, \alpha_{\max}]$. *Assume that* f_0 *in* (1) *is convex and the solution set* X^* *of problem* (1) *is not empty. Let* $\{x^{(k)}\}_{k\in\mathbb{N}}$ *be the sequence generated as*

$$x^{(k+1)} = x^{(k)} + \lambda^{(k)} d^{(k)}, \quad d^{(k)} = \tilde{y}^{(k)} - x^{(k)},$$

where $\lambda^{(k)}$ *is obtained by means of the backtracking procedure in Algorithm LS, with* $\tilde{y}^{(k)}$ *satisfying* $\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) < 0$. *Moreover assume that*

(H1) $\tilde{y}^{(k)}$ *satisfies* (31), *where the sequence* $\{\epsilon_k\}_{k\in\mathbb{N}}$ *is summable, i.e.,* $\sum_{k=0}^{\infty} \epsilon_k < \infty$;

(H2) $\{D_k\}_{k\in\mathbb{N}} \subset \mathcal{M}_\mu$, *where* $\mu \geq 1$ *and*

$$D_{k+1} \preceq (1+\zeta_k)D_k, \quad \{\zeta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}, \quad \textit{and} \quad \sum_{k=0}^{\infty} \zeta_k < \infty.$$

Then the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ *converges to a solution of* (1).

Proof. First of all we recall the basic norm equality

$$\|a-b\|_D^2 + \|b-c\|_D^2 - \|a-c\|_D^2 = 2(a-b)^T D(c-b), \tag{39}$$

which holds for any $a, b, c \in \mathbb{R}^n$. Let $\hat{x} \in X^*$. By definition of $\tilde{y}^{(k)}$ we have

$$f_1(w) \geq f_1(\tilde{y}^{(k)}) + \frac{1}{\alpha_k}(z^{(k)} - \tilde{y}^{(k)})^T D_k(w - \tilde{y}^{(k)}) - \epsilon_k \quad \forall w \in \mathbb{R}^n,$$

which, recalling that $z^{(k)} = x^{(k)} - \alpha_k D_k^{-1}\nabla f_0(x^{(k)})$, can also be written as

$$(\tilde{y}^{(k)} - x^{(k)})^T D_k(w - \tilde{y}^{(k)}) \geq \alpha_k \left(f_1(\tilde{y}^{(k)}) - f_1(w) + \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - w)\right) - \alpha_k\epsilon_k \;\; \forall w \in \mathbb{R}^n.$$

For $w = \hat{x}$, the previous inequality gives

$$\begin{aligned}
(\tilde{y}^{(k)} - x^{(k)})^T D_k(\hat{x} - x^{(k)}) &\geq \alpha_k \left(f_1(\tilde{y}^{(k)}) - f_1(\hat{x}) + \nabla f_0(x^{(k)})^T(x^{(k)} - \hat{x})\right) - \alpha_k\epsilon_k \\
&\quad + \left(\tilde{y}^{(k)} - x^{(k)} + \alpha_k D_k^{-1}\nabla f_0(x^{(k)})\right)^T D_k(\tilde{y}^{(k)} - x^{(k)}) \\
&\geq \alpha_k \left(f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + f(x^{(k)}) - f(\hat{x})\right) + \|\tilde{y}^{(k)} - x^{(k)}\|_{D_k}^2 \qquad (40)\\
&\quad + \alpha_k \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) - \alpha_k\epsilon_k \\
&\geq \|\tilde{y}^{(k)} - x^{(k)}\|_{D_k}^2 - \alpha_k\epsilon_k \\
&\quad + \alpha_k \left(f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)})\right) \\
&= \frac{1}{(\lambda^{(k)})^2}\|x^{(k+1)} - x^{(k)}\|_{D_k}^2 - \alpha_k\epsilon_k \qquad (41)\\
&\quad + \alpha_k \left(f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)})\right),
\end{aligned}$$

where the second inequality is obtained by adding and subtracting $f_1(x^{(k)})$ and by the convexity of f_0, the third one from the fact that $\hat{x}$ is a minimum point, and the last one by definition of $x^{(k+1)}$. By equality (39) with $a = x^{(k+1)}$, $b = x^{(k)}$, $c = \hat{x}$, $D = D_k$ we obtain

$$\begin{aligned}
\|x^{(k+1)} - \hat{x}\|_{D_k}^2 &= \|x^{(k)} - \hat{x}\|_{D_k}^2 + \|x^{(k+1)} - x^{(k)}\|_{D_k}^2 - 2(x^{(k)} - x^{(k+1)})^T D_k(x^{(k)} - \hat{x}) \\
&= \|x^{(k)} - \hat{x}\|_{D_k}^2 + \|x^{(k+1)} - x^{(k)}\|_{D_k}^2 - 2\lambda^{(k)}(\tilde{y}^{(k)} - x^{(k)})^T D_k(\hat{x} - x^{(k)}) \\
&\overset{(41)}{\leq} \|x^{(k)} - \hat{x}\|_{D_k}^2 + \left(1 - \frac{2}{\lambda^{(k)}}\right)\|x^{(k+1)} - x^{(k)}\|_{D_k}^2 \\
&\quad - 2\alpha_k\lambda^{(k)}\left(\nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)})\right) + 2\alpha_k\lambda^{(k)}\epsilon_k \\
&= \|x^{(k)} - \hat{x}\|_{D_k}^2 + \left(1 - \frac{2}{\lambda^{(k)}} + \frac{\gamma}{\lambda^{(k)}}\right)\|x^{(k+1)} - x^{(k)}\|_{D_k}^2 \\
&\quad - 2\alpha_k\lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) + 2\alpha_k\lambda^{(k)}\epsilon_k \\
&\leq \|x^{(k)} - \hat{x}\|_{D_k}^2 - 2\alpha_k\lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) + 2\alpha_k\lambda^{(k)}\epsilon_k, \qquad (42)
\end{aligned}$$

where the third equality is obtained by adding and subtracting the term $\gamma\lambda^{(k)}\|\tilde{y}^{(k)} - x^{(k)}\|^2_{D_k} = \gamma/\lambda^{(k)}\|x^{(k+1)} - x^{(k)}\|^2_{D_k}$ and the last inequality follows from the fact that $\gamma \in [0, 1]$. From assumption (H2) we obtain

$$\begin{aligned} \|x^{(k+1)} - \hat{x}\|^2_{D_{k+1}} &\leq (1+\zeta_k)\|x^{(k+1)} - \hat{x}\|^2_{D_k} \\ &\leq (1+\zeta_k)\|x^{(k)} - \hat{x}\|^2_{D_k} - 2\alpha_k(1+\zeta_k)\lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \\ &\quad + 2\alpha_k\lambda^{(k)}(1+\zeta_k)\epsilon_k \\ &\leq (1+\zeta_k)\|x^{(k)} - \hat{x}\|^2_{D_k} - 2\alpha_{\max}\zeta\lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) + 2\alpha_{\max}\zeta\epsilon_k, \end{aligned} \tag{43}$$

where we set $\zeta = 1 + \max_k \zeta_k$. Then from [27, Lemma 2.2.2] we can conclude that the sequence $\{\|x^{(k)} - \hat{x}\|^2_{D_k}\}_{k\in\mathbb{N}}$ converges. In particular, since $D_k \in \mathcal{M}_\mu$, $\{x^{(k)}\}_{k\in\mathbb{N}}$ is bounded and thus has at least one limit point. Let us denote such a limit point by x^∞. By Corollary 3.1, x^∞ is stationary; in particular, since f is convex, it is a minimum point, i.e., $x^\infty \in X^*$, and thus $\{\|x^{(k)} - x^\infty\|^2_{D_k}\}_{k\in\mathbb{N}}$ converges. Let $\{x^{(k_i)}\}_{i\in\mathbb{N}}$ be a subsequence of $\{x^{(k)}\}_{k\in\mathbb{N}}$ which converges to x^∞. By the norm inequality (5)

$$\|x^{(k_i)} - x^\infty\|^2_{D_{k_i}} \leq \mu\|x^{(k_i)} - x^\infty\| \overset{i\to\infty}{\longrightarrow} 0.$$

Since $\{\|x^{(k)} - x^\infty\|^2_{D_k}\}_{k\in\mathbb{N}}$ converges, this implies that its limit is zero. Invoking again (5), we obtain

$$\frac{1}{\mu}\|x^{(k)} - x^\infty\|^2 \leq \|x^{(k)} - x^\infty\|^2_{D_k} \overset{k\to\infty}{\longrightarrow} 0,$$

which allows us to conclude that $\{x^{(k)}\}_{k\in\mathbb{N}}$ converges to x^∞. ∎

In the following we present a variation of Theorem 3.3 where the tolerance parameters ϵ_k are adaptively chosen, instead of being a predefined summable sequence.

THEOREM 3.4. *Let* $0 < \alpha_{\min} \leq \alpha_{\max}$, $\gamma \in [0, 1]$, $\{\alpha_k\}_{k\in\mathbb{N}} \subset [\alpha_{\min}, \alpha_{\max}]$. *Assume that* f_0 *in* (1) *is convex and the solution set* X^* *of problem* (1) *is not empty. Let* $\{x^{(k)}\}_{k\in\mathbb{N}}$ *be the sequence generated as*

$$x^{(k+1)} = x^{(k)} + \lambda^{(k)}d^{(k)}, \quad d^{(k)} = \tilde{y}^{(k)} - x^{(k)},$$

where $\lambda^{(k)}$ *is obtained by means of the backtracking procedure in Algorithm LS, with* $\tilde{y}^{(k)}$ *satisfying* $\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) < 0$. *Moreover assume that*

(H1′) $\tilde{y}^{(k)}$ *satisfies* (31), *where the sequence* $\{\epsilon_k\}_{k\in\mathbb{N}}$ *satisfies*

$$\epsilon_k \leq -\tau\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \tag{44}$$

for some $\tau > 0$,

and that hypothesis (H2) *of Theorem* 3.3 *holds. Then the sequence* $\{x^{(k)}\}_{k\in\mathbb{N}}$ *converges to a solution of* (1).

Proof. By substituting (44) in (43) we obtain

$$\|x^{(k+1)} - \hat{x}\|^2_{D_{k+1}} \leq (1+\zeta_k)\|x^{(k)} - \hat{x}\|^2_{D_k} - 2\alpha_{\max}\zeta(1+\tau)\lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}).$$

The remainder of the proof follows exactly from the same arguments employed in Theorem 3.3. ∎

We show in section 4.4 how the conditions (31) and (44) can be satisfied in practice.

Assumption (H2) is analogous to the one proposed in [16, 19]. A special case of it is given by

$$\text{(H2}'\text{)}\ \ \{D_k\}_{k\in\mathbb{N}} \subset \mathcal{M}_{\mu_k}, \text{ where } \mu_k^2 = 1+\xi_k, \quad \xi_k \geq 0, \quad \sum_{k=0}^{\infty}\xi_k < \infty.$$

Thanks to inequality (5), for any $x \in \mathbb{R}^n$ we have

$$x^T(D_{k+1} - \mu_k\mu_{k+1}D_k)x = x^T D_{k+1}x - \mu_k\mu_{k+1}x^T D_k x \leq \mu_{k+1}\|x\|^2 - \mu_k\mu_{k+1}\frac{\|x\|^2}{\mu_k} = 0,$$

which implies $D_{k+1} \preceq \mu_k\mu_{k+1}D_k$. Moreover, $\mu_k\mu_{k+1}$ can be written as $\mu_k\mu_{k+1} = 1+\zeta_k$, where $\zeta_k = \sqrt{(1+\xi_k)(1+\xi_{k+1})} - 1$. Since $\lim_{x\to 0}\sqrt{1+x}/x = 1/2$, it follows that $\sum_{k=0}^{\infty}\xi_k$ and $\sum_{k=0}^{\infty}\zeta_k$ have the same behavior. Then we can conclude that (H2$'$) implies (H2).

We also observe that, employing the same arguments above, we can prove that $\mu_{k+1}\mu_k D_{k+1} \succeq D_k$, and as a consequence, (H2$'$) also implies that $(1+\zeta_k)D_{k+1} \succeq D_k$ with $\sum_{k=0}^{\infty}\zeta_k < \infty$. In practice, (H2$'$) says that the scaling matrices have to converge to the identity matrix at a certain rate, while (H2) implies the convergence to some symmetric positive definite matrix (see Lemma 2.3 in [19]).

3.5.2. Convergence rate analysis. In this section we analyze the convergence rate of the objective function values $f(x^{(k)})$ to the optimal value f^*, proving that $f(x^{(k+1)}) - f^* = \mathcal{O}(\frac{1}{k})$. This complexity result is obtained using the same settings as Theorem 3.4, but further assumes that the gradient of f_0 is Lipschitz continuous on the domain of f_1. This Lipschitz assumption guarantees that the sequence $\{\lambda^{(k)}\}_{k\in\mathbb{N}}$ is bounded away from zero. Before giving the main results, we need to prove the following lemma, which actually does not require the Lipschitz assumption.

LEMMA 3.3. *Let* $x^{(k)}, \tilde{y}^{(k)} \in \Omega$. *If* $\tilde{y}^{(k)}$ *satisfies* (31), *with* $0 < \alpha_k \leq \alpha_{\max}$ *and* $D_k \in \mathcal{M}_\mu$, *then*

$$\frac{1}{2\alpha_{\max}\mu}\|\tilde{y}^{(k)} - x^{(k)}\|^2 \leq -\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) + \epsilon_k. \tag{45}$$

Proof. For any $w \in \partial_{\epsilon_k} f_1(\tilde{y}^{(k)})$ we have

$$\begin{aligned}
h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) &= \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + \frac{1}{2\alpha_k}\|\tilde{y}^{(k)} - x^{(k)}\|^2_{D_k} + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) \\
&\leq \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + \frac{1}{2\alpha_k}\|\tilde{y}^{(k)} - x^{(k)}\|^2_{D_k} + w^T(\tilde{y}^{(k)} - x^{(k)}) + \epsilon_k.
\end{aligned}$$

In particular, the previous inequality holds true for $w = \frac{1}{\alpha_k}D_k(z^{(k)} - \tilde{y}^{(k)})$ (see (31)). This results in

$$\begin{aligned}
\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) &\leq h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) \\
&\leq \nabla f_0(x^{(k)})^T(\tilde{y}^{(k)} - x^{(k)}) + \frac{1}{2\alpha_k}\|\tilde{y}^{(k)} - x^{(k)}\|^2_{D_k} \\
&\quad + \frac{1}{\alpha_k}(x^{(k)} - \alpha_k D_k^{-1}\nabla f_0(x^{(k)}) - \tilde{y}^{(k)})^T D_k(\tilde{y}^{(k)} - x^{(k)}) + \epsilon_k \\
&= -\frac{1}{2\alpha_k}\|\tilde{y}^{(k)} - x^{(k)}\|^2_{D_k} + \epsilon_k \leq -\frac{1}{2\alpha_{\max}\mu}\|\tilde{y}^{(k)} - x^{(k)}\|^2 + \epsilon_k,
\end{aligned}$$

where the last inequality follows from (5). ∎

PROPOSITION 3.2. *Let $\{x^{(k)}\}_{k\in\mathbb{N}}$ be a sequence of points in Ω and $\{d^{(k)}\}_{k\in\mathbb{N}}$ a sequence of descent directions such that $d^{(k)} = \tilde{y}^{(k)} - x^{(k)}$ and (45) holds. Let $\{\lambda^{(k)}\}_{k\in\mathbb{N}}$ be the steplength sequence computed by Algorithm LS and assume that ∇f_0 is Lipschitz continuous on Ω and that (44) holds. Then there exists $\lambda_{\min} \in \mathbb{R}_{>0}$ such that*

$$\lambda^{(k)} \geq \lambda_{\min} \quad \forall k \in \mathbb{N}. \tag{46}$$

Proof. In view of (44)–(45), setting $a = \alpha_{\max}\mu$, one obtains

$$\|d^{(k)}\|^2 \leq -2a(1+\tau)\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}). \tag{47}$$

If ∇f_0 is Lipschitz continuous on Ω with Lipschitz constant L, then from the descent lemma [6, p. 667] we have

$$f_0(x^{(k)} + \lambda d^{(k)}) \leq f_0(x^{(k)}) + \lambda \nabla f_0(x^{(k)})^T d^{(k)} + \frac{L}{2}\lambda^2 \|d^{(k)}\|^2, \tag{48}$$

where $\lambda \in [0, 1]$. By combining inequalities (47) and (48) we further obtain

$$f_0(x^{(k)} + \lambda d^{(k)}) \leq f_0(x^{(k)}) + \lambda \nabla f_0(x^{(k)})^T d^{(k)} - a(1+\tau)L\lambda^2 \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}).$$

Summing $f_1(x^{(k)} + \lambda d^{(k)})$ on both sides of the previous relation and applying the Jensen inequality $f_1(x^{(k)} + \lambda d^{(k)}) \leq (1-\lambda) f_1(x^{(k)}) + \lambda f_1(\tilde{y}^{(k)})$ to the right-hand side yields

$$\begin{aligned}
f(x^{(k)} + \lambda d^{(k)}) &\leq f(x^{(k)}) - \lambda f_1(x^{(k)}) + \lambda f_1(\tilde{y}^{(k)}) + \lambda \nabla f_0(x^{(k)})^T d^{(k)} \\
&\quad - aL\lambda^2(1+\tau)\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \\
&\leq f(x^{(k)}) - \lambda f_1(x^{(k)}) + \lambda f_1(\tilde{y}^{(k)}) + \lambda \nabla f_0(x^{(k)})^T d^{(k)} \\
&\quad - aL\lambda^2(1+\tau)\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) + \frac{\lambda\gamma}{2}\|d^{(k)}\|^2_{D_k} \\
&= f(x^{(k)}) + \lambda \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) - aL\lambda^2(1+\tau)\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) \\
&= f(x^{(k)}) + \lambda\left(1 - aL(1+\tau)\lambda\right)\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}).
\end{aligned}$$

The previous inequality ensures that the Armijo condition

$$f(x^{(k)} + \lambda d^{(k)}) \leq f(x^{(k)}) + \lambda\beta \tilde{h}_{\sigma,\gamma}(\tilde{y}^{(k)}, x^{(k)}) \tag{49}$$

is satisfied, for all $k \in \mathbb{N}$, when $1 - aL(1+\tau)\lambda \geq \beta$, that is, for all λ such that $\lambda \leq (1-\beta)/(aL(1+\tau))$. If $\lambda^{(k)}$ is the steplength computed by Algorithm LS and the backtracking loop is performed at least once, then $\lambda = \lambda^{(k)}/\delta$ does not satisfy inequality (49), which means $\lambda^{(k)} > (1-\beta)\delta/(aL(1+\tau))$. Thus, the steplength sequence $\{\lambda^{(k)}\}_{k\in\mathbb{N}}$ satisfies inequality (46) with $\lambda_{\min} = (1-\beta)\delta/(aL(1+\tau))$. ∎

Based on these premises, we are now ready to prove the convergence rate result. The proof of the theorem follows the arguments developed in Theorem 3.1 of [5] by modifying them to the case of inexact proximal computation, variables metrics, and the presence of the Armijo line search. In [16] where the authors also take into account variable metrics and inexact computation, no convergence rate results are given. In the case of exact proximal computation and without variable metrics, a different proof of the convergence rate is given in [12]. In contrast to the results mentioned, the explicit expression of the Lipschitz constant of ∇f_0 is not needed in order to choose the line-search parameter in the correct range (the Armijo rule guarantees sufficient decrease and the steplength parameters are free).

THEOREM 3.5. *Assume that the hypotheses of Theorem* 3.4 *hold and, in addition, that the gradient of* f_0 *is Lipschitz continuous on* Ω. *Let* f^* *be the optimal function value for problem* (1). *Then*

$$f(x^{(k+1)}) - f^* = \mathcal{O}\left(\frac{1}{k}\right).$$

Proof. If we do not neglect the term $f(x^{(k)}) - f(\hat{x}) = f(x^{(k)}) - f^*$ in (40) and in all the subsequent inequalities, instead of (42) we obtain

$$\begin{aligned}\|x^{(k+1)} - \hat{x}\|_{D_k}^2 \le \|x^{(k)} - \hat{x}\|_{D_k}^2 &+ 2\alpha_k\lambda^{(k)}\left(-\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) + \epsilon_k\right)\\ &- 2\lambda^{(k)}\alpha_k(f(x^{(k)}) - f^*),\end{aligned}$$

and hence

$$\begin{aligned}\|x^{(k+1)} - \hat{x}\|_{D_{k+1}}^2 &\le (1+\zeta_k)\|x^{(k+1)} - \hat{x}\|_{D_k}^2\\ &\le (1+\zeta_k)\|x^{(k)} - \hat{x}\|_{D_k}^2 + 2\alpha_k\lambda^{(k)}(1+\zeta_k)(-\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}) + \epsilon_k)\\ &\quad - 2\lambda^{(k)}(1+\zeta_k)\alpha_k(f(x^{(k)}) - f^*)\\ &\overset{(44)}{\le} (1+\zeta_k)\|x^{(k)} - \hat{x}\|_{D_k}^2 - 2\alpha_{\max}(1+\tau)\zeta\lambda^{(k)}\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)})\\ &\quad + a(f^* - f(x^{(k)})),\end{aligned}$$

where we set $\zeta = 1 + \max_k \zeta_k$, $a = 2\lambda_{\min}\alpha_{\min}$, where $\lambda_{\min}$ is defined in Proposition 3.2. Summing the previous inequality from 0 to k gives

$$\begin{aligned}&\|x^{(k+1)} - \hat{x}\|_{D_{k+1}}^2 \le \|x^{(0)} - \hat{x}\|_{D_0}^2 + \sum_{j=0}^{k}\zeta_j\|x^{(j)} - \hat{x}\|_{D_j}^2\\ &- 2\alpha_{\max}(1+\tau)\zeta\sum_{j=0}^{k}\lambda^{(j)}\tilde{h}_{\sigma^{(j)},\gamma}(\tilde{y}^{(j)}, x^{(j)}) + a\left((k+1)f^* - \sum_{j=0}^{k}f(x^{(j)})\right)\\ &\le \|x^{(0)} - \hat{x}\|_{D_0}^2 + M\bar{\zeta} - \frac{2\alpha_{\max}(1+\tau)\zeta}{\beta}(f(x^{(0)}) - f^*) + a\left((k+1)f^* - \sum_{j=0}^{k}f(x^{(j)})\right),\end{aligned}$$

where the second inequality follows

- by setting $\bar{\zeta} = \sum_{j=0}^{\infty}\zeta_j$;
- from the fact that $\{\|x^{(k)} - \hat{x}\|_{D_k}^2\}_{k\in\mathbb{N}}$ is a convergent sequence (see Theorem 3.4), and thus there exists M such that $\|x^{(j)} - \hat{x}\|_{D_j}^2 \le M$; and
- from (22).

Adding the positive quantity $a(f(x^{(0)}) - f^*)$ to the right-hand side of the last inequality, we obtain

$$\begin{aligned}\|x^{(k+1)} - \hat{x}\|_{D_{k+1}}^2 \le \|x^{(0)} - \hat{x}\|_{D_0}^2 + M\bar{\zeta} &- \frac{2\alpha_{\max}(1+\tau)\zeta}{\beta}(f(x^{(0)}) - f^*)\\ &+ a\left(kf^* - \sum_{j=1}^{k}f(x^{(j)})\right).\end{aligned}$$

Moreover, exploiting the inequality

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

gives

$$\|x^{(k+1)} - \hat{x}\|^2_{D_{k+1}} \le \|x^{(0)} - \hat{x}\|^2_{D_0} + M\bar{\zeta} - \frac{2\alpha_{\max}(1+\tau)\zeta}{\beta}(f(x^{(0)}) - f^*) + ak(f^* - f(x^{(k+1)})).$$

Rearranging terms yields

$$f(x^{(k+1)}) - f(\hat{x}) \le \frac{1}{ak}\left(\|x^{(0)} - \hat{x}\|^2_{D_0} + M\bar{\zeta} - 2\frac{\alpha_{\max}(1+\tau)\zeta}{\beta}(f(x^{(0)}) - f(\hat{x}))\right),$$

establishing the result. ∎

4. Practical computation of η- and ϵ-approximations.

4.1. Computation of the inexact proximal: η-approximations. In this section we discuss how to compute a point $\tilde{y}^{(k)}$ such that (34) holds, i.e., satisfying

$$h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) \le \eta h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}), \quad \text{with } y^{(k)} = p(x^{(k)}; h_\sigma), \tag{50}$$

for a given $\eta \in (0, 1]$, without knowing $y^{(k)}$. A special case of this problem, corresponding to the case $f_1 = \iota_\Omega$, where Ω is the intersection of closed, convex sets and the metric is given by (7), is considered in [7].

Satisfying inequality (50) is possible when, for each k, one can compute a sequence $\{a_l\}_{l\in\mathbb{N}} \subset \mathbb{R}$ such that

$$a_l \le h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \ \ \forall l \in \mathbb{N} \quad \text{and} \ \lim_{l\to\infty} a_l = h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}), \tag{51}$$

and a sequence of points $\{\tilde{y}^{(k,l)}\}_{l\in\mathbb{N}}$ such that

$$\lim_{l\to\infty} h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)}) = h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}). \tag{52}$$

In practice, l should be considered as the index of an inner loop for computing $\tilde{y}^{(k)}$. Indeed, when (51) holds, we also have

$$\eta a_l \le \eta h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \ \ \forall l \in \mathbb{N}. \tag{53}$$

Moreover, for all sufficiently large l we have $a_l > h_{\sigma^{(k)}}(y^{(k)}, x^{(k)})/\eta$, which, together with (53), gives

$$h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) < \eta a_l \le \eta h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}).$$

Then, if one considers any method generating a sequence $\tilde{y}^{(k,l)}$ such that (52) holds, the stopping criterion

$$h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)}) \le \eta a_l \tag{54}$$

for the inner iterations is well defined. If l is the smallest integer such that (54) is satisfied, then the point $\tilde{y}^{(k)} = \tilde{y}^{(k,l)}$ satisfies (50). In the following sections we show how to compute a sequence a_l satisfying (51) in the interesting case of the Euclidean metric.

4.2. Composition with a linear operator. In this section we assume that $f_1(x)$ is given by

$$f_1(x) = g(Ax), \tag{55}$$

where $A \in \mathbb{R}^{m\times n}$ and $g : \mathbb{R}^m \to \bar{\mathbb{R}}$ is a convex function. Moreover, we choose d_σ as in (7). We consider the minimum problem (11), which can be written in equivalent primal-dual and dual form as

$$\min_{y\in\mathbb{R}^n} h_{\sigma^{(k)}}(y, x^{(k)}) = \min_{y\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} F_{\sigma^{(k)}}(y, v, x^{(k)}) = \max_{v\in\mathbb{R}^m} \Psi_{\sigma^{(k)}}(v, x^{(k)}).$$

The primal-dual problem can be obtained from the primal one by applying the definition of the convex conjugate g^*, which gives $g(Ax) = \max_{v\in\mathbb{R}^m} v^T Ax - g^*(v)$, obtaining

$$F_{\sigma^{(k)}}(y, v, x^{(k)}) = \frac{1}{2\alpha_k}\|y - z^{(k)}\|^2_{D_k} + y^T A^T v - g^*(v) - f_1(x^{(k)}) - \frac{\alpha_k}{2}\|\nabla f_0(x^{(k)})\|^2_{D_k^{-1}} \tag{56}$$

with $z^{(k)} = x^{(k)} - \alpha_k D_k^{-1}\nabla f_0(x^{(k)})$. The dual problem is obtained by computing the minimum of the primal-dual function with respect to y, which is given by $y = z^{(k)} - \alpha_k D_k^{-1} A^T v$, and substituting it in (56), obtaining the explicit expression of the dual function

$$\begin{aligned}\Psi_{\sigma^{(k)}}(v, x^{(k)}) = &-\frac{1}{2\alpha_k}\|\alpha_k D_k^{-1} A^T v - z^{(k)}\|^2_{D_k} - g^*(v) - f_1(x^{(k)}) - \frac{\alpha_k}{2}\|\nabla f_0(x^{(k)})\|^2_{D_k^{-1}}\\ &+ \frac{1}{2\alpha_k}\|z^{(k)}\|^2_{D_k}.\end{aligned}$$

By definition of the primal-dual and dual functions, the following inequalities hold:

$$h_{\sigma^{(k)}}(y, x^{(k)}) \geq F_{\sigma^{(k)}}(y, v, x^{(k)}) \geq \Psi_{\sigma^{(k)}}(v, x^{(k)}) \quad \forall y \in \mathbb{R}^n, v \in \mathbb{R}^m.$$

In particular, the previous inequality holds for $y = y^{(k)}$. Then an approximation $\tilde{y}^{(k)}$ of $y^{(k)}$ can be computed by applying any method to the dual problem

$$\max_{v\in\mathbb{R}^m} \Psi_{\sigma^{(k)}}(v, x^{(k)}), \tag{57}$$

generating a sequence $\{v^{(l)}\}_{l\in\mathbb{N}}$ such that $\Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)})$ converges to the maximum of the dual function $\Psi_{\sigma^{(k)}}(\,\cdot\,, x^{(k)})$. As a consequence of this, setting $\tilde{y}^{(k,l)} = z^{(k)} - \alpha_k D_k^{-1} A^T v^{(l)}$, a point satisfying (50) can be found by stopping the dual iterations when

$$h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)}) \leq \eta \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)}) \tag{58}$$

is satisfied, i.e., (54) with $a_l = \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)})$.

For example, one can apply a forward–backward method [18], such as ISTA or its accelerated version (FISTA [5]), to the dual problem. As an alternative, the saddle point problem

$$\min_{y\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} F_{\sigma^{(k)}}(y, v, x^{(k)})$$

can be addressed, for example, with a primal-dual method such as [14, 26], using (58) as stopping condition. More generally, a point $\tilde{y}^{(k)} \in P_\eta(x^{(k)}; h_{\sigma^{(k)}})$ can be obtained by computing two sequences, $\{v^{(l)}\}_{l\in\mathbb{N}}$, $\{\tilde{y}^{(k,l)}\}_{l\in\mathbb{N}}$, such that

$$\lim_{l\to\infty} \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)}) = \max_{v\in\mathbb{R}^m} \Psi_{\sigma^{(k)}}(v, x^{(k)}) = \min_{y\in\mathbb{R}^n} h_{\sigma^{(k)}}(y, x^{(k)}) = \lim_{l\to\infty} h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)}),$$

stopping the iterates when (58) is met.

Remarks. We observe that (55) includes also the case where $f_1(x)$ is defined as $f_1(x) = \sum_{i=1}^r g_i(A_i x)$, where $A_i \in \mathbb{R}^{m_i\times n}$, $g_i : \mathbb{R}^{m_i} \to \mathbb{R}$. Indeed, formulation (55) is recovered by setting $A = [A_1^T\ A_2^T\ \dots\ A_r^T]^T \in \mathbb{R}^{m\times n}$ with $m = \sum_{i=1}^r m_i$. In this case the dual variable v can be partitioned as $v = [v_1^T\ v_2^T\ \dots\ v_r^T]^T$, where $v_i \in \mathbb{R}^{m_i}$ and $g^*(v) = \sum_{i=1}^r g_i^*(v_i)$ (see [38, Theorem 2.3.1 (iv)]).

4.3. Preserving feasibility. Clearly, any point $\tilde{y}^{(k,l)}$ satisfying (58), where $v^{(l)}$ is generated by any converging algorithm applied to the dual or the primal-dual problem, belongs to the domain of $h_\sigma(\cdot, x^{(k)})$, i.e., to the set Ω. Indeed, for any l, $v^{(l)}$ belongs to the domain of the dual function $\Psi_{\sigma^{(k)}}(\cdot, x^{(k)})$ and, as a consequence, (54) implies that $h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)})$ is finite. However, the stopping criterion (54) may require a very large number of inner iterations l to be satisfied, and, in addition, the primal sequence points $\tilde{y}^{(k,l)}$ may be feasible only in the limit. For these reasons, we propose also considering the sequence $\bar{y}^{(k,l)} = P_\Omega(\tilde{y}^{(k,l)})$, where P_Ω denotes the Euclidean projection onto the set Ω. If, at some inner iteration l, the inequality

$$\tilde{h}_{\sigma^{(k)},\bar{\gamma}}(\bar{y}^{(k,l)}, x^{(k)}) \leq \eta \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)}) \tag{59}$$

is satisfied, this clearly means that $\bar{y}^{(k,l)} \in P_\eta(x^{(k)}; h_\sigma)$ (i.e., (34) is satisfied), and we can set $\tilde{y}^{(k)} = \bar{y}^{(k,l)}$. We observe that when $\tilde{y}^{(k,l)}$ converges to $y^{(k)}$ as l diverges, the stopping criterion (59) is well defined, since $\bar{y}^{(k,l)}$ also converges to $y^{(k)}$.

4.4. Computation of the inexact proximal: ϵ-approximations. In this section we show how to compute a point satisfying inclusion (31), for any given $\epsilon_k \in \mathbb{R}_{\geq 0}$, when the convex function f_1 in (1) has the form (55). Our arguments are obtained by extending those in [36], which are recovered by setting $D_k = I$. As done in section 4.2, we will make use of the duality theory. In particular, we define the primal-dual gap function as

$$\mathcal{G}_{\sigma^{(k)}}(y, v, x^{(k)}) = h_{\sigma^{(k)}}(y, x^{(k)}) - \Psi_{\sigma^{(k)}}(v, x^{(k)}). \tag{60}$$

We also have the following results.

LEMMA 4.1. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, $g : \mathbb{R}^m \to \bar{\mathbb{R}}$ be two convex functions, $A \in \mathbb{R}^{m\times n}$. If $f(x) = g(Ax)$, then $f^*(A^T y) \leq g^*(y)$ for all $y \in \mathbb{R}^m$.*

Proof. By the definition of the convex conjugate we have

$$\begin{aligned} f^*(A^T y) &= \sup_{x\in\mathbb{R}^n} x^T A^T y - f(x) = \sup_{x\in\mathbb{R}^n} (Ax)^T y - g(Ax) \\ &= \sup_{z\in\mathbb{R}^m, z=Ax} z^T y - g(z) \leq \sup_{z\in\mathbb{R}^m} z^T y - g(z) = g^*(y). \qquad \square \end{aligned}$$

PROPOSITION 4.1. *Let $\epsilon_k \in \mathbb{R}_{\geq 0}$. If*

$$\mathcal{G}_{\sigma^{(k)}}(\tilde{y}^{(k)}, v, x^{(k)}) \leq \epsilon_k, \tag{61}$$

with $\tilde{y}^{(k)} = z^{(k)} - \alpha_k D_k^{-1} A^T v$, for some $v \in \mathbb{R}^m$, then (31) is satisfied.

Proof. From the definition of the primal-dual gap, a simple computation shows that

$$\begin{aligned}\mathcal{G}_{\sigma^{(k)}}(\tilde{y}^{(k)}, v, x^{(k)}) &= \frac{1}{\alpha_k}\|\alpha_k D_k^{-1}A^T v\|_{D_k}^2 - v^T A z^{(k)} + f_1(z^{(k)} - \alpha_k D_k^{-1}A^T v) + g^*(v)\\ &= \sup_{w\in\mathbb{R}^m} \frac{1}{\alpha_k}\|\alpha_k D_k^{-1}A^T v\|_{D_k}^2 - v^T A z^{(k)} + w^T(z^{(k)} - \alpha_k D_k^{-1}A^T v)\\ &\quad - f_1^*(w) + g^*(v)\\ &= \sup_{w\in\mathbb{R}^m} (w - A^T v)^T(z^{(k)} - \alpha_k D_k^{-1}A^T v) - f_1^*(w) + g^*(v)\\ &\geq \sup_{w\in\mathbb{R}^m} (w - A^T v)^T(z^{(k)} - \alpha_k D_k^{-1}A^T v) - f_1^*(w) + f_1^*(A^T v),\end{aligned}$$

where the last inequality follows from Lemma 4.1. Thus, if (61) holds, the previous inequality yields

$$(w-A^T v)^T(z^{(k)}-\alpha_k D_k^{-1}A^T v)-f_1^*(w)+f_1^*(A^T v) \leq \mathcal{G}_{\sigma^{(k)}}(\tilde{y}^{(k)}, v, x^{(k)}) \leq \epsilon_k \quad \forall w\in\mathbb{R}^m.$$

Rearranging terms, the previous inequality writes also as

$$f_1^*(w) \geq f_1^*(A^T v) + (w - A^T v)^T(z^{(k)} - \alpha_k D_k^{-1}A^T v) - \epsilon_k \quad \forall w\in\mathbb{R}^m,$$

which, from definition (29), is equivalent to $z^{(k)} - \alpha_k D_k^{-1}A^T v \in \partial_{\epsilon_k} f_1^*(A^T v)$. Finally, by applying [38, Theorem 2.4.4 (iv)], we obtain $A^T v \in \partial_{\epsilon_k} f_1(z^{(k)} - \alpha_k D_k^{-1}A^T v)$. Recalling that $\tilde{y}^{(k)} = z^{(k)} - \alpha_k D_k^{-1}A^T v$, which implies $A^T v = D_k(z^{(k)} - \tilde{y}^{(k)})/\alpha_k$, (31) follows. ∎

Proposition 4.1 suggests that for computing $\tilde{y}^{(k)}$ satisfying the assumptions of Corollary 3.1, we can use the same iterative approaches described at the end of section 4.1, stopping the iterates when

$$\mathcal{G}_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, v^{(l)}, x^{(k)}) \leq \epsilon_k \quad \text{and } \tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k,l)}, x^{(k)}) < 0. \tag{62}$$

4.5. Equivalence between η- and ϵ-approximations. Any η-approximation $\tilde{y}^{(k)}$ satisfying (58) for some $v^{(l)} \in \mathbb{R}^m$ is also an ϵ-approximation, where $\epsilon = -\tau h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)})$ and $\tau = -1 + 1/\eta$. In fact, in these settings, (58) implies $h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) - \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)}) \leq -\tau h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)})$, and, as shown in section 4.4, this means that $\tilde{y}^{(k)}$ is an ϵ-approximation with $\epsilon = -\tau h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)})$. Thus, any point computed by an iterative procedure stopped when (58) is satisfied is both an η- and an ϵ-approximation.

5. Numerical illustration. In order to validate the proposed approach, we consider two relevant image restoration problems, whose variational formulations consist in minimizing the sum of a discrepancy functional plus a regularization term. The first test problem is convex; the second test problem is nonconvex. An outline of the algorithm that we will use is shown in Algorithm VMILA, and it is the same for both convex and nonconvex problems.

5.1. Image deconvolution in presence of Poisson noise. Following the Bayesian paradigm, when the noise affecting the data is of Poisson type, a typical choice for measuring the discrepancy of a given image x from the observed data b is the following Kullback–Leibler divergence:

$$\mathrm{KL}(x, b) = \sum_{i=1}^n b_i \log\left(\frac{b_i}{x_i}\right) + x_i - b_i.$$

Taking into account also the distortion due to the image acquisition system, which we assume to be modeled through a linear operator $H \in \mathbb{R}^{n\times n}$, and a constant background term g, the data discrepancy is defined as

$$f_0(x) = \mathrm{KL}(Hx + g\mathbb{1}, b),$$

where $\mathbb{1} \in \mathbb{R}^n$ is the vector of all ones. Moreover, when one wants to preserve edges in the restored image and also the nonnegativity of the pixel values, the regularization term can be chosen as

$$f_1(x) = \rho \sum_{i=1}^{n} \|\nabla_i x\| + \iota_{\mathbb{R}^n_{\geq 0}}(x), \tag{63}$$

where $\rho > 0$ is a regularization parameter multiplying the total variation functional [33] and $\nabla_i \in \mathbb{R}^{2\times n}$ represents the discrete-gradient operator at the pixel i. Clearly, the function $f_1(x)$ has the form (55), with $A = \begin{pmatrix}\nabla_1^T & \cdots & \nabla_n^T & I\end{pmatrix}^T \in \mathbb{R}^{3n\times n}$. In this case $v \in \mathbb{R}^{3n}$ and g^* is the indicator function of the set $B^2_{0,\rho} \times \cdots \times B^2_{0,\rho} \times \mathbb{R}^n_{\leq 0}$, where $B^2_{0,\rho} \subset \mathbb{R}^2$ is the 2-dimensional Euclidean ball centered in 0 with radius ρ.

In our experiments we assume that H corresponds to a convolution operator associated to a Gaussian kernel, with reflective boundary conditions, so that the matrix-vector products involving H can be performed via the discrete cosine transform [24].

We define a set of test problems in the following way: A reference image is rescaled so that the pixel values lie in a specified range (this is for simulating different noise levels); then it is blurred by convolution with a Gaussian kernel with standard deviation σ_{psf} and the background is added. Finally, Poisson noise is simulated with the MATLAB `imnoise` function, obtaining the noisy blurred image b. The details of each test problem are listed in Table 1. The regularization parameter ρ has been manually tuned to obtain a visually satisfactory solution. For each test problem we numerically compute the optimal value f^* by running the considered algorithms for a huge number of iterations, retaining the smallest value found.

Algorithm VMILA Variable metric inexact line-search algorithm (VMILA).

Choose $0 < \alpha_{\min} \leq \alpha_{\max}$, $\mu \geq 1$, $\delta, \beta \in (0,1)$, $\gamma \in [0,1]$, $\eta \in (0,1]$, $x^{(0)} \in \Omega$.
For $k = 0, 1, 2, \ldots$

1. Choose $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$, $1 \leq \mu_k \leq \mu$, and $D_k \in \mathcal{M}_{\mu_k}$;
2. Compute $\tilde{y}^{(k)}$: compute a dual vector $v^{(l)} \in \mathbb{R}^m$ and the corresponding primal vector $\tilde{y}^{(k,l)}$ such that (58) is satisfied, then set $\tilde{y}^{(k)} = \tilde{y}^{(k,l)}$.
3. Set $d^{(k)} = \tilde{y}^{(k)} - x^{(k)}$;
4. Compute the steplength parameter $\lambda^{(k)}$ with Algorithm LS;
5. Set $x^{(k+1)} = x^{(k)} + \lambda^{(k)} d^{(k)}$.

We implement our inexact algorithm, which is summarized in Algorithm VMILA, in a MATLAB environment with the following settings:

Step 1*: Metric selection.* The scaling matrix D_k is chosen mimicking the split-gradient idea [25] that was developed for smooth optimization problems under a nonnegativity constraint. In particular, the gradient of the smooth part of the objective function is split into a positive and a negative part, which gives rise to a class of iterative optimization algorithms that can be interpreted as scaled gradient methods. As the gradient of the KL function can easily be split in the same way, we propose

TABLE 1
Test problems description.

Problem	Ref. image	Size	Range	σ_{psf}	g	ρ
`cameraman`	MATLAB cameraman	256^2	$[0, 1000]$	1.4	5	0.0091
`micro`	[37, Figure 8]	128^2	[1,69]	3.2	0.5	0.09
`phantom`	Shepp–Logan phantom	256^2	$[0, 1000]$	1.4	10	0.004

adopting this procedure in this case, too. Skipping the technical details, the scaling matrix is defined as the diagonal matrix with positive entries:

$$[D_k]_{ii} = \max\left(\min\left(\frac{x_i^{(k)}}{[H^T\mathbb{1}]_i}, \mu_k\right), \frac{1}{\mu_k}\right)^{-1},$$

where $\mu_k = \sqrt{1 + 10^{10}/k^2}$, so that assumption (H2′) is satisfied. We choose a large initial range for the scaling matrix selection to allow more freedom of choice at the first iterates, where the benefits of the scaling matrix are more relevant [9].

Step 1*: Steplength selection.* The parameter α_k is chosen by the same strategy used, e.g., in [11, 30, 29] (based on a suitable alternation of the Barzilai–Borwein rules), and its value is constrained in the interval $[\alpha_{\min}, \alpha_{\max}]$ with $\alpha_{\min} = 10^{-5}$, $\alpha_{\max} = 10^2$.

Step 2*: Computation of the approximated proximal point* $\tilde{y}^{(k)}$*.* We experimented with different inner solvers applied on the primal-dual or on the dual formulation of the inner problem. The best performances have been obtained using FISTA applied to the dual problem (57), in the variant proposed in [13], which ensures the convergence not only of the objective function values to the optimal one but also of the iterates to the minimum point. In particular, we set $t_l = (l + a - 1)/2$, with $a = 2.1$ [13, formula (5)]. For brevity, we report only the results obtained stopping the inner iterates when criterion (59) is met, which corresponds to both an η- and an ϵ-approximation (see section 4.5). A maximum number of 1500 inner iterations is imposed. The initial guess of the inner loop at the first outer iterate is the vector of all zeros, while at all successive iterates the inner solver is initialized with the dual solution computed at the previous iterate.

Other parameter setting. The line-search parameters δ, β, γ have been set, respectively, equal to $0.5, 10^{-4}, 1$. These are the standard choices for the Armijo parameters in the constrained optimization setting [11, 29, 30], where it has been remarked that the performance of the line-search algorithm is not sensitive to modification of the standard values.

All the following results have been obtained on a PC equipped by an Intel Core i7-2620M processor with CPU at 2.70GHz and 8GB of RAM, running Windows 7 OS and MATLAB R2010b. The MATLAB routines and the datasets are available at the website http://www.oasis.unimore.it/site/home/software.html.

We investigate first the impact of the inexactness parameter η choice on the overall method. In Figure 3 the relative decrease of the objective function values in the first 500 iterates is reported with respect to both the iteration number (first row) and the computational time, in seconds (second row). It can be observed that a higher precision can accelerate the progress toward the solution, but this usually results in a very large number of inner iterations, and, consequently, it is extremely time consuming (for example, for the test problem `cameraman` with $\eta = 10^{-6}, 10^{-2}, 5 \cdot 10^{-1}$ the mean number of inner iterations per outer iteration is 28, 54, 409, respectively). This

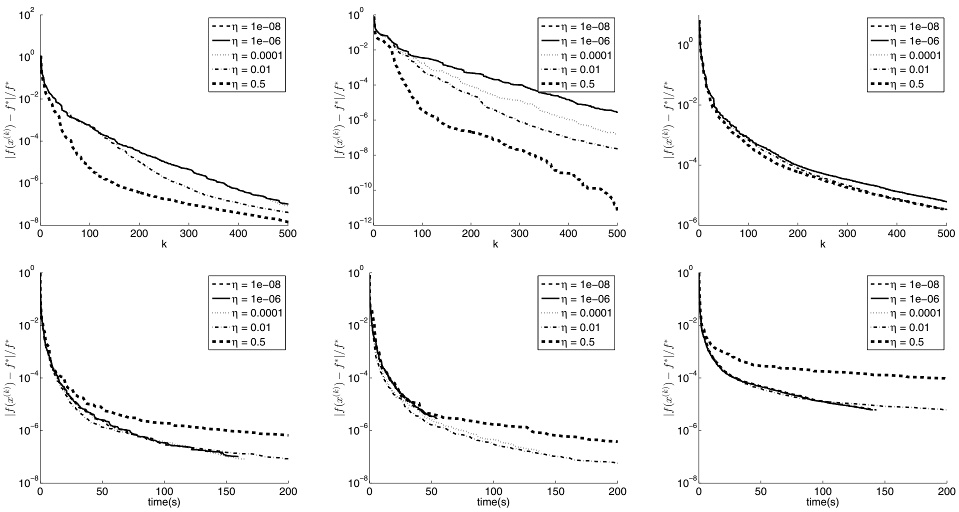

FIG. 3. *Algorithm VMILA with different choices for η. Relative decrease of the objective function values with respect to the outer iteration number (top row) and to the computational time (bottom row). Left column:* `cameraman`*. Middle column:* `micro`*. Right column:* `phantom`*.*

is typical of inexact algorithms based on the iterative solution of an inner subproblem. We find that a good balance between convergence speed and computational cost is obtained by allowing a relatively large tolerance, corresponding to $\eta = 10^{-6}$.

As a further benchmark, we compare our algorithm to Chambolle and Pock's method (CP) [14], which, referring to the notation used in their paper, has been implemented setting $G(x) = \iota_{\mathbb{R}^n_{\geq 0}(x)}$ and $F(Kx) = \mathrm{KL}(Hx + g, b) + \beta \sum_{i=1}^n \|\nabla_i x\|$, with $K = \left(H^T, \nabla_1^T, \ldots, \nabla_n^T\right)^T$. In this way the resolvent operator associated to F^* can be computed in closed form. In Figure 4, we compare the behavior of our approach (with $\eta = 10^{-6}$) with CP (2000 iterations) for different choices of its two parameters, σ and τ (once τ is selected, σ is chosen such that $\tau\sigma L^2 = 1$, where $L = \|K\|$). We observe that CP is quite sensitive to these parameters, and it is difficult to devise, in general, a more convenient choice, while our approach with the parameter settings described above seems to be always comparable to the best results obtained by CP in terms of objective function decrease with respect to both the iteration number and the computational time.

We have run the simulations also for $D_k = I$. In this case, the identity scaling matrix gave rise to slower convergence as compared to the scaling matrix described above. The advantages of using a variable metric have already been noticed in many smooth optimization problems arising in different applications [8, 9, 10, 11]. Our results are coherent with what has been already observed in the recent literature.

5.2. Image deconvolution in the presence of signal-dependent Gaussian noise. In the second example, we study the image reconstruction problem described in [15], where the observed data $b \in \mathbb{R}^n$ are governed by the model

$$b_i = (Hx_{\mathrm{true}})_i + \sigma_i((Hx_{\mathrm{true}})_i)w_i,$$

where $x_{\mathrm{true}} \in \mathbb{R}^n$ is the true image to be recovered, $H \in \mathbb{R}^{n\times n}$ is a matrix with nonnegative entries associated to the acquisition system, $w = (w_1, \ldots, w_n)^T$ are drawn

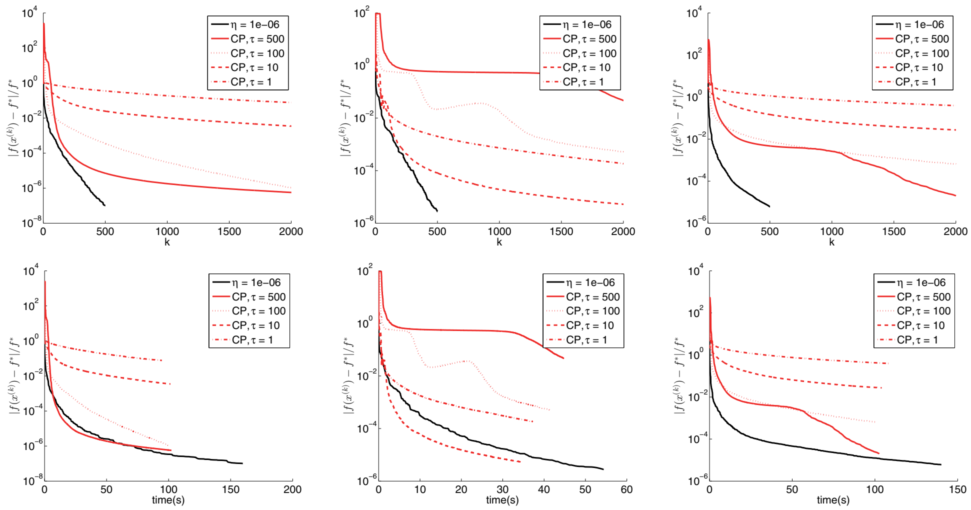

FIG. 4. *Comparison between Algorithm VMILA ($\eta = 10^{-6}$) and the CP algorithm with different choices of parameters. Relative decrease of the objective function values with respect to the outer iteration number (top row) and to the computational time (bottom row). Left column:* `cameraman`. *Middle column:* `micro`. *Right column:* `phantom`.

from Gaussian distribution with zero mean and covariance matrix I_n, and $\sigma_i : \mathbb{R} \to \mathbb{R}_{>0}$ is defined as $\sigma_i(u) = \sqrt{a_i u + c_i}$, with $a_i \in \mathbb{R}_{\geq 0}$, $c_i \in \mathbb{R}_{>0}$, for all $i = 1, \dots, n$.

As explained in [15], an approximation of the true image x_{true} can be found as a solution of the minimization problem (1) where the data discrepancy function is given by

$$f_0(x) = \frac{1}{2} \sum_{i=1}^{n} \frac{((Hx)_i - b_i)^2}{a_i(Hx)_i + c_i} + \log(a_i(Hx)_i + c_i), \tag{64}$$

which is nonconvex and smooth in $\mathrm{dom}(f_0) = \{x \in \mathbb{R}^n : a_i(Hx)_i + c_i > 0 \ \forall i = 1, \dots, n\}$, and the regularization term is chosen as in (63).

Since $c_i > 0$ for all $i = 1, \dots, n$ and H has nonnegative entries, it holds that $\mathrm{dom}(f_0) \supset \mathrm{dom}(f_1)$. We also have that ∇f_0 is Lipschitz continuous in $\mathrm{dom}(f_1)$, but there is no explicit expression for the Lipschitz constant.

We evaluate the performance of the suggested method in comparison with the variable-metric forward–backward (VMFB) algorithm [15] (the implementation is provided by the authors [31]). In particular, we analyze the test problem `jet plane` [31]. In this case, the operator H is a convolution with a truncated Gaussian function of size 7×7, $a_i = c_i = 1$ for all $i = 1, \dots, n$, and $\rho = 0.03$. The approximation $\tilde{y}^{(k)}$ of the proximal operator is handled in the same way as in the previous test problem. The parameters $\alpha_{\min} = 10^{-5}$, $\alpha_{\max} = 10^2$, $\delta = 0.5$, $\beta = 10^{-4}$, $\gamma = 1$, $\eta = 10^{-6}$ used in the MATLAB implementation are also the same as in the previous test problem.

We consider two diagonal scaling matrices D_k corresponding to two choices for the metric:

MM $(D_k)_{ii}^{-1} = \max\{\min\{(A_k)_{ii}, \mu\}, \frac{1}{\mu}\}$, where A_k is defined in [15, formula (36)] with $\varepsilon = 0$. This matrix A_k is chosen such that the quadratic function $Q(x, x^{(k)}) = f_0(x^{(k)}) + \nabla f_0(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2}\|x - x^{(k)}\|_{A_k}^2$ is a majorant function for f_0, i.e., $f_0(x) \leq Q(x, x^{(k)})$ for all $x \in \mathrm{dom}(f_1)$.

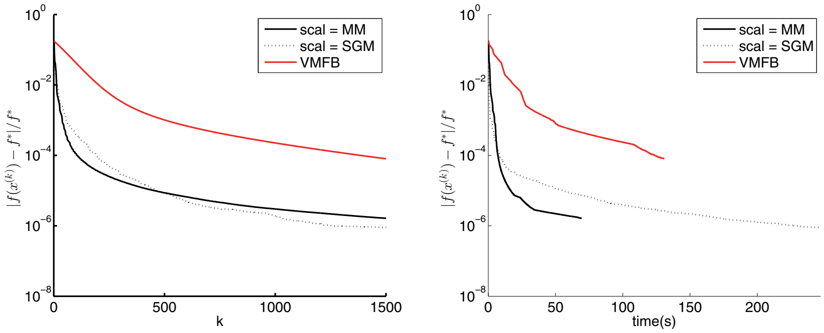

FIG. 5. *Image deconvolution in the presence of signal-dependent Gaussian noise. Relative decrease of the objective function toward the minimum value with respect to the iteration number (left) and computational time in seconds (right).*

SGM $(D_k)_{ii}^{-1} = \max\{\min\{\frac{x_i^{(k)}}{V_i(x^{(k)})+\epsilon}, \mu\}, \frac{1}{\mu}\}$, where ϵ is set to the machine precision and $V(x^{(k)})$ is defined as $V(x^{(k)}) = H^T s^{(k)}$ with

$$s_i^{(k)} = (Hx)_i \frac{a_i((Hx)_i + b_i) + 2c_i}{2(a_i(Hx)_i + c_i)^2} + \frac{a_i}{2(a_i(Hx)_i + c_i)}.$$

This scaling matrix is again chosen mimicking the split-gradient idea [25]. The bound μ of the diagonal entries of D_k is set equal to 10^{10}. The stepsize parameter α_k is chosen using the Barzilai–Borwein rules as before.

In our experiments, both methods achieve the same value of the objective function in the limit, which we denote by f^* (this is not true in general for nonconvex problems). Thus we can compare the convergence properties of the algorithms by showing the decay toward this value (it has been precomputed by running 5000 iterations of all methods and retaining the smallest value).

The results have been obtained on a PC equipped by an Intel Core i7-3667U processor with CPU at 2.0GHz and 8GB of RAM, running Linux Ubuntu 64-bit OS and MATLAB R2012a.

Figure 5 reports the relative decrease of the objective function with respect to the minimum value f^* as a function of the iteration number and of the computational time. We observe a faster decrease of the objective function for Algorithm VMILA.

6. Conclusions and future work. In this paper we presented and analyzed an inexact variable-metric forward–backward method based on an Armijo-type line-search along a suitable descent direction. The inexactness of the method lies in the possibility of using an approximation of the proximal operator, while the underlying metric may change at each iteration and also non-Euclidean metrics are allowed. We performed the convergence analysis of the method, obtaining results in both the nonconvex and convex cases and providing also a convergence rate estimate in the latter one. The main strengths of the method are listed below.

- For nonconvex problems we proved a weak convergence result (Theorem 3.1) which holds true for inexact computation of the proximal point (for which we provide implementable conditions). This result is not given in [1, 15], nor under the weaker Kurdyka–Łojasiewicz hypothesis.
- The convergence is ensured by a line-search procedure, which does not depend on any user-supplied parameter (actually the constants γ, β, δ have to be

chosen, but the behavior of the whole algorithm is not sensitive to these choices). On the other side, the "free" parameter σ in (11) could be exploited to accelerate the convergence speed.

- The possibility of using at each iterate an approximation of $p(x^{(k)}; h_\sigma)$ makes the method well suited for the solution of a wide variety of structured problems.
- The numerical results on large-scale problems show that the performances of the inexact method are promising and comparable with those of state-of-the-art methods.

Future work will be addressed especially to deepen the theoretical and numerical analysis in the nonconvex case, investigating the possibility of obtaining convergence results stronger than those stated in Theorems 3.1 and 3.2, at least for some classes of nonconvex functions (e.g., Kurdyka–Łojasiewicz functions).

Acknowledgment. We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

REFERENCES

[1] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., 137 (2013), pp. 91–129.

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

[3] A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optim., 16 (2006), pp. 697–725.

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

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

[6] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

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

[8] 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.

[9] 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.

[10] S. Bonettini and M. Prato, *New convergence results for the scaled gradient projection method*, Inverse Problems, 31 (2015), 095008.

[11] S. Bonettini, R. Zanella, and L. Zanni, *A scaled gradient projection method for constrained image deblurring*, Inverse Problems, 25 (2009), 015002.

[12] K. Bredies and D. A. Lorenz, *Linear convergence of iterative soft-thresholding*, J. Fourier Anal. Appl., 14 (2008), pp. 813–837.

[13] A. Chambolle and C. Dossal, *On the Convergence of the Iterates of "FISTA,"* hal-01060130v3, 2014.

[14] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vis., 40 (2011), pp. 120–145.

[15] 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.

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

[17] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[18] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, H. H. Bauschke,

R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, eds., Springer Optim. Appl. 49, Springer, New York, 2011, pp. 185–212.

[19] P. L. Combettes and B. C. Vũ, *Variable metric quasi-Féjer monotonicity*, Nonlinear Anal., 78 (2013), pp. 17–31.

[20] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.

[21] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.

[22] P. Frankel, G. Garrigos, and J. Peypouquet, *Splitting methods with variable metric for Kurdyka-Lojasiewicz functions and general convergence rates*, J. Optim. Theory Appl., 165 (2015), pp. 874–900.

[23] 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.

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

[25] H. Lantéri, M. Roche, and C. Aime, *Penalized maximum likelihood image restoration with positivity constraints: Multiplicative algorithms*, Inverse Problems, 18 (2002), pp. 1397–1419.

[26] I. Loris and C. Verhoeven, *On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty*, Inverse Problems, 27 (2011), 125007.

[27] B. Polyak, *Introduction to Optimization*, Optimization Software Inc., Publication Division, New York, 1987.

[28] F. Porta and I. Loris, *On some steplength approaches for proximal algorithms*, Appl. Math. Comput., 253 (2015), pp. 345–362.

[29] 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.

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

[31] A. Repetti and E. Chouzenoux, *RestoVMFB_Lab: Matlab Toolbox for Image Restoration with the Variable Metric Forward-Backward Algorithm*, 2013, http://www-syscom.univ-mlv.fr/~chouzeno/Logiciel.html.

[32] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[33] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, J. Phys. D, 60 (1992), pp. 259–268.

[34] S. Salzo and S. Villa, *Inexact and accelerated proximal point algorithms*, J. Convex Anal., 19 (2012), pp. 1167–1192.

[35] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.

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

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

[38] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.