

This is the peer reviewed version of the following article:

Variable metric inexact line-search based methods for nonsmooth optimization / Bonettini, Silvia; Loris, Ignace; Porta, Federica; Prato, Marco. - In: SIAM JOURNAL ON OPTIMIZATION. - ISSN 1052-6234. - STAMPA. - 26:2(2016), pp. 891-921. [10.1137/15M1019325]

*Terms of use:*

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

24/12/2025 10:37

# 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 non differentiable, 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 step size 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 we present the results of a numerical experience on a convex total variation based image restoration problem, showing that the proposed approach is competitive with another state-of-the-art method.

**Key words.** Proximal algorithms, nonsmooth optimization, generalized projection, nonconvex optimization.

**AMS subject classifications.** 65K05, 90C30

**1. Introduction.** In this paper we consider the problem

$$\min_{x \in \mathbb{R}^n} f(x) \equiv f_0(x) + f_1(x) \quad (1.1)$$

where  $f_1$  is a proper, convex, lower semicontinuous function and  $f_0$  is smooth, i.e. continuously differentiable, on an open subset  $\Omega_0$  of  $\mathbb{R}^n$  containing  $\text{dom}(f_1) = \{x \in \mathbb{R}^n : f_1(x) < +\infty\}$ .

We also assume that  $f_1$  is bounded from below and that  $\text{dom}(f_1)$  is non-empty and closed. Formulation (1.1) includes also 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  $\Omega$ , i.e.  $f_1 = \iota_\Omega$  with

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

a simple and well studied algorithm for the solution of (1.1) is the gradient projection (GP) method, which is particularly appealing for large scale problems. In the last years, several variants of such method have been proposed [7, 10, 18, 21], 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 in

$$x^{(k+1)} = x^{(k)} + \lambda^{(k)}(y^{(k)} - x^{(k)}), \quad (1.2)$$

where  $y^{(k)}$  is the Euclidean projection of the point  $x^{(k)} - \nabla f_0(x^{(k)})$  onto the feasible set  $\Omega$  and  $\lambda^{(k)} \in [0, 1]$  is a steplength parameter ensuring the sufficient decrease of the objective function. Typically,  $\lambda^{(k)}$  is determined by means of a backtracking loop until an Armijo-type inequality is satisfied. Variants of the

\*This work has been partially supported by MIUR under the two projects FIRB - Futuro in Ricerca 2012, contract RBFR12M3AC and PRIN 2012, contract 2012MTE38N. Ignace Loris is a Research Associate of the Fonds de la Recherche Scientifique - FNRS. The Italian GNCS - INdAM is also acknowledged.

\*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).

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

basic scheme are obtained by introducing a further variable stepsize parameter  $\alpha_k$ , 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

$$y^{(k)} = \arg \min_{y \in \Omega} \nabla f_0(x^{(k)})^T (y - x^{(k)}) + \frac{1}{2\alpha_k} (y - x^{(k)})^T D_k (y - x^{(k)}) \quad (1.3)$$

where  $\alpha_k$  is a positive parameter and  $D_k \in \mathbb{R}^{n \times n}$  is a symmetric positive definite matrix. The stepsizes  $\alpha_k$  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 behaviour [7, 8, 10].

In this paper we generalize the GP scheme (1.2)–(1.3), 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 (1.1). In particular, we focus on the case when the descent direction has the form  $y^{(k)} - x^{(k)}$ , with

$$y^{(k)} = \arg \min_{y \in \mathbb{R}^n} \nabla f_0(x^{(k)})^T (y - x^{(k)}) + d_{\sigma^{(k)}}(y, x^{(k)}) + f_1(y) - f_1(x^{(k)}), \quad (1.4)$$

where  $d_{\sigma^{(k)}}(\cdot, \cdot)$  plays the role of a distance function, depending on the parameter  $\sigma^{(k)} \in \mathbb{R}^q$ . Clearly, (1.4) is a generalization of (1.3), 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 (1.2)–(1.4) is a forward–backward (or proximal gradient) method [15, 16] depending on the parameters  $\lambda^{(k)}, \sigma^{(k)}$ .

In particular, we deeply investigate the variant of the scheme (1.2)–(1.4) where the minimization problem in (1.4) 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  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  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 (1.1). 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 (1.1), if it exists. In the latter case, under the further assumption that  $\nabla f_0$  is Lipschitz continuous, we give a  $\mathcal{O}(\frac{1}{k})$  convergence rate estimate for the objective function values. Our analysis includes as special cases several state-of-the-art methods, as those in [7, 9, 10, 26, 32].

Forward–backward algorithms based on a variable metric have been recently studied also in [14] for the convex case and in [13] for the nonconvex case under the Kurdyka–Łojasiewicz assumption (see also [20]). Even if our scheme is formally very similar to those in [13, 14], 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.  $\alpha_k$  and  $D_k$ ), without necessarily relating them to the Lipschitz constant of  $\nabla f_0$  (actually, our analysis, except the convergence rate estimate, is performed without this assumption). We believe that this is also one of the main strength of our method, since acceleration techniques based on suitable choices of  $\alpha_k$  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 (1.4): this issue has been considered in several papers in the context of proximal and proximal gradient methods (see for example [1, 13, 31, 33] and references therein). The approach we follow in this paper is more similar to the one proposed in [33] and has the advantage to provide 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, 19] for the use of non Euclidean distances in the context of forward–backward and proximal methods.

The paper is organized as follows: some background material is collected in Section 2, while the concept of descent direction for problem (1.1) is presented and developed in Section 3. In Section 4, 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 of  $\epsilon$ -type and  $\eta$ -type are proposed in Sections 4.2 and 4.3, and the related implementation is discussed in Sections 5.1 and 5.4. Section 4.5 deals with the convex case, where the convergence of an  $\epsilon$ -approximation based algorithm is proved and the related convergence rate is analyzed. The results of a numerical experience on a total variation based image restoration problem are presented in Section 6 while our conclusions are given in Section 7.

*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 non-negative and positive real numbers, respectively. The scaled Euclidean norm of an  $n$ -vector  $x$ , associated to a symmetric positive definite matrix  $D$  is  $\|x\|_D = \sqrt{x^T D x}$ . 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

$$\frac{1}{\mu} \|x\|^2 \leq \|x\|_D^2 \leq \mu \|x\|^2 \quad (1.5)$$

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

**2. Definitions and basic properties.** We recall the following definitions.

**DEFINITION 2.1** [29, p.213] *Let  $f$  be any function from  $\mathbb{R}^n$  to  $\bar{\mathbb{R}}$ . The one sided directional derivative of  $f$  at  $x$  with respect to a vector  $d$  is defined as*

$$f'(x; d) = \lim_{\lambda \downarrow 0} \frac{f(x + \lambda d) - f(x)}{\lambda} \quad (2.1)$$

*if the limit on the right-hand side exists in  $\bar{\mathbb{R}}$ .*

When  $f$  is smooth at  $x$ , then  $f'(x; d) = \nabla f(x)^T d$ . When  $f$  is convex, its directional derivative has the following property.

**THEOREM 2.1** [29, Theorem 23.1] *If  $f$  is convex and  $x \in \text{dom}(f)$ , then for any  $d \in \mathbb{R}^n$  the limit at the right-hand side of (2.1) exists and  $f'(x; d) = \inf_{\lambda > 0} \frac{f(x + \lambda d) - f(x)}{\lambda}$ .*

As a consequence of the previous theorem, for any convex function  $f$  we have that  $f'(x; d)$  exists for any  $x \in \text{dom}(f)$ ,  $d \in \mathbb{R}^n$  and

$$f'(x; d) \leq f(x + d) - f(x). \quad (2.2)$$

**DEFINITION 2.2** [32, p. 394] *A point  $x$  is stationary for problem (1.1) if  $x \in \text{dom}(f)$  and*

$$f'(x; d) \geq 0 \quad \forall d \in \mathbb{R}^n. \quad (2.3)$$

**DEFINITION 2.3** [20, §2.3] *The proximity or resolvent operator associated to a convex function  $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  in the metric induced by a symmetric positive definite matrix  $D$  is defined as*

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

We remark that  $\text{prox}_f^D$  is a Lipschitz continuous function whose Lipschitz constant is  $\|D\|$ .

**DEFINITION 2.4** *Let  $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  be a convex function. The conjugate function of  $f$  is the function  $f^* : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  defined as  $f^*(y) = \sup_{x \in \mathbb{R}^n} x^T y - f(x) \quad \forall y \in \mathbb{R}^n$ .*

The following proposition states a useful property of the conjugate.

**PROPOSITION 2.1** *Let  $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ ,  $g : \mathbb{R}^m \rightarrow \bar{\mathbb{R}}$  be two convex functions,  $A \in \mathbb{R}^{m \times n}$ . If  $f(x) = g(Ax)$ , then  $f^*(A^T y) \leq g^*(y) \forall y \in \mathbb{R}^m$ .*

*Proof.* By Definition 2.4 we have

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

□

**DEFINITION 2.5** [35, p. 82] *Given  $\epsilon \in \mathbb{R}_{\geq 0}$ , the  $\epsilon$ -subdifferential of a convex function  $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  at a point  $z \in \mathbb{R}^n$  is the set*

$$\partial_\epsilon f(z) = \{w \in \mathbb{R}^n : f(x) \geq f(z) + (x - z)^T w - \epsilon, \forall x \in \mathbb{R}^n\}. \quad (2.4)$$

If  $z \in \text{dom}(f)$ , then  $\partial_\epsilon f(z) \neq \emptyset$ . For  $\epsilon = 0$  the usual subdifferential set  $\partial f(z)$  is recovered. A useful property of the  $\epsilon$ -subdifferential is the following one.

**PROPOSITION 2.2** [35, Theorem 2.4.4 (iv)] *Let  $f : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  be a convex, proper, lower semicontinuous function. Then for any  $\epsilon \in \mathbb{R}_{\geq 0}$  and for any  $x \in \mathbb{R}^n$  we have  $x^* \in \partial_\epsilon f(x) \Leftrightarrow x \in \partial_\epsilon f^*(x^*)$ .*

**3. A family of descent directions.** When  $f$  is smooth, a vector  $d \in \mathbb{R}^n$  is said a descent direction for  $f$  at  $x$  when  $\nabla f(x)^T d < 0$ . In the nonsmooth case (1.1), we give the following definition, based on the directional derivative.

**DEFINITION 3.1** *A vector  $d \in \mathbb{R}^n$  is a descent direction for  $f$  at  $x \in \text{dom}(f)$  if  $f'(x; d) < 0$ .*

Thanks to Theorem 2.1, the previous definition is well posed. In this section we define a family of descent directions for problem (1.1). To this end, we define the following set of non-negative 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  $d_\sigma : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0} \cup \{+\infty\}$  continuously depending on  $\sigma \in S$  such that for all  $z, x \in \Omega$  we have:

- (D<sub>1</sub>)  $d_\sigma(z, x)$  is continuous in  $(\sigma, z, x)$ ;
- (D<sub>2</sub>)  $d_\sigma(z, x)$  is smooth w.r.t.  $z \in \Omega$ ;
- (D<sub>3</sub>)  $d_\sigma(z, x)$  is strongly convex w.r.t.  $z$ :

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

where  $m > 0$  does not depend on  $\sigma$  or  $x$  (here  $\nabla_1$  denotes the gradient with respect to the first argument of a function);

- (D<sub>4</sub>)  $d_\sigma(z, x) = 0$  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

$$d_\sigma(x, y) = \frac{1}{2\alpha} \|x - y\|_D^2 \quad (3.1)$$

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.

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

$$h_\sigma(z, x) = \nabla f_0(x)^T (z - x) + d_\sigma(z, x) + f_1(z) - f_1(x) \quad \forall z, x \in \mathbb{R}^n, \quad (3.2)$$

where  $d_\sigma \in \mathcal{D}(\Omega, S)$  and  $\Omega = \text{dom}(f_1)$ . We remark that  $h_\sigma$  depends continuously on  $\sigma$ , as  $d_\sigma$  does. 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

$$h'_\sigma(x, x; d) = f'(x; d), \quad (3.3)$$

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  $p : \Omega_0 \rightarrow \Omega$  associated to any function  $h_\sigma$  of the form (3.2)

$$p(x; h_\sigma) = \arg \min_{z \in \mathbb{R}^n} h_\sigma(z, x). \quad (3.4)$$

When  $d_\sigma$  is chosen as in (3.1), the operator (3.4) becomes

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

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

**PROPOSITION 3.1** *Let  $d_\sigma \in \mathcal{D}(\Omega, S)$  and  $h_\sigma$  be defined as in (3.2). Then  $p(x; h_\sigma)$  depends continuously on  $(x, \sigma)$ .*

*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  $f_1(u) \geq f_1(y) + w^T(u - y)$  for all  $u \in \mathbb{R}^n$  or:

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

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

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

Together, these two inequalities yield:

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

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:

$$m \|y_1 - y_2\|^2 \leq d_{\sigma_1}(y_2, x_1) - d_{\sigma_1}(y_1, x_1) + d_{\sigma_2}(y_1, x_2) - d_{\sigma_2}(y_2, x_2) + (\nabla f_0(x_1) - \nabla f_0(x_2))^T(y_2 - y_1)$$

and hence:

$$m \|y_1 - y_2\|^2 \leq d_{\sigma_2}(y_1, x_2) - d_{\sigma_1}(y_1, x_1) + d_{\sigma_1}(y_2, x_1) - d_{\sigma_2}(y_2, x_2) + \|\nabla f_0(x_1) - \nabla f_0(x_2)\| \|y_2 - y_1\|.$$

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  $\lim_{x_2 \rightarrow x_1} b = 0$ . As  $d_\sigma(z, x)$  is continuous in  $(\sigma, z, x)$ , one also has that  $\lim_{x_2 \rightarrow x_1} c = 0$ . This shows then that  $\lim_{x_2 \rightarrow x_1} \|y_2 - y_1\| = 0$ , in other words  $p(x_1; h_{\sigma_1})$  is continuous in  $(\sigma_1, x_1)$ .  $\square$

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

$$\tilde{h}_{\sigma, \gamma}(z, x) = \nabla f_0(x)^T(z - x) + \gamma d_\sigma(z, x) + f_1(z) - f_1(x) \quad \forall z, x \in \mathbb{R}^n \quad (3.5)$$

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

$$\tilde{h}_{\sigma, \gamma}(y, x) \leq h_\sigma(y, x) \quad \forall x, y \in \mathbb{R}^n \quad (3.6)$$

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

- the stationarity condition (2.3) 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  $h_\sigma$  and the associated operator  $p(\cdot; h_\sigma)$ .

**PROPOSITION 3.2** *Let  $\sigma \in S \subseteq \mathbb{R}^q$ ,  $\gamma \in [0, 1]$ ,  $d_\sigma \in \mathcal{D}$  and  $h_\sigma, \tilde{h}_{\sigma,\gamma}$  be defined as in (3.2), (3.5), 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)  $\tilde{h}_{\sigma,\gamma}(y, x) \leq 0$  and  $\tilde{h}_{\sigma,\gamma}(y, x) = 0$  if and only if  $y = x$ ;
- (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) is a direct consequence of definition (3.5) and condition  $(\mathcal{D}_3)$  on  $d_\sigma$ .

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

$$0 \geq -\gamma d_\sigma(z, x) > \nabla f_0(x)^T(z - x) + f_1(z) - f_1(x) \geq \nabla f_0(x)^T(z - x) + f'_1(x; z - x) = f'(x; z - x),$$

where the second inequality follows from definition (3.5) of  $\tilde{h}_{\sigma,\gamma}$  and the third one from (2.2).

(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 (3.6), 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 (3.6) 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 (1.1) 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 3.3** *Let  $S \subseteq \mathbb{R}^q$ ,  $\sigma \in S$ ,  $h_\sigma, \tilde{h}_{\sigma,\gamma}$  be defined as in (3.2),  $\gamma \in [0, 1]$ ,  $x \in \Omega$  and  $y = p(x; h_\sigma)$ . The following statements are equivalent:*

- (a)  $x$  is stationary for problem (1.1);
- (b)  $x = y$ ;
- (c)  $\tilde{h}_{\sigma,\gamma}(y, x) = 0$ .

*Proof.* (a)  $\iff$  (b) Assume that  $x = y$ . Then,  $h_\sigma(\cdot, x)$  achieves its minimum at  $x$  and inequality (2.3) applied to it yields  $h'_\sigma(x, x; z - x) \geq 0 \quad \forall z \in \mathbb{R}^n$ . Recalling (3.3) we have  $h'_\sigma(x, x; z - x) = f'(x; z - x)$ , hence  $x$  is a stationary point for problem (1.1).

Conversely, let  $x \in \Omega$  be a stationary point of (1.1) and assume by contradiction that  $x \neq y$ . Then, by Proposition 3.2 (d) we obtain  $f'(x, y - x) < 0$ , which contradicts the stationarity assumption on  $x$ .

(b)  $\iff$  (c) See Proposition 3.2 (c). □

**4. 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 [32]. Indeed the rule proposed in [32] is recovered when  $d_\sigma$  is chosen as in (3.1) 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.

**Algorithm LS** Modified Armijo linesearch algorithm

Let  $\{x^{(k)}\}_{k \in \mathbb{N}}$ ,  $\{\tilde{y}^{(k)}\}_{k \in \mathbb{N}}$  be two sequences of points in  $\Omega$ , and  $\{\sigma^{(k)}\}_{k \in \mathbb{N}}$  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

$$f(x^{(k)} + \lambda^{(k)} d^{(k)}) \leq f(x^{(k)}) + \beta \lambda^{(k)} \Delta^{(k)} \quad (4.1)$$

where

$$\Delta^{(k)} = \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) \quad (4.2)$$

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 will define the function  $h_\sigma(\cdot, \cdot)$  as in (3.2) and, for sake of simplicity, we will make the following assumption

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

**PROPOSITION 4.1** *Let  $\{x^{(k)}\}_{k \in \mathbb{N}}$ ,  $\{\tilde{y}^{(k)}\}_{k \in \mathbb{N}}$  be two sequences of points in  $\Omega$ ,  $\{\sigma^{(k)}\}_{k \in \mathbb{N}}$  a sequence of parameters in  $S \subseteq \mathbb{R}^q$  and  $\gamma \in [0, 1]$ . Assume that*

$$\tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) < 0 \quad (4.3)$$

*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  $f(x^{(k+1)}) \leq f(x^{(k)})$ , then we have that  $\Delta^{(k)} = \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)})$  is bounded. Assuming also that*

$$\lim_{k \rightarrow \infty} f(x^{(k)}) - f(x^{(k)} + \lambda^{(k)} d^{(k)}) = 0, \quad (4.4)$$

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

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

*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

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

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  $j \rightarrow \infty$  we obtain

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



where the second inequality follows from the non-negativity of  $d_\sigma \in \mathcal{D}(\Omega, S)$  and the last one from (4.3). 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  $f(x^{(k+1)}) \leq f(x^{(k)})$ . We show that  $\Delta^{(k)} = \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)})$  is bounded. By assumption (4.3),  $\tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)})$  is bounded from above. We show that it is also bounded from below. Indeed we have

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

where the first inequality follows from the non-negativity of  $d_\sigma$ , the second one is obtained by adding and subtracting  $f_0(x^{(k)})$  and the last one is a consequence of  $f(x^{(k+1)}) \leq f(x^{(k)})$ .

As  $f_1$  is proper and convex, there exists a supporting hyperplane, i.e.  $\exists a, b \in \mathbb{R}^n$  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.

Let us show that the only limit point of  $\Delta^{(k)}$  is zero. We observe that from (4.3) and (4.4) we obtain

$$0 = \lim_{k \rightarrow \infty} f(x^{(k)}) - f(x^{(k)} + \lambda^{(k)} d^{(k)}) = \beta \lim_{k \rightarrow \infty} \Delta^{(k)} \lambda^{(k)}. \quad (4.5)$$

Assume that there exists a subset of indices  $K \subseteq \mathbb{N}$  such that  $\lim_{k \in K, k \rightarrow \infty} \Delta^{(k)} = \bar{\Delta} \in \mathbb{R}$ , with  $\bar{\Delta} < 0$ . By (4.5), this implies that

$$\lim_{k \in K, k \rightarrow \infty} \lambda^{(k)} = 0. \quad (4.6)$$

Denote by  $\bar{K} \subseteq K$  a set of indices such that  $\lim_{k \in \bar{K}, k \rightarrow \infty} \sigma^{(k)} = \bar{\sigma}$ ,  $\lim_{k \in \bar{K}, k \rightarrow \infty} x^{(k)} = \bar{x}$  and  $\lim_{k \in \bar{K}, k \rightarrow \infty} \tilde{y}^{(k)} = \bar{y}$  for some  $\bar{\sigma} \in S$ ,  $\bar{x}, \bar{y} \in \Omega$ . From (4.6) 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  $k \in \bar{K}, k \rightarrow \infty$ , since  $\{d^{(k)} = \tilde{y}^{(k)} - x^{(k)}\}_{k \in \mathbb{N}}$  is bounded and by (4.6) we obtain  $\beta\bar{\Delta} \leq \bar{\Delta} < 0$ , which is an absurdum, being  $0 < \beta < 1$ .  $\square$

We prove also the following useful Lemma.

**LEMMA 4.1** *Let  $\{x^{(k)}\}_{k \in \mathbb{N}}$ ,  $\{\tilde{y}^{(k)}\}_{k \in \mathbb{N}}$  be two sequences of points in  $\Omega$ ,  $\{\sigma^{(k)}\}_{k \in \mathbb{N}}$  a sequence of parameters in  $S \subseteq \mathbb{R}^q$  and  $\gamma \in [0, 1]$ . Assume that*

$$f(x^{(k+1)}) \leq f(x^{(k)} + \lambda^{(k)} d^{(k)}), \quad d^{(k)} = \tilde{y}^{(k)} - x^{(k)} \quad (4.7)$$

where  $\tilde{y}^{(k)}$  satisfies (4.3) and  $\lambda^{(k)}$  is computed by Algorithm LS for any  $k \in \mathbb{N}$ . Suppose that  $f$  is bounded from below. Then, we have

$$0 \leq - \sum_{k=0}^{\infty} \lambda^{(k)} \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) < \infty. \quad (4.8)$$

*Proof.* Denote by  $\ell \in \mathbb{R}$  a lower bound for  $f$ , i.e.  $\ell \leq f(x) \forall x \in \mathbb{R}^n$ . Inequalities (4.1) and (4.7) 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

$$-\beta \sum_{k=0}^j \lambda^{(k)} \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) \leq \sum_{k=0}^j (f(x^{(k)}) - f(x^{(k+1)})) = f(x^{(0)}) - f(x^{(j+1)}) \leq f(x^{(0)}) - \ell. \quad (4.9)$$

Thus, inequality (4.8) follows.  $\square$

**4.1. A class of line-search based algorithms.** Proposition 4.1 allows the convergence analysis of a wide class of descent methods based on the Armijo condition (4.1). 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 at least to  $\lambda^{(k)} \tilde{h}_{\sigma, \gamma}(\tilde{y}^{(k)}, x^{(k)})$ , where  $\lambda^{(k)}$  is determined by the backtracking procedure given in Algorithm LS.

**THEOREM 4.1** *Let  $\{x^{(k)}\}_{k \in \mathbb{N}}$ ,  $\{\tilde{y}^{(k)}\}_{k \in \mathbb{N}}$  be two sequences of points in  $\Omega$ ,  $\{\sigma^{(k)}\}_{k \in \mathbb{N}} \subset S$  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  $\lim_{k \in K', k \rightarrow \infty} x^{(k)} = \bar{x} \in \Omega$ . 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 (4.3) and there exists  $K'' \subseteq K'$  such that

$$\lim_{k \in K'', k \rightarrow \infty} h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) = 0, \quad \text{with } y^{(k)} = p(x^{(k)}; h_{\sigma^{(k)}}). \quad (4.10)$$

Then  $\bar{x}$  is a stationary point for problem (1.1).

*Proof.* First, we notice that Algorithm LS is well defined, since (4.3) 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

$$\frac{m}{2} \|z - y^{(k)}\|^2 \leq h_{\sigma^{(k)}}(z, x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \quad \forall z \in \mathbb{R}^n. \quad (4.11)$$

Setting  $z = \tilde{y}^{(k)}$  in the previous inequality and using (4.10) gives

$$\lim_{k \in K'', k \rightarrow \infty} \|\tilde{y}^{(k)} - y^{(k)}\| = 0. \quad (4.12)$$

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, (4.12) 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  $\lim_{k \in K, k \rightarrow \infty} \tilde{y}^{(k)} = \bar{y}$  and  $\lim_{k \in K, k \rightarrow \infty} \sigma^{(k)} = \bar{\sigma}$ . By continuity of the operator  $p(x; h_{\sigma})$  with respect to all its arguments, (4.12) implies that  $\bar{y} = p(\bar{x}; h_{\bar{\sigma}})$ .

Consider now the sequence  $\{f(x^{(k)})\}_{k \in \mathbb{N}}$ . From assumption (4.7) it follows that

$$f(x^{(k+1)}) \leq f(x^{(k)} + \lambda^{(k)} d^{(k)}) \leq f(x^{(k)}). \quad (4.13)$$

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

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

The previous inequality implies that  $\bar{f} \in \mathbb{R}$  and this fact, together with inequality (4.13), gives

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

Thus we can apply Proposition 4.1 and obtain

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

Combining the previous equality with (3.6) and (4.10) yields

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

Since  $h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \leq 0$ , this implies  $\lim_{k \rightarrow \infty, k \in K} h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) = 0$ . Expressing inequality (4.11) for  $z = x^{(k)}$ , we can write

$$\frac{m}{2} \|x^{(k)} - y^{(k)}\|^2 \leq h_{\sigma^{(k)}}(x^{(k)}, x^{(k)}) - h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) = -h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \xrightarrow{k \rightarrow \infty, k \in K} 0.$$

Thus, we proved that  $\bar{y} = \bar{x}$ , and, by Proposition 3.3 we have that  $\bar{x}$  is stationary.  $\square$

Let us now discuss assumption (4.10) in the previous theorem, concerning the inexact solution of the minimum problem in (3.4). Assumption (4.3) 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, but we need also to assume that (4.10) holds.

As 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 4.1 except (4.10). However, starting from  $x^{(0)} = 2$ , the sequence writes as  $x^{(k)} = 1 + (1/2)^k \xrightarrow{k \rightarrow \infty} 1$ , while the only stationary point is 0.

We remark that assumption (4.10) could be replaced by requiring that  $f_1$  is continuous and (4.12) holds. Clearly, (4.10) cannot be checked directly, but it is very general. In the following sections, we will consider two implementable conditions which imply (4.10) and in Sections 5.1–5.4 we show how  $\tilde{y}^{(k)}$  can be computed in practice without knowing  $p(x^{(k)}; h_{\sigma^{(k)}})$ .

**4.2.  $\epsilon$ -approximations.** In this section we will assume that  $d_\sigma$  has the form (3.1) and, in this case, we will describe a sufficient condition for (4.10).

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

$$\frac{1}{\alpha} D(z - y) \in \partial f_1(y), \quad (4.14)$$

where  $z = x - \alpha D^{-1} \nabla f_0(x)$ . Borrowing the ideas in [31, 33], we consider a relaxed version of (4.14) and we study the properties of any point  $\tilde{y}$  satisfying the following inclusion

$$\frac{1}{\alpha} D(z - \tilde{y}) \in \partial_\epsilon f_1(\tilde{y}), \quad (4.15)$$

where  $\epsilon \in \mathbb{R}_{\geq 0}$ .

LEMMA 4.2 *Let  $d_\sigma$  be defined as in (3.1) and  $x \in \Omega$ . Assume that  $y = p(x; h_\sigma)$  and that  $\tilde{y}$  satisfies (4.15) 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 [35, Theorem 2.4.2 viii]), inclusion (4.15) implies  $0 \in \partial_\epsilon h_\sigma(\tilde{y}, x)$  which, by definition (2.4) of  $\epsilon$ -subdifferential, is equivalent to

$$h_\sigma(w, x) \geq h_\sigma(\tilde{y}, x) - \epsilon \quad \forall w \in \mathbb{R}^n. \quad (4.16)$$

We 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 (4.16) with  $w = y$ .  $\square$

The previous result combined with Theorem 4.1 directly implies the following Corollary.

COROLLARY 4.1 *Let  $0 < \alpha_{\min} \leq \alpha_{\max}$ ,  $\gamma \in [0, 1]$ ,  $\mu \geq 1$ . Assume that  $\{\alpha_k\}_{k \in \mathbb{N}} \subset [\alpha_{\min}, \alpha_{\max}]$ ,  $\{D_k\}_{k \in \mathbb{N}} \subset \mathcal{M}_\mu$ ,  $\{\epsilon_k\}_{k \in \mathbb{N}} \subset \mathbb{R}_{\geq 0}$ ,  $\lim_{k \rightarrow \infty} \epsilon_k = 0$ . Let  $\{x^{(k)}\}_{k \in \mathbb{N}}$ ,  $\{\tilde{y}^{(k)}\}_{k \in \mathbb{N}}$  be two sequences of points in  $\Omega$  such that, for any  $k \in \mathbb{N}$ , (4.7) holds, where  $\lambda^{(k)}$  is computed by Algorithm LS and  $\tilde{y}^{(k)}$  satisfies (4.3) and*

$$\frac{1}{\alpha_k}D_k(z^{(k)} - \tilde{y}^{(k)}) \in \partial_{\epsilon_k} f_1(\tilde{y}^{(k)}), \quad (4.17)$$

*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 (1.1).*

**4.3.  $\eta$ -approximations.** A different approach to define a suitable approximation of the operator (3.4) is based on the following definition.

$$P_\eta(x; h_\sigma) = \{\tilde{y} \in \Omega : h_\sigma(\tilde{y}, x) \leq \eta h_\sigma(y, x), \text{ where } y = p(x; h_\sigma)\} \quad (4.18)$$

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

$$\tilde{y} \in P_\eta(x; h_\sigma), \quad (4.19)$$

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 4.2 *Let  $\eta \in (0, 1]$ ,  $0 \leq \gamma \leq 1$ ,  $\{\sigma^{(k)}\}_{k \in \mathbb{N}} \subset S$  and  $\{x^{(k)}\}_{k \in \mathbb{N}} \subset \Omega$  satisfying (4.7), where  $\lambda^{(k)}$  is computed by Algorithm LS, with*

$$\tilde{y}^{(k)} \in P_\eta(x^{(k)}; h_{\sigma^{(k)}}). \quad (4.20)$$

*Then, either for some  $k$  the iterate  $x^{(k)}$  is stationary for problem (1.1), or any limit point  $\bar{x}$  of  $\{x^{(k)}\}_{k \in \mathbb{N}}$  is stationary for problem (1.1).*

*Proof.* We set  $y^{(k)} = p(x^{(k)}; h_{\sigma^{(k)}})$  and we first observe that  $\gamma \leq 1$  and (4.20) imply

$$\tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) \leq h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) \leq \eta h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \leq 0. \quad (4.21)$$

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 3.3,  $x^{(k)}$  is a stationary point for problem (1.1).

Otherwise  $\tilde{h}_{\sigma(k), \gamma}(\tilde{y}^{(k)}, x^{(k)}) < 0$  for all  $k \in \mathbb{N}$  and, thus, (4.3) holds. Consider now a limit point  $\bar{x} \in \Omega$  of  $\{x^{(k)}\}_{k \in \mathbb{N}}$  (if one exists) such that  $\lim_{k \rightarrow \infty, k \in K'} x^{(k)} = \bar{x}$  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 (4.20) we have

$$h_{\sigma(k)}(\tilde{y}^{(k)}, x^{(k)}) - h_{\sigma(k)}(y^{(k)}, x^{(k)}) \leq (\eta - 1)h_{\sigma(k)}(y^{(k)}, x^{(k)}). \quad (4.22)$$

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

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

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 4.1, we have that  $\{\tilde{h}_{\sigma(k), \gamma}(\tilde{y}^{(k)}, x^{(k)})\}_{k \in K'}$  is bounded from below; then, using inequalities (4.21), 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  $\lim_{k \in K, k \rightarrow +\infty} \sigma^{(k)} = \bar{\sigma}$ ,  $\lim_{k \in K, k \rightarrow +\infty} y^{(k)} = \bar{y}$  for some  $\bar{\sigma} \in S$ ,  $\bar{y} \in \Omega$ . Thanks to the continuity of the operator (3.4), 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 4.1, the existence of a limit point guarantees that (4.4) is satisfied. Then, by Proposition 4.1, we obtain  $\lim_{k \rightarrow \infty, k \in K} \tilde{h}_{\sigma(k), \gamma}(\tilde{y}^{(k)}, x^{(k)}) = 0$ . Combining this with (4.20), we also have

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

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

$$\lim_{k \rightarrow \infty, k \in K} h_{\sigma(k)}(y^{(k)}, x^{(k)}) = 0 \quad (4.23)$$

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

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

with, together with (4.23) gives  $\lim_{k \rightarrow \infty, k \in K} \|y^{(k)} - x^{(k)}\|^2 = 0$ . Thus,  $\bar{y} = \bar{x}$  and by Proposition 3.3, we conclude that  $\bar{x}$  is stationary.  $\square$

**4.4. Remarks.** Different notions of inexactness have been proposed in the literature (see [31, 33] and references therein), especially in the context of proximal point methods, with the aim of approximating the resolvent operator, and some of them could be considered also in our framework. A synthetic description of possible inexactness notions and their relationships is given in Figure 4.1.

It is difficult to insert the inexactness criterion (4.19) in the scheme in Figure 4.1, since the shape of  $P_\eta$  in (4.19) depends on  $x$ , while the implications in Figure 4.1 are independent of  $x$ . In general, we observe that from inequality (4.22) and by definition of  $\epsilon$ -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 (4.19) and (4.15) in Figure 4.2. 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. (3.1) with

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

FIG. 4.1. Connection of different inexactness notions, under the assumption (3.1). The proof of the implications are given in Lemma 4.2 and in [31, Proposition 1].

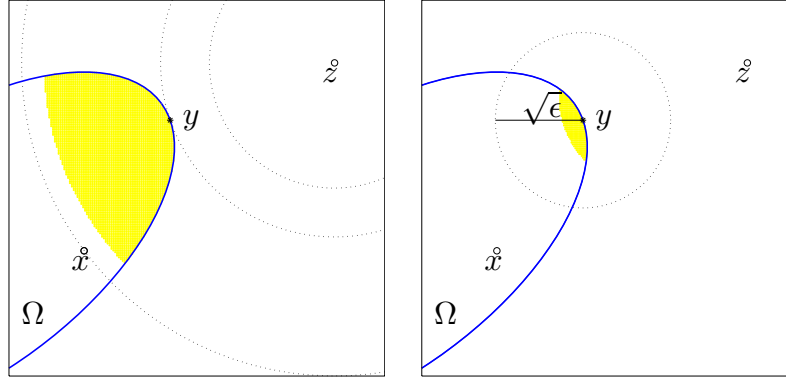


FIG. 4.2. Example with  $f_1(x) = \iota_{\Omega}(x)$ ,  $d_{\sigma}$  as in (3.1) with  $\alpha = 1$ ,  $D = I$ . Left panel: in yellow, the set  $P_{\eta}(x; h_{\sigma})$  defined in (4.18). Right panel: in yellow, the set of points  $\tilde{y}$  satisfying (4.15).

$D = I$ ,  $\alpha = 1$ , as distance function, the operator  $p(x; h_{\sigma})$  reduces to the Euclidean projection of the point  $z = x - \nabla f_0(x)$  onto  $\Omega$ . Moreover, condition (4.15) becomes

$$\tilde{y} \in \Omega \text{ and } (w - \tilde{y})^T(z - \tilde{y}) \leq \epsilon. \quad \forall w \in \Omega. \quad (4.24)$$

As well explained in [31, 33], from a geometrical point of view, a point  $\tilde{y} \in \Omega$  satisfies (4.24) if and only if  $\Omega$  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  $\tilde{y}$ .

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  $\Omega$  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 (4.20) and (4.17) 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  $\epsilon_k$ , 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.

#### 4.5. Convergence analysis in the convex case with $\epsilon$ -approximations.

**4.5.1. 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  $\epsilon$ -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 [9]. Further results on forward-backward variable metric algorithms which apply to problems of the form (1.1) when  $f_0$  has Lipschitz continuous gradient can be found in the recent papers [14, 17]. 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 [14, Theorem 4.1] is given under the stronger assumption that  $\|\tilde{y}^{(k)} - y^{(k)}\|$  is summable.

**THEOREM 4.3** *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 (1.1) is convex and the solution set  $X^*$  of problem (1.1) 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 (4.17), 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

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

Then the sequence  $\{x^{(k)}\}_{k \in \mathbb{N}}$  converges to a solution of (1.1).

*Proof.* First of all we recall the basic norm equality

$$\|a - b\|_D^2 + \|b - c\|_D^2 - \|a - c\|_D^2 = 2(a - b)^T D(c - b) \quad (4.25)$$

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  $z^{(k)} = x^{(k)} - \alpha_k D_k^{-1} \nabla f_0(x^{(k)})$ , writes also as

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

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

$$\begin{aligned} (\tilde{y}^{(k)} - x^{(k)})^T D_k (\hat{x} - x^{(k)}) &\geq \alpha_k \left( f_1(\tilde{y}^{(k)}) - f_1(\hat{x}) + \nabla f_0(x^{(k)})^T (x^{(k)} - \hat{x}) \right) - \alpha_k \epsilon_k + \\ &\quad + \left( \tilde{y}^{(k)} - x^{(k)} + \alpha_k D_k^{-1} \nabla f_0(x^{(k)}) \right)^T D_k (\tilde{y}^{(k)} - x^{(k)}) \\ &\geq \alpha_k \left( f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + f(x^{(k)}) - f(\hat{x}) \right) + \|\tilde{y}^{(k)} - x^{(k)}\|_{D_k}^2 \\ &\quad + \alpha_k \nabla f_0(x^{(k)})^T (\tilde{y}^{(k)} - x^{(k)}) - \alpha_k \epsilon_k \\ &\geq \|\tilde{y}^{(k)} - x^{(k)}\|_{D_k}^2 - \alpha_k \epsilon_k \\ &\quad + \alpha_k \left( f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + \nabla f_0(x^{(k)})^T (\tilde{y}^{(k)} - x^{(k)}) \right) \\ &= \frac{1}{(\lambda^{(k)})^2} \|x^{(k+1)} - x^{(k)}\|_{D_k}^2 - \alpha_k \epsilon_k \\ &\quad + \alpha_k \left( f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) + \nabla f_0(x^{(k)})^T (\tilde{y}^{(k)} - x^{(k)}) \right), \end{aligned} \quad (4.26)$$

where the second inequality is obtained 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

(4.25) with  $a = x^{(k+1)}$ ,  $b = x^{(k)}$ ,  $c = \hat{x}$ ,  $D = D_k$  we obtain

$$\begin{aligned}
\|x^{(k+1)} - \hat{x}\|_{D_k}^2 &= \|x^{(k)} - \hat{x}\|_{D_k}^2 + \|x^{(k+1)} - x^{(k)}\|_{D_k}^2 - 2(x^{(k)} - x^{(k+1)})^T D_k (x^{(k)} - \hat{x}) \\
&= \|x^{(k)} - \hat{x}\|_{D_k}^2 + \|x^{(k+1)} - x^{(k)}\|_{D_k}^2 - 2\lambda^{(k)}(\tilde{y}^{(k)} - x^{(k)})^T D_k (\hat{x} - x^{(k)}) \\
&\stackrel{(4.27)}{\leq} \|x^{(k)} - \hat{x}\|_{D_k}^2 + \left(1 - \frac{2}{\lambda^{(k)}}\right) \|x^{(k+1)} - x^{(k)}\|_{D_k}^2 + \\
&\quad - 2\alpha_k \lambda^{(k)} \left( \nabla f_0(x^{(k)})^T (\tilde{y}^{(k)} - x^{(k)}) + f_1(\tilde{y}^{(k)}) - f_1(x^{(k)}) \right) + 2\alpha_k \lambda^{(k)} \epsilon_k \\
&= \|x^{(k)} - \hat{x}\|_{D_k}^2 + \left(1 - \frac{2}{\lambda^{(k)}} + \frac{\gamma}{\lambda^{(k)}}\right) \|x^{(k+1)} - x^{(k)}\|_{D_k}^2 + \\
&\quad - 2\alpha_k \lambda^{(k)} \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) + 2\alpha_k \lambda^{(k)} \epsilon_k \\
&\leq \|x^{(k)} - \hat{x}\|_{D_k}^2 - 2\alpha_k \lambda^{(k)} \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) + 2\alpha_k \lambda^{(k)} \epsilon_k,
\end{aligned} \tag{4.28}$$

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

$$\begin{aligned}
\|x^{(k+1)} - \hat{x}\|_{D_{k+1}}^2 &\leq (1 + \zeta_k) \|x^{(k+1)} - \hat{x}\|_{D_k}^2 \\
&\leq (1 + \zeta_k) \|x^{(k)} - \hat{x}\|_{D_k}^2 - 2\alpha_k (1 + \zeta_k) \lambda^{(k)} \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) \\
&\quad + 2\alpha_k \lambda^{(k)} (1 + \zeta_k) \epsilon_k \\
&\leq (1 + \zeta_k) \|x^{(k)} - \hat{x}\|_{D_k}^2 - 2\alpha_{\max} \zeta \lambda^{(k)} \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) + 2\alpha_{\max} \zeta \epsilon_k
\end{aligned} \tag{4.29}$$

where we set  $\zeta = 1 + \max_k \zeta_k$ . Then, from [25, Lemma 2.2.2] we can conclude that the sequence  $\{\|x^{(k)} - \hat{x}\|_{D_k}^2\}_{k \in \mathbb{N}}$  converges. In particular, since  $D_k \in \mathcal{M}_\mu$ ,  $\{x^{(k)}\}_{k \in \mathbb{N}}$  is bounded and, thus, it has at least one limit point. Let us denote such limit point by  $x^\infty$ . By Corollary 4.1,  $x^\infty$  is stationary; in particular, since  $f$  is convex, it is a minimum point, i.e.  $x^\infty \in X^*$  and, thus,  $\{\|x^{(k)} - x^\infty\|_{D_k}^2\}_{k \in \mathbb{N}}$  converges. Let  $\{x^{(k_i)}\}_{i \in \mathbb{N}}$  be a subsequence of  $\{x^{(k)}\}_{k \in \mathbb{N}}$  which converges to  $x^\infty$ . By the norm inequality (1.5) we can write

$$\|x^{(k_i)} - x^\infty\|_{D_{k_i}}^2 \leq \mu \|x^{(k_i)} - x^\infty\| \xrightarrow{i \rightarrow \infty} 0$$

Since  $\{\|x^{(k)} - x^\infty\|_{D_k}^2\}_{k \in \mathbb{N}}$  converges, this implies that its limit is zero. Invoking again (1.5) we can write

$$\frac{1}{\mu} \|x^{(k)} - x^\infty\|^2 \leq \|x^{(k)} - x^\infty\|_{D_k}^2 \xrightarrow{k \rightarrow \infty} 0$$

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

In the following we present a variation of Theorem 4.3 where the tolerance parameters  $\epsilon_k$  are adaptively chosen, instead of being a prefixed summable sequence.

**THEOREM 4.4** *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 (1.1) is convex and the solution set  $X^*$  of problem (1.1) 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 (4.17), where the sequence  $\{\epsilon_k\}_{k \in \mathbb{N}}$  satisfies

$$\epsilon_k \leq -\tau \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) \tag{4.30}$$

for some  $\tau > 0$ ,



and that hypothesis (H2) of Theorem 4.3 holds. Then, the sequence  $\{x^{(k)}\}_{k \in \mathbb{N}}$  converges to a solution of (1.1).

*Proof.* By substituting (4.30) in (4.29) we obtain

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

The rest of the proof follows exactly from the same arguments employed in Theorem 4.3.  $\square$

We will show in Section 5.4 how the conditions (4.17) and (4.30) can be satisfied in practice.

Assumption (H2) is analogous to the one proposed in [14, 17]. A special case of it consists in the following

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

Thanks to the inequality (1.5), 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  $\lim_{x \rightarrow 0} \sqrt{1 + x}/x = 1/2$ , it follows that  $\sum_{k=0}^{\infty} \xi_k$  and  $\sum_{k=0}^{\infty} \zeta_k$  have the same behaviour. Then, we can conclude that (H2') implies (H2).

We also observe that, employing the same arguments above, we can also 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 [17]).

**4.5.2. Convergence rate analysis.** In this section we analyze the convergence rate of the objective function values  $f(x^{(k)})$  to the optimal one,  $f^*$ , proving that  $f(x^{(k+1)}) - f^* = \mathcal{O}(\frac{1}{k})$ . This complexity result is obtained in the same settings of Theorem 4.4, but further assuming 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 4.3** *Let  $x^{(k)}, \tilde{y}^{(k)} \in \Omega$ . If  $\tilde{y}^{(k)}$  satisfies (4.17), with  $0 < \alpha_k \leq \alpha_{\max}$  and  $D_k \in \mathcal{M}_{\mu}$ , then,*

$$\frac{1}{2\alpha_{\max}\mu} \|\tilde{y}^{(k)} - x^{(k)}\|^2 \leq -\tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) + \epsilon_k. \quad (4.31)$$

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

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

In particular, the previous inequality holds true for  $w = \frac{1}{\alpha_k} D_k (z^{(k)} - \tilde{y}^{(k)})$  (see (4.17)). This results in

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

where the last inequality follows from (1.5).  $\square$

PROPOSITION 4.2 *Let  $\{x^{(k)}\}_{k \in \mathbb{N}}$  be a sequence of points in  $\Omega$  and  $\{d^{(k)}\}_{k \in \mathbb{N}}$  a sequence of descent directions such that  $d^{(k)} = \tilde{y}^{(k)} - x^{(k)}$  and (4.31) holds. Let  $\{\lambda^{(k)}\}_{k \in \mathbb{N}}$  be the steplength sequence computed by Algorithm LS and assume that  $\nabla f_0$  is Lipschitz continuous on  $\Omega$  and that (4.30) holds. Then, there exists  $\lambda_{\min} \in \mathbb{R}_{>0}$  such that*

$$\lambda^{(k)} \geq \lambda_{\min} \quad \forall k \in \mathbb{N}. \quad (4.32)$$

*Proof.* In view of (4.30)–(4.31), setting  $a = \alpha_{\max}\mu$ , one obtains

$$\|d^{(k)}\|^2 \leq -2a(1+\tau)\tilde{h}_{\sigma^{(k)},\gamma}(\tilde{y}^{(k)}, x^{(k)}). \quad (4.33)$$

If  $\nabla f_0$  is Lipschitz continuous on  $\Omega$  with Lipschitz constant  $L$ , then from the descent lemma [6, p.667] we have

$$f_0(x^{(k)} + \lambda d^{(k)}) \leq f_0(x^{(k)}) + \lambda \nabla f_0(x^{(k)})^T d^{(k)} + \frac{L}{2} \lambda^2 \|d^{(k)}\|^2, \quad (4.34)$$

where  $\lambda \in [0, 1]$ . By combining inequalities (4.33) and (4.34) 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 r.h.s. yields

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

The previous inequality ensures that the Armijo condition

$$f(x^{(k)} + \lambda d^{(k)}) \leq f(x^{(k)}) + \lambda \beta \tilde{h}_{\sigma,\gamma}(\tilde{y}^{(k)}, x^{(k)}) \quad (4.35)$$

is satisfied, for all  $k \in \mathbb{N}$ , when  $1 - aL(1+\tau)\lambda \geq \beta$ , that is for all  $\lambda$  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 (4.35), which means  $\lambda^{(k)} > (1-\beta)\delta/(aL(1+\tau))$ . Thus, the steplength sequence  $\{\lambda^{(k)}\}_{k \in \mathbb{N}}$  satisfies inequality (4.32) with  $\lambda_{\min} = (1-\beta)\delta/(aL(1+\tau))$ .  $\square$

Based on these premises, we are now ready to prove the convergence rate result.

THEOREM 4.5 *Assume that the hypotheses of Theorem 4.4 hold and, in addition, that the gradient of  $f_0$  is Lipschitz continuous on  $\Omega$ . Let  $f^*$  be the optimal function value for problem (1.1). Then, we have*

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

*Proof.* If we do not neglect the term  $f(x^{(k)}) - f(\hat{x}) = f(x^{(k)}) - f^*$  in (4.26) and in all the subsequent inequalities, instead of (4.28) we obtain

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

and hence:

$$\begin{aligned}
\|x^{(k+1)} - \hat{x}\|_{D_{k+1}}^2 &\leq (1 + \zeta_k) \|x^{(k+1)} - \hat{x}\|_{D_k}^2 \\
&\leq (1 + \zeta_k) \|x^{(k)} - \hat{x}\|_{D_k}^2 + 2\alpha_k \lambda^{(k)} (1 + \zeta_k) (-\tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) + \epsilon_k) + \\
&\quad - 2\lambda^{(k)} (1 + \zeta_k) \alpha_k (f(x^{(k)}) - f^*) \\
&\stackrel{(4.30)}{\leq} (1 + \zeta_k) \|x^{(k)} - \hat{x}\|_{D_k}^2 - 2\alpha_{\max} (1 + \tau) \zeta \lambda^{(k)} \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k)}, x^{(k)}) + a(f^* - f(x^{(k)})),
\end{aligned}$$

where we set  $\zeta = 1 + \max_k \zeta_k$ ,  $a = 2\lambda_{\min} \alpha_{\min}$ , where  $\lambda_{\min}$  is defined in Proposition 4.2. Summing the previous inequality from 0 to  $k$  gives

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

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 4.4), thus there exists  $M$  such that  $\|x^{(j)} - \hat{x}\|_{D_j}^2 \leq M$ , and from (4.9). Adding the positive quantity  $a(f(x^{(0)}) - f^*)$  to the right hand side of the last inequality we obtain

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

Moreover, exploiting the inequality

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

gives

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

Rearranging terms, this finally yields

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

establishing the result.  $\square$

## 5. Practical computation of $\eta$ - and $\epsilon$ - approximations.

**5.1. Computing  $\eta$ -approximations.** In this section we discuss how to compute a point  $\tilde{y}^{(k)}$  such that (4.20) holds, i.e. satisfying

$$h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)}) \leq \eta h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}), \quad \text{with } y^{(k)} = p(x^{(k)}; h_{\sigma}), \quad (5.1)$$

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  $\Omega$  is the intersection of closed, convex sets and the metric is given by (3.1), is considered in [7]. This is possible when, for each  $k$ , one can compute a sequence  $\{a_l\}_{l \in \mathbb{N}} \subset \mathbb{R}$  such that

$$a_l \leq h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}), \quad \forall l \in \mathbb{N}, \quad \text{and} \quad \lim_{l \rightarrow \infty} a_l = h_{\sigma}(y^{(k)}, x^{(k)}), \quad (5.2)$$

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

$$\lim_{l \rightarrow \infty} h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)}) = h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}). \quad (5.3)$$

In practice,  $l$  should be considered as the index of an inner loop for computing  $\tilde{y}^{(k)}$ . Indeed, when (5.2) holds, we also have

$$\eta a_l \leq \eta h_{\sigma^{(k)}}(y^{(k)}, x^{(k)}) \quad \forall l \in \mathbb{N}. \quad (5.4)$$

Moreover, for all sufficiently large  $l$  we have  $a_l > h_{\sigma^{(k)}}(y^{(k)}, x^{(k)})/\eta$  which, together with (5.4) gives

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

Then, if one considers any method generating a sequence  $\tilde{y}^{(k,l)}$  such that (5.3) holds, the stopping criterion

$$h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)}) \leq \eta a_l \quad (5.5)$$

for the inner iterations is well defined. If  $l$  is the smallest integer such that (5.5) is satisfied, then the point  $\tilde{y}^{(k)} = \tilde{y}^{(k,l)}$  satisfies (5.1). In the following sections we show how to compute a sequence  $a_l$  satisfying (5.2) in an interesting case.

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

$$f_1(x) = g(Ax), \quad (5.6)$$

where  $A \in \mathbb{R}^{m \times n}$  and  $g : \mathbb{R}^m \rightarrow \bar{\mathbb{R}}$  is a convex function. Moreover, we choose  $d_\sigma$  as in (3.1). Let us consider the minimum problem (3.4) 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 Definition 2.4 of the convex conjugate, which gives  $g(Ax) = \max_{v \in \mathbb{R}^m} v^T Ax - g^*(v)$ , obtaining

$$F_{\sigma^{(k)}}(y, v, x^{(k)}) = \frac{1}{2\alpha_k} \|y - z^{(k)}\|_{D_k}^2 + y^T A^T v - g^*(v) - f_1(x^{(k)}) - \frac{\alpha_k}{2} \|\nabla f_0(x^{(k)})\|_{D_k^{-1}}^2 \quad (5.7)$$

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 (5.7), obtaining the explicit expression of the dual function

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

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

$$\max_{v \in \mathbb{R}^m} \Psi_{\sigma^{(k)}}(v, x^{(k)}), \quad (5.8)$$

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 (5.1) can be found by stopping the dual iterations when

$$h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)}) \leq \eta \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)}) \quad (5.9)$$

is satisfied, i.e. (5.5) with  $a_l = \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)})$ .

For example, one can apply a forward-backward method [16], called also ISTA or its accelerated version (FISTA, [5]) to the dual problem. As an alternative, also 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 faced, for example with a primal-dual method such as [12, 24], using (5.9) as stopping condition. More in general, 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

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

stopping the iterates when (5.9) is met.

*Remarks.* We observe that (5.6) includes also the case where  $f_1(x)$  is defined as  $f_1(x) = \sum_{i=1}^r g_i(A_i x)$ , where  $A_i \in \mathbb{R}^{m_i \times n}$ ,  $g_i : \mathbb{R}^{m_i} \rightarrow \mathbb{R}$ . Indeed, formulation (5.6) 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 [35, Theorem 2.3.1 (iv)]).

**5.3. Preserving feasibility.** Clearly, any point  $\tilde{y}^{(k,l)}$  satisfying (5.9), 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  $\Omega$ . Indeed, for any  $l$ ,  $v^{(l)}$  belongs to the domain of the dual function  $\Psi_{\sigma^{(k)}}(\cdot, x^{(k)})$  and, as a consequence, (5.5) implies that  $h_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, x^{(k)})$  is finite. However, the stopping criterion (5.5) 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 to consider also the sequence  $\bar{y}^{(k,l)} = P_{\Omega}(\tilde{y}^{(k,l)})$ , where  $P_{\Omega}$  denotes the Euclidean projection onto the set  $\Omega$ . If, at some inner iteration  $l$ , the inequality

$$\tilde{h}_{\sigma^{(k)}, \bar{y}}(\bar{y}^{(k,l)}, x^{(k)}) \leq \eta \Psi_{\sigma^{(k)}}(v^{(l)}, x^{(k)}) \quad (5.10)$$

is satisfied, this clearly means that  $\bar{y}^{(k,l)} \in P_{\eta}(x^{(k)}; h_{\sigma})$  (i.e., (4.20) is satisfied) and we can set  $\bar{y}^{(k)} = \bar{y}^{(k,l)}$ . We observe that, when  $\tilde{y}^{(k,l)}$  converges to  $y^{(k)}$  as  $l$  diverges, the stopping criterion (5.10) is well defined, since  $\bar{y}^{(k,l)}$  also converges to  $y^{(k)}$ .

**5.4. Computing  $\epsilon$ -approximations.** In this section we show how to compute a point satisfying inclusion (4.17), for any given  $\epsilon_k \in \mathbb{R}_{\geq 0}$ , when the convex function  $f_1$  in (1.1) has the form (5.6). Our arguments are obtained by extending those in [33], which are recovered setting  $D_k = I$ . As done in Section 5.2, we will make use of the duality theory. In particular, we define the primal-dual gap function as

$$\mathcal{G}_{\sigma^{(k)}}(y, v, x^{(k)}) = h_{\sigma^{(k)}}(y, x^{(k)}) - \Psi_{\sigma^{(k)}}(v, x^{(k)}). \quad (5.11)$$

We also have the following result.

**PROPOSITION 5.1** *Let  $\epsilon_k \in \mathbb{R}_{\geq 0}$ . If*

$$\mathcal{G}_{\sigma^{(k)}}(\bar{y}^{(k)}, v, x^{(k)}) \leq \epsilon_k, \quad (5.12)$$

*with  $\bar{y}^{(k)} = z^{(k)} - \alpha_k D_k^{-1} A^T v$ , for some  $v \in \mathbb{R}^m$ , then (4.17) is satisfied.*

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

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

where the last inequality follows from Proposition 2.1. Thus, if (5.12) 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 (2.4), 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 Proposition 2.2, 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  $A^T v = D_k(z^{(k)} - \tilde{y}^{(k)})/\alpha_k$ , (4.17) follows.  $\square$

The previous result suggests that for computing  $\tilde{y}^{(k)}$  satisfying the assumptions of Corollary 4.1 we can use the same iterative approaches described at the end of Section 5.1, stopping the iterates when

$$\mathcal{G}_{\sigma^{(k)}}(\tilde{y}^{(k,l)}, v^{(l)}, x^{(k)}) \leq \epsilon_k \quad \text{and} \quad \tilde{h}_{\sigma^{(k)}, \gamma}(\tilde{y}^{(k,l)}, x^{(k)}) < 0. \quad (5.13)$$

**5.5. Equivalence between  $\eta$  and  $\epsilon$  approximations.** Any  $\eta$ -approximation  $\tilde{y}^{(k)}$  satisfying (5.9) for some  $v^{(l)} \in \mathbb{R}^m$  is also an  $\epsilon$ -approximation, where  $\epsilon = -\tau h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)})$  and  $\tau = -1 + 1/\eta$ . In fact, in these settings, (5.9) 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 5.4, this means that  $\tilde{y}^{(k)}$  is an  $\epsilon$ -approximation with  $\epsilon = -\tau h_{\sigma^{(k)}}(\tilde{y}^{(k)}, x^{(k)})$ . Thus, any point computed by an iterative procedure stopped when (5.9) is satisfied, is both an  $\eta$ - and  $\epsilon$ -approximation.

**6. Numerical illustration.** In order to validate the proposed approach, we consider a relevant image restoration problem, whose variational formulation consists in minimizing the sum of a discrepancy functional plus a regularization term. 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

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

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  $bg$ , the data discrepancy is defined as

$$f_0(x) = KL(Hx + bg\mathbf{1}, b),$$

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

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

| problem   | ref. image          | size    | range       | $\sigma_{\text{psf}}$ | $bg$ | $\rho$ |
|-----------|---------------------|---------|-------------|-----------------------|------|--------|
| cameraman | Matlab cameraman    | $256^2$ | $[0, 1000]$ | 1.4                   | 5    | 0.0091 |
| micro     | [34, 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  |

TABLE 6.1  
Test problems description

where  $\rho \in \mathbb{R}_{>0}$  is a regularization parameter multiplying the total variation functional [30] 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 (5.6), with  $A = (\nabla_1^T \ \dots \ \nabla_n^T \ I)^T \in \mathbb{R}^{3n \times n}$ . In this case  $v \in \mathbb{R}^{3n}$  and  $g^*$  is the indicator function of the set  $B_{0,\rho}^2 \times \dots \times B_{0,\rho}^2 \times \mathbb{R}_{\leq 0}^n$ , where  $B_{0,\rho}^2 \subset \mathbb{R}^2$  is the 2-dimensional Euclidean ball centered in 0 with radius  $\rho$ .

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 [22].

We define a set of test problems in the following way: a reference image has been rescaled so that the pixel values lie in a specified range (this is for simulating different noise levels), then it has been blurred by convolution with a Gaussian kernel with standard deviation  $\sigma_{\text{psf}}$  and the background has been added. Finally, Poisson noise has been simulated with the Matlab `imnoise` function, obtaining the noisy blurred image  $b$ . The details of each test problem are listed in Table 6.1. The regularization parameter  $\rho$  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. We implement our inexact algorithm, which is summarized in Algorithm VMILA,

---

**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  $k = 0, 1, 2, \dots$

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 (5.9) 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  $x^{(k+1)} = x^{(k)} + \lambda^{(k)} d^{(k)}$ .
- 

in Matlab environment with the following settings:

*Step 1, metric selection:* the scaling matrix  $D_k$  is chosen mimicking the split-gradient idea [23]. In particular, at each outer iteration it is defined as the diagonal matrix with positive entries as follows

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

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 [8].

*Step 1, steplength selection:* the parameter  $\alpha_k$  is chosen by the same strategy used e.g. in [10, 28, 27], 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 experienced different inner solvers applied on the primal-dual or on the dual formulation of the inner problem. The best performances have been obtained choosing FISTA applied to the dual problem (5.8), in the variant proposed in [11] 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  $a = 2.1$  in [11, formula (5)]. For brevity, in the following, we report only the results obtained stopping the inner iterates when criterion



(5.10) is met, which corresponds to both an  $\eta$  and  $\epsilon$  approximation (see section 5.5). A maximum number of 1500 inner iterations is also 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 parameters setting:* the line-search parameters  $\delta, \beta, \gamma$  have been set respectively equal to  $0.5, 10^{-4}, 1$ .

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 Version 7 (R2010b).

We investigate first the impact of the inexactness parameter  $\eta$  choice on the overall method. In Figure 6.1 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 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 further benchmark, we compare our algorithm to a well established state-of-the-art method, the Chambolle and Pock's method (CP) [12], which, referring to the notations used in their paper, has been implemented setting  $G(x) = \iota_{\mathbb{R}_{\geq 0}^n}(x)$  and  $F(Kx) = KL(Hx + bg, b) + \beta \sum_{i=1}^n \|\nabla_i x\|$ , with  $K = (H^T, \nabla_1^T, \dots, \nabla_n^T)^T$ . In this way the resolvent operator associated to  $F^*$  can be computed in closed form. In Figure 6.2, we compare the behaviour of our approach (with  $\eta = 10^{-6}$ ) with CP (2000 iterations) for different choices of its two parameters,  $\sigma$  and  $\tau$  (once  $\tau$  is selected,  $\sigma$  is chosen such that  $\tau\sigma L^2 = 1$ , where  $L = \|K\|$ ). We can observe that CP is quite sensitive to these parameters, and it is difficult to devise, in general, the more convenient choice, while our approach with the parameters 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.

**7. 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 relies in the possibility of using an approximation of the proximal operator, while the underlying metric may change at each iterations 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.

- The convergence is ensured by a line-search procedure, which does not depend on any user supplied parameter (actually the constants  $\gamma, \beta, \delta$  have to be chosen, but the behaviour of the whole algorithm is not sensitive to these choices). On the other side, the “free” parameter  $\sigma$  in (3.4) 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 a large scale convex problems shows that the performances of the inexact method are promising and comparable with those of a state-of-the-art method.

Future work will be addressed especially to deepen the theoretical and numerical analysis in the nonconvex case, investigating the possibility to obtain convergence results stronger than the ones stated in Theorems 4.1 and 4.2, at least for some classes of nonconvex functions (e.g. Kurdyka-Łojasiewicz functions).

## REFERENCES

- [1] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forwardbackward splitting, and regularized GaussSeidel methods*, Math. Program., 137 (2013), pp. 91–129.



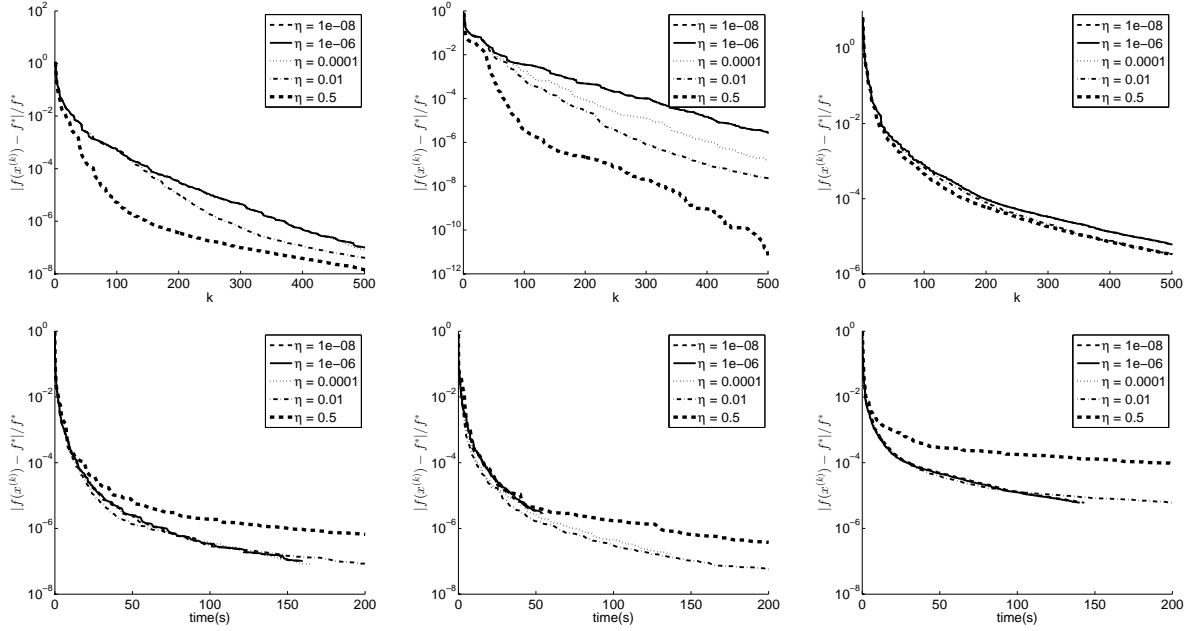


FIG. 6.1. Algorithm VMILA with different choices for  $\eta$ . Relative decrease of the objective function values with respect to the outer iteration number (top row) and to the computational time (bottom row). Left column: cameraman. Middle column: micro. Right column: phantom.

- [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] ———, *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, 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, G. LANDI, E. LOLI PICCOLOMINI, AND L. ZANNI, *Scaling techniques for gradient projection-type methods in astronomical image deblurring*, Int. J. Comput. Math., 90 (2013), pp. 9–29.
- [9] S. BONETTINI AND M. PRATO, *New convergence results for the scaled gradient projection method*, submitted, available on <http://arxiv.org/abs/1406.6601> (2015).
- [10] S. BONETTINI, R. ZANELLA, AND L. ZANNI, *A scaled gradient projection method for constrained image deblurring*, Inverse Probl., 25 (2009), p. 015002.
- [11] A. CHAMBOLE AND C. DOSSAL, *On the convergence of the iterates of “FISTA”*. <hal-01060130v3>, Sept. 2014.
- [12] A. CHAMBOLE 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.
- [13] 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.
- [14] P.L. COMBETTES AND B.C. VŨ, *Variable metric forward-backward splitting with applications to monotone inclusions in duality*, Optimization, 63 (2014), pp. 1289–1318.
- [15] P.L. COMBETTES AND V. R. WAJS, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
- [16] 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 Optimization and Its Applications, Springer, New York, NY, 2011, pp. 185–212.
- [17] P. L. COMBETTES AND B. C. VŨ, *Variable metric quasi-Féjer monotonicity*, Nonlinear Anal.-Theor., 78 (2013), pp. 17–31.
- [18] J. DUCHI, E. HAZAN, AND Y. SINGER, *Adaptive subgradient methods for online learning and stochastic optimization*, Journal of Machine Learning Research, 12 (2011), pp. 2121–2159.

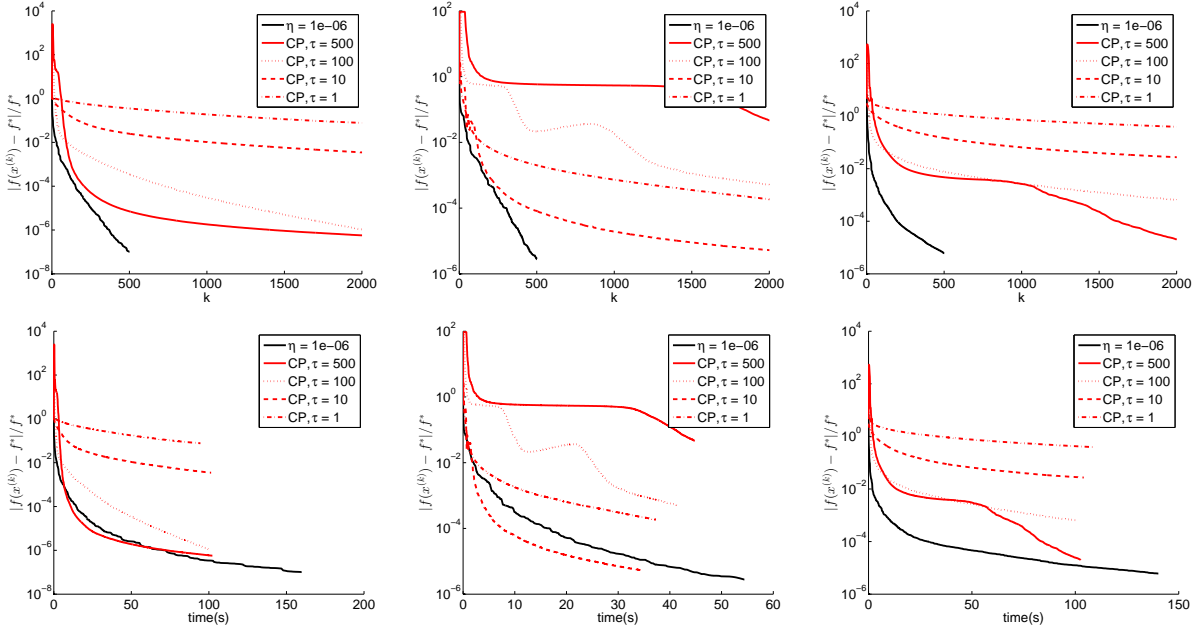


FIG. 6.2. Comparison between Algorithm VMILA ( $\eta = 10^{-6}$ ) and the CP algorithm with different choices of its parameters. Relative decrease of the objective function values with respect to the outer iteration number (top row) and to the computational time (bottom row). Left column: cameraman. Middle column: micro. Right column: phantom.

- [19] J. ECKSTEIN, *Nonlinear proximal point algorithms using bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.
- [20] P. FRANKEL, G. GARRIGOS, AND J. PEYPOUQUET, *Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates*, J. Opt. Theory Appl., (to appear). DOI: 10.1007/s10957-014-0642-3.
- [21] 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.
- [22] P. C. HANSEN, J. G. NAGY, AND D. P. O’LEARY, *Deblurring Images: Matrices, Spectra and Filtering*, SIAM, Philadelphia, 2006.
- [23] H. LANTÉRI, M. ROCHE, AND C. AIME, *Penalized maximum likelihood image restoration with positivity constraints: multiplicative algorithms*, Inverse Probl., 18 (2002), pp. 1397–1419.
- [24] I. LORIS AND C. VERHOEVEN, *On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty*, Inverse Probl., 27 (2011), p. 125007.
- [25] B. POLYAK, *Introduction to optimization*, Optimization Software - Inc., Publication Division, New York, 1987.
- [26] F. PORTA AND I. LORIS, *On some steplength approaches for proximal algorithms*, Appl. Math. Comput., 253 (2015), pp. 345–362.
- [27] M. PRATO, A. LA CAMERA, S. BONETTINI, AND M. BERTERO, *A convergent blind deconvolution method for post-adaptive-optics astronomical imaging*, Inverse Probl., 29 (2013), p. 065017.
- [28] 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), p. A133.
- [29] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [30] L.I. RUDIN, S. OSHER, AND E. FATEMI, *Nonlinear total variation based noise removal algorithms*, J. Phys. D., 60 (1992), pp. 259–268.
- [31] S. SALZO AND S. VILLA, *Inexact and accelerated proximal point algorithms*, J. Convex Anal., 19 (2012), pp. 1167–1192.
- [32] P. TSENG AND S. YUN, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.
- [33] S. VILLA, S. SALZO, L. BALDASSARRE, AND A. VERRI, *Accelerated and inexact forward-backward algorithms*, SIAM J. Optim., 23 (2013), pp. 1607–1633.
- [34] 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.
- [35] A. ZALINESCU, *Convex analysis in general vector spaces*, World Scientific Publishing Co. Inc., River Edge, NJ, 2002.