Barzilai-Borwein-like rules in proximal gradient schemes for ℓ_1-regularized problems

Serena Crisci[a,1,*], Simone Rebegoldi[b,1], Gerardo Toraldo[a,1], Marco Viola[a,1]

[a] *Department of Mathematics and Physics, University of Campania "L. Vanvitelli",*
Viale A. Lincoln 5, Caserta, Italy
[b] *Department of Industrial Engineering, University of Florence, Viale Morgagni 40, Florence, Italy*
[c] *School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland*

Abstract

We propose a novel steplength selection rule in proximal gradient methods for minimizing the sum of a differentiable function plus an ℓ_1-norm penalty term. The proposed rule modifies one of the classical Barzilai-Borwein steplength, extending analogous results obtained in the context of gradient projection methods for constrained optimization. We analyze the spectral properties of the Barzilai-Borwein-like steplength when the differentiable part is quadratic, showing that its reciprocal lies in the spectrum of the submatrix of the Hessian matrix that depends on both the nonzero and the nonoptimal zero components of the current iterate, allowing for acceleration effects when the optimal zero components start to be identified. Furthermore, we insert the modified rule into a proximal gradient method with a nonmonotone line search, for which we prove global convergence towards a stationary point. Numerical experiments show the ability of the proposed rule to sweep the spectrum of the reduced Hessian on a series of quadratic ℓ_1-regularized problems, as well as its effectiveness in recovering the ground truth in a least squares regularized problem arising in image restoration.

1. Introduction

In this paper, we are interested in solving the following minimization problem

$$\min_{x\in\mathbb{R}^n} f_0(x) + \lambda\|x\|_1, \tag{1}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $\|x\|_1 = \sum_{i=1}^n |x_i|$ is the ℓ_1-norm of the vector $x \in \mathbb{R}^n$, and $\lambda > 0$ is a regularization parameter. Problems of the form (1) are widespread in several domains of applied science, including image reconstruction [2, 22], machine learning [11], and portfolio selection [12, 15], just to mention a few. In such applications, the discrepancy function f_0 describes how close the model fits the observed data, whereas the ℓ_1-norm is used to foster the sparsity of the optimal solution.

Problem (1) belongs to the more general class of composite convex problems

$$\min_{x\in\mathbb{R}^n} f(x) \equiv f_0(x) + f_1(x), \tag{2}$$

where $f_0 : \Omega \to \mathbb{R}$ is continuously differentiable on an open set Ω containing $\overline{\mathrm{dom}(f_1)}$, and $f_1 : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ is proper, convex, and lower semicontinuous. A standard approach to address problem (2) is the proximal gradient (or forward backward) method [2, 8, 13, 14], which typically alternates a

[*] Corresponding author

Email addresses: serena.crisci@unicampania.it (Serena Crisci), simone.rebegoldi@unifi.it (Simone Rebegoldi), gerardo.toraldo@unicampania.it (Gerardo Toraldo), marco.viola@ucd.ie (Marco Viola)

gradient step on f_0 with a proximal step on f_1. One popular strategy consists of combining the proximal gradient method with a line search procedure [6, 7, 13, 40, 49], leading to the following general iteration

$$\begin{aligned} y^{(k)} &= \operatorname{prox}_{\alpha_k f_1}\left(x^{(k)} - \alpha_k \nabla f_0(x^{(k)})\right) && (3)\\ x^{(k+1)} &= x^{(k)} + \nu_k(y^{(k)} - x^{(k)}), \qquad k = 0, 1, \dots && (4) \end{aligned}$$

where $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ is the steplength parameter, with $0 < \alpha_{\min} \leq \alpha_{\max}$, $\operatorname{prox}_{\alpha_k f_1} : \mathbb{R}^n \to \mathbb{R}^n$ is the proximal operator of $\alpha_k f_1$, i.e.,

$$\operatorname{prox}_{\alpha_k f_1}(x) = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\|z - x\|^2 + \alpha_k f_1(z),$$

and $\nu_k \in (0, 1]$ is the line search parameter. Convergence for this class of methods usually relies on appropriate choices of the parameters α_k, ν_k, which can be practically computed by means of backtracking procedures enforcing some sufficient decrease condition for the f_0 term or the objective function f, respectively.

The selection of the steplength α_k is of crucial importance for improving the practical convergence behavior of method (3)-(4). If the function f_1 is identically zero, and thus problem (2) is differentiable and unconstrained, the steplength α_k can be computed by means of the well-known Barzilai-Borwein (BB) rules, originally developed in the seminal paper [1]. The BB rules capture second order information of the objective function f in a quasi-Newton fashion, by imposing the following secant conditions

$$\alpha_k^{\mathrm{BB1}} = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \|\alpha^{-1} s^{(k-1)} - z^{(k-1)}\|, \qquad \alpha_k^{\mathrm{BB2}} = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \|s^{(k-1)} - \alpha z^{(k-1)}\|, \tag{5}$$

where $s^{(k-1)} = x^{(k)} - x^{(k-1)}$ and $z^{(k-1)} = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$. From (5), the following two alternative steplength rules descend

$$\begin{aligned} \alpha_k^{\mathrm{BB1}} &= \frac{{s^{(k-1)}}^T s^{(k-1)}}{{s^{(k-1)}}^T z^{(k-1)}}, && (6)\\ \alpha_k^{\mathrm{BB2}} &= \frac{{s^{(k-1)}}^T z^{(k-1)}}{{z^{(k-1)}}^T z^{(k-1)}}. && (7) \end{aligned}$$

The special advantage of BB rules lies in their ability of sweeping the spectrum of the inverse of the Hessian matrix $\nabla^2 f(x^{(k)})$ [30, 45]. This spectral property was first deduced in the strictly convex quadratic case, where the BB steplengths can be written as reciprocals of Rayleigh quotients of the Hessian, and are proven to satisfy the following inequality [31, 47]

$$\frac{1}{\lambda_{\max}(A)} \leq \alpha_k^{\mathrm{BB2}} \leq \alpha_k^{\mathrm{BB1}} \leq \frac{1}{\lambda_{\min}(A)}, \tag{8}$$

with $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denoting the minimum and the maximum eigenvalues of the Hessian matrix A, respectively. Assuming that ${s^{(k-1)}}^T z^{(k-1)} > 0$, the inequality (8) still holds for non-quadratic objective function; in this case, the inverse of BB rules can be interpreted as Rayleigh quotients related to the average of the Hessian matrix along the segment $s^{(k-1)}$. In view of (8), adaptive strategies can emphasize the practical effectiveness of the BB rules in gradient methods, by means of ad-hoc switching criteria between BB1 and BB2 [54, 32]. The idea of exploiting the alternation of small and large steplengths to promote a more suitable sweeping of the Hessian spectrum along the iterative procedure, has been adopted in several spectral-based steplength selections, resulting in significant improvements of the practical convergence rate of gradient methods, for both quadratic and non quadratic applications [20, 52, 24, 23, 33, 38].

The BB rules and the related alternation strategies have been employed for constrained optimization as well [4, 10, 21, 47, 28, 5, 44], despite the fact that they were conceived for unconstrained problems, and the lack of a related spectral analysis. In the constrained setting, the function f_1 in (2) is the indicator function of a closed convex constraint set, and the scheme (3)-(4) reduces to the Gradient Projection (GP) method. Recent works [19, 16, 17] have shown that the spectral theory outlined above requires suitable adaptations for specific constrained problems, in order to take advantage of the presence of the constraints, exploiting first-order optimality conditions in a proper manner. In particular, in [19], the authors focused on GP methods of the form (3)-(4) applied to differentiable minimization problems subject to box constraints, showing that the sequence of steplengths provided by the BB1 rule is able to sweep the spectrum of the so-called *reduced Hessian*, which is the sub-matrix of the Hessian that depends on the inactive constraints at the current iterate. This is coherent with the Karush-Kuhn-Tucker optimality conditions, which require only the gradient components of the inactive constraints at the solution to be zero. As a consequence, a steplength rule that is able to neglect the gradient information that depends on the active constraints can accelerate GP methods, and this effect becomes more and more remarkable as the active set is stabilizing. Based on these observations, suitable modifications to the BB2 rule were introduced in [19, 16, 17] to exploit the presence of the constraints, which led to novel BB2-like rules for box-constrained problems [19], and singly linearly constrained problems subject to lower and upper bounds [16], respectively. The novel steplength rules have been shown to outperform the standard BB2 rule in practical implementations, leading to a faster reduction of the gradient components that must be null at the solution, especially when employed within alternating strategies.

Interestingly, the efficiency of the BB rules has been observed also for regularized problems of the form (2), where f_1 is a convex penalty term. More precisely, the BB rules (or their alternated implementations) have been successfully applied to the ℓ_1-regularized problem (1), see e.g. [39, 48, 50, 51], and problems where f_1 is the composition of the ℓ_1-norm with some linear operator, as is the case with Total Variation based image deblurring [6, 7, 9, 36]. However, to the best of our knowledge, an analysis of the spectral properties of the BB rules for the regularized problem (2) has yet to be proposed in the literature.

In this paper, we analyze the spectral properties of the BB rules when they are adopted within a proximal gradient method with line search applied to the ℓ_1-regularized problem (1). Our work is inspired by the spectral analysis carried out for constrained problems [19, 16], as we adapt its arguments to the presence of the ℓ_1-penalty term. Under the assumption that the differentiable term f_0 is quadratic, we prove that the BB1 rule possesses the property of automatically discarding the second-order information related to the zero components of the current iterate that satisfy the optimality condition. Vice versa, we show that the standard BB2 rule may not well approximate the second-order information of f_0, due to the presence of an error term. Based on this observation, we design a new BB2 rule that mimic the natural behaviour of BB1. From the numerical viewpoint, we show that the proposed modification can foster the acceleration of the proximal gradient method on a series of quadratic ℓ_1-regularized problems, enabling a faster fulfillment of the optimality conditions. Additionally, we demonstrate the efficiency of the proposed rule on a least squares problem regularized with the Total Variation function and the ℓ_1-penalty term, which arises from image deblurring in presence of Gaussian noise.

The paper is organized as follows. In Section 2, we report some preliminary notions of subdifferential calculus. In Section 3, we provide the spectral analysis of the proposed BB-like rule for ℓ_1-regularized problems. In Section 4, we propose and analyze a proximal gradient method with a nonmonotone Armijo-like line search, whose steplength can be computed by means of our BB-like rule. Numerical experiments of the proposed proximal gradient method are reported in Section 5. Our final remarks and future work are given in Section 6.

2. Preliminaries

In the following, we denote with $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ the extended real numbers set. The symbol $\|\cdot\|$ refers to the standard Euclidean norm on $\mathbb{R}^n$. The domain of a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is the set $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : \ f(x) < +\infty\}$, and f is called proper if $\mathrm{dom}(f) \neq \emptyset$ and f is finite on $\mathrm{dom}(f)$. The notation $\bigtimes_{i=1}^{n} \Omega_i$ stands for the Cartesian product of the n sets $\Omega_1, \dots, \Omega_n$. Finally, the indicator function of a set $\Omega \neq \emptyset$ is given by

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

We start with the definition of subdifferential for a general (possibly non convex) function.

Definition 2.1. *The* Fréchet subdifferential *of a proper, lower semicontinuous function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ at the point $x \in \mathrm{dom}(f)$ is given by [46, Definition 8.3(a)]*

$$\hat{\partial} f(x) = \left\{ w \in \mathbb{R}^n : \liminf_{z \to x, z \neq x} \frac{f(z) - f(x) - (z-x)^T w}{\|z - x\|} \geq 0 \right\}.$$

The limiting subdifferential *of f at x is the set [46, Definition 8.3(b)]*

$$\partial f(x) = \{ w \in \mathbb{R}^n : \exists\, x^{(k)} \to x,\ f(x^{(k)}) \to f(x),\ w^{(k)} \in \hat{\partial} f(x^{(k)}) \to w \text{ as } k \to \infty \}.$$

Below, we recall that the limiting subdifferential reduces to the usual Fenchel subdifferential when f is a convex function.

Lemma 2.1. *[46, Proposition 8.12] Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper, convex function and $x \in \mathrm{dom}(f)$. Then we have*

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

In the following lemma, we report a useful subdifferential calculus rule holding for the sum of two functions, one of which is continuously differentiable.

Lemma 2.2. *[46, Exercise 8.8] Suppose that the function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ can be written as $f = f_0 + f_1$, where $f_1 : \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper, and $f_0 : \Omega_0 \to \bar{\mathbb{R}}$ is continuously differentiable on an open set $\Omega_0 \supseteq \overline{\mathrm{dom}(f_1)}$. Then we have*

$$\partial f(x) = \{\nabla f_0(x)\} + \partial f_1(x), \quad \forall\, x \in \mathrm{dom}(f_1).$$

We now introduce the definition of stationary point.

Definition 2.2. *Given a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, a point $x \in \mathbb{R}^n$ is stationary for f if $x \in \mathrm{dom}(f)$ and $0 \in \partial f(x)$.*

Remark 2.1. *If $x \in \mathbb{R}^n$ is a local minimum point, then x is a stationary point; if f is convex, then $x \in \mathbb{R}^n$ is a (global) minimum point if and only x is stationary. The latter remark is a straightforward consequence of Lemma 2.1.*

Definition 2.3. *[42] The proximal operator $\mathrm{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ associated to a proper, convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is defined as*

$$\mathrm{prox}_f(x) = \operatorname*{argmin}_{z \in \mathbb{R}^n} \frac{1}{2}\|z - x\|^2 + f(z), \quad \forall\, x \in \mathbb{R}^n.$$

We conclude the section by reporting a nice result holding when a function f is given by a separable sum of convex functions.

Lemma 2.3. *Suppose that the function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is given by*

$$f(x) = \sum_{i=1}^{r} f_i(x_i),$$

where $f_i : \mathbb{R}^{n_i} \to \bar{\mathbb{R}}$ is proper, lower, and semicontinuous for all $i = 1, \ldots, r$ and $\sum_{i=1}^{r} n_i = n$.

(i) The subdifferential of f can be written as

$$\partial f(x) = \bigtimes_{i=1}^{r} \partial f_i(x_i) = (\partial f_1(x_1), \dots, \partial f_r(x_r)), \quad \forall\, x \in \mathrm{dom}(f).$$

(ii) The proximal operator of f is given by

$$\mathrm{prox}_f(x) = \bigtimes_{i=1}^{n} \mathrm{prox}_{f_i}(x_i) = (\mathrm{prox}_{f_1}(x_1), \dots, \mathrm{prox}_{f_r}(x_r)), \quad \forall\, x \in \mathbb{R}^n.$$

Proof. Item (i) is proved in [53, Corollary 2.4.5]. Regarding Item (ii), we can combine Definition 2.3, Remark 2.1, and Lemma 2.2, to obtain the following equivalence

$$y = \mathrm{prox}_f(x) \quad \Leftrightarrow \quad x - y \in \partial f(y).$$

Then, item (ii) follows by applying item (i) to the above equivalence. □

3. A Barzilai-Borwein-like rule for ℓ_1−regularized problems

In this section, we analyse the behavior of the BB rules (6)-(7) within the proximal gradient method (3)-(4) applied to the ℓ_1−regularized problem (1), and propose a novel BB2-like rule that takes into account the optimality conditions of the problem. The spectral analysis and the proposed steplength rule are first given under the assumption that f_0 is quadratic, and then extended to the general non quadratic case.

3.1. The quadratic case

We turn our attention to a special instance of the ℓ_1−regularized problem (1), that is

$$\min_{x \in \mathbb{R}^n} f(x) \equiv \underbrace{\frac{1}{2} x^T A x - b^T x + c}_{:= f_0(x)} + \lambda \|x\|_1, \tag{9}$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$, and $\lambda > 0$ is the regularization parameter.

Since the objective function f is convex, it follows by Remark 2.1 and Lemma 2.2 that a solution of problem (9) is identified by the following subdifferential inclusion

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^n} f_0(x) + f_1(x) \quad \Leftrightarrow \quad -\nabla f_0(x^*) \in \partial(\lambda \| \cdot \|_1)(x^*), \tag{10}$$

where the subdifferential of the ℓ_1−norm can be computed by Lemma 2.3(i) as $\partial(\lambda\|\cdot\|_1)(x) = \bigtimes_{i=1}^{n} \partial(\lambda|\cdot|)(x_i)$ with

$$\partial(\lambda|\cdot|)(x_i) = \begin{cases} \lambda\, \mathrm{sign}(x_i), & \text{if } x_i \neq 0, \\ [-\lambda, \lambda], & \text{if } x_i = 0. \end{cases} \tag{11}$$

Therefore, the optimality condition (10) can be reformulated component-wise as

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^n} f_0(x) + f_1(x) \quad \Leftrightarrow \quad -(\nabla f_0(x^*))_i \in \partial(\lambda|\cdot|)(x_i^*), \quad i = 1, \dots, n. \tag{12}$$

Let us assume to apply the line search based proximal gradient scheme defined in (3) with $f_1(x) = \lambda\|x\|_1$ to solve problem (9). Due to the separable structure of the ℓ_1−norm and Lemma 2.3(ii), the proximal operator of f_1 can be decomposed in the following manner

$$\mathrm{prox}_{\alpha_{k-1} f_1}(x) = \mathrm{prox}_{\alpha_{k-1}\lambda\|\cdot\|_1}(x) = \left(\mathrm{prox}_{\alpha_{k-1}\lambda|\cdot|}(x_1), \dots, \mathrm{prox}_{\alpha_{k-1}\lambda|\cdot|}(x_n)\right),$$

where $\mathrm{prox}_{\alpha_{k-1}\lambda|\cdot|}$ denotes the classical *soft-thresholding operator* of parameter $\alpha_{k-1}\lambda$, i.e.,

$$\mathrm{prox}_{\alpha_{k-1}\lambda|\cdot|}(x_i) = \begin{cases} x_i - \alpha_{k-1}\lambda, & \text{if } x_i > \alpha_{k-1}\lambda, \\ 0, & \text{if } x_i \in [-\alpha_{k-1}\lambda, \alpha_{k-1}\lambda], \\ x_i + \alpha_{k-1}\lambda, & \text{if } x_i < -\alpha_{k-1}\lambda. \end{cases} \tag{13}$$

Then, the $k-$th iteration of the proximal gradient method (3)-(4) applied to problem (9) can be written component-wise as

$$y_i^{(k-1)} = \mathrm{prox}_{\alpha_{k-1}\lambda|\cdot|}\left(x_i^{(k-1)} - \alpha_{k-1} g_i^{(k-1)}\right) \tag{14}$$

$$x_i^{(k)} = x_i^{(k-1)} + \nu_{k-1}(y_i^{(k-1)} - x_i^{(k-1)}), \qquad i = 1, \ldots, n, \tag{15}$$

where we use the notation $g^{(k-1)} = (g_1^{(k-1)} \ \cdots \ g_n^{(k-1)})^T = \nabla f_0(x^{(k-1)})$ for all $k \geq 1$.

We now proceed with the spectral analysis of the BB rules when employed inside method (14)-(15). To this aim, by analogy with the corresponding definitions derived for box-constrained problems in [19], we consider the following partition of the indices set

$$\mathcal{J}_{k-1} = \left\{ i : \ x_i^{(k-1)} = 0 \ \wedge \ -g_i^{(k-1)} \in \partial(\lambda|\cdot|)(x_i^{(k-1)}) \right\} = \left\{ i : \ x_i^{(k-1)} = 0 \ \wedge \ -g_i^{(k-1)} \in [-\lambda, \lambda] \right\}$$

$$\mathcal{I}_{k-1} = \{1, \ldots, n\} \setminus \mathcal{J}_{k-1}.$$

The set $\mathcal{J}_{k-1}$ contains the indexes corresponding to the zero components of the iterate $x^{(k-1)}$ that satisfy the scalar optimality condition (12), namely, the *active* components of $x^{(k-1)}$ that are optimal. If $i \in \mathcal{J}_{k-1}$, $x_i^{(k-1)} - \alpha_{k-1} g_i^{(k-1)} \in [-\alpha_{k-1}\lambda, \alpha_{k-1}\lambda]$, and from (13) and (14) we have $y_i^{(k-1)} = \mathrm{prox}_{\alpha_{k-1}\lambda|\cdot|}(x_i^{(k-1)} - \alpha_{k-1} g_i^{(k-1)}) = 0$, which implies

$$x_i^{(k)} = x_i^{(k-1)} + \nu_{k-1}(y_i^{(k-1)} - x_i^{(k-1)}) = x_i^{(k-1)}, \quad i \in \mathcal{J}_{k-1}.$$

Hence, we can split the components of the $k-$th iterate as follows

$$x_i^{(k)} = \begin{cases} 0 & \text{if } i \in \mathcal{J}_{k-1} \\ x_i^{(k-1)} - \nu_{k-1}\alpha_{k-1} g_i^{(k-1)} - \nu_{k-1}\alpha_{k-1} v_i^{(k)}, & \text{if } i \in \mathcal{I}_{k-1} \end{cases} \tag{16}$$

where $v_i^{(k)} \in \partial(\lambda|\cdot|)(x_i^{(k)})$ takes the values

$$v_i^{(k)} = \begin{cases} \lambda, & \text{if } x_i^{(k-1)} - \alpha_{k-1} g_i^{(k-1)} > \alpha_{k-1}\lambda, \\ -\lambda, & \text{if } x_i^{(k-1)} - \alpha_{k-1} g_i^{(k-1)} < -\alpha_{k-1}\lambda. \end{cases} \tag{17}$$

Consequently, the vector $s^{(k-1)} = x^{(k)} - x^{(k-1)}$ can be partitioned as:

$$s^{(k-1)} = \begin{bmatrix} s_{\mathcal{J}_{k-1}}^{(k-1)} \\ s_{\mathcal{I}_{k-1}}^{(k-1)} \end{bmatrix} = \begin{bmatrix} 0 \\ -\nu_{k-1}\alpha_{k-1}(g_i^{(k-1)} + v_i^{(k)}) \end{bmatrix}, \tag{18}$$

where, without loss of generality, we assume that the first components of $s^{(k-1)}$ are indexed in $\mathcal{J}_{k-1}$ and the last ones are indexed in $\mathcal{I}_{k-1}$. Analogously, each component of the vector $z^{(k-1)} = g^{(k)} - g^{(k-1)}$

expressing the difference between the corresponding gradients becomes

$$\begin{aligned}
z_i^{(k-1)} &= g_i^{(k)} - g_i^{(k-1)} = \\
&= \sum_{j=1}^{n} a_{ij} x_j^{(k)} - \sum_{j=1}^{n} a_{ij} x_j^{(k-1)} \\
&= \sum_{j\in\mathcal{I}_{k-1}} a_{ij}\left(x_j^{(k-1)} - \nu_{k-1}\alpha_{k-1} g_j^{(k)} - \nu_{k-1}\alpha_{k-1} v_j^{(k)}\right) - \sum_{j\in\mathcal{I}_{k-1}} a_{ij} x_j^{(k-1)} \\
&= \sum_{j\in\mathcal{I}_{k-1}} a_{ij}\left(-\nu_{k-1}\alpha_{k-1}(g_j^{(k)} + v_j^{(k)})\right) \\
&= \sum_{j\in\mathcal{I}_{k-1}} a_{ij} s_j^{(k-1)} \qquad (19)
\end{aligned}$$

,

and we may write

$$z^{(k-1)} = \begin{bmatrix} z_{\mathcal{J}_{k-1}}^{(k-1)} \\ z_{\mathcal{I}_{k-1}}^{(k-1)} \end{bmatrix} = \begin{bmatrix} A_{\mathcal{J}_{k-1}\,\mathcal{I}_{k-1}} s_{\mathcal{I}_{k-1}}^{(k-1)} \\ A_{\mathcal{I}_{k-1}\,\mathcal{I}_{k-1}} s_{\mathcal{I}_{k-1}}^{(k-1)} \end{bmatrix}. \qquad (20)$$

When $\mathcal{I}_k = \mathcal{I}_{k-1}$, i.e., when the active optimal components are stabilized from one iteration to the other, the previous equation yields the following recurrence formula

$$g_{\mathcal{I}_k}^{(k+1)} = g_{\mathcal{I}_k}^{(k)} - \nu_k\alpha_k A_{\mathcal{I}_k\,\mathcal{I}_k} g_{\mathcal{I}_k}^{(k)} - \nu_k\alpha_k A_{\mathcal{I}_k\,\mathcal{I}_k} v_{\mathcal{I}_k}^{(k+1)}. \qquad (21)$$

Then, letting $(u_1,\dots,u_r)$ denote a basis of orthonormal eigenvectors for $A_{\mathcal{I}_k\,\mathcal{I}_k}$ with associated eigenvalues $(\gamma_1,\dots,\gamma_r)$, where $r = \sharp\mathcal{I}_k$, we can write $g_{\mathcal{I}_k}^{(k+1)} = \sum_{i=1}^r \mu_i^{(k+1)} u_i$, $g_{\mathcal{I}_k}^{(k)} = \sum_{i=1}^r \mu_i^{(k)} u_i$ and $v_{\mathcal{I}_k}^{(k+1)} = \sum_{i=1}^r \zeta_i^{(k+1)} u_i$ as unique representations of $g_{\mathcal{I}_k}^{(k+1)}$, $g_{\mathcal{I}_k}^{(k)}$, and $v_{\mathcal{I}_k}^{(k+1)}$ with respect to the basis $(u_1,\dots,u_r)$. In this way, equation (21) allows us to obtain the following relation

$$\mu_i^{(k+1)} = (1-\nu_k\alpha_k\gamma_i)\mu_i^{(k)} - \nu_k\alpha_k\gamma_i\zeta_i^{(k+1)}, \quad i = 1,\dots,r. \qquad (22)$$

Hence, when a steplength selection rule provides a suitable approximation of $1/\gamma_i$ and $\nu_k = 1$, it follows that $\mu_i^{(k+1)} + \zeta_i^{(k+1)} = 0$, i.e., the corresponding $i-$th eigencomponent of the subgradient $g^{(k+1)} + v^{(k+1)} \in \partial f(x^{(k+1)})$ is annihilated, thus improving optimality. As a consequence, at the $(k+1)$-th iteration, a steplength selection rule must aim at approximating the inverses of the eigenvalues of the submatrix $A_{\mathcal{I}_k\,\mathcal{I}_k}$ in order to be effective.

In view of these considerations, the BB1 rule computed at a given k-th iteration possesses the intrinsic property of sweeping the spectrum of the matrix $A^{-1}_{\mathcal{I}_{k-1}\,\mathcal{I}_{k-1}}$. This fact can be proved, by analogy with [19, Theorem 2], in the following theorem.

Theorem 3.1. *Given the problem* (9)*, if the matrix A is symmetric positive definite, then*

$$\frac{1}{\gamma_{\max}(A_{\mathcal{I}_{k-1}\,\mathcal{I}_{k-1}})} \le \alpha_k^{\mathrm{BB1}} \le \frac{1}{\gamma_{\min}(A_{\mathcal{I}_{k-1}\,\mathcal{I}_{k-1}})} \qquad (23)$$

where $\gamma_{\max}(A_{\mathcal{I}_{k-1}\,\mathcal{I}_{k-1}})$ and $\gamma_{\min}(A_{\mathcal{I}_{k-1}\,\mathcal{I}_{k-1}})$ are the maximum and the minimum eigenvalue of the matrix $A_{\mathcal{I}_{k-1}\,\mathcal{I}_{k-1}}$, respectively.

*Proof.*From definition (6) and equations (18)-(20)

$$
\begin{aligned}
\alpha_k^{\mathrm{BB1}} = \frac{{s^{(k-1)}}^T s^{(k-1)}}{{s^{(k-1)}}^T z^{(k-1)}} &= \frac{{s^{(k-1)}_{\mathcal{J}_{k-1}}}^T s^{(k-1)}_{\mathcal{J}_{k-1}} + {s^{(k-1)}_{\mathcal{I}_{k-1}}}^T s^{(k-1)}_{\mathcal{I}_{k-1}}}{{s^{(k-1)}_{\mathcal{J}_{k-1}}}^T z^{(k-1)}_{\mathcal{J}_{k-1}} + {s^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}} \\
&= \frac{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T s^{(k-1)}_{\mathcal{I}_{k-1}}}{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}} \\
&= \frac{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T s^{(k-1)}_{\mathcal{I}_{k-1}}}{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}} s^{(k-1)}_{\mathcal{I}_{k-1}}},
\end{aligned}
$$

then, the steplength α_k^{BB1} corresponds to the inverse of a Rayleigh quotient of the submatrix $A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}}$, and the thesis follows from the extremal properties of the Rayleigh quotients (see [37], Theorem 4.2.2). □

From the previous theorem, we can recover the quasi-Newton interpretation of the BB1 rule as the steplength satisfying the following secant condition

$$
\alpha_k^{\mathrm{BB1}} = \underset{\alpha\in\mathbb{R}}{\operatorname{argmin}} \|\alpha^{-1} s^{(k-1)}_{\mathcal{I}_{k-1}} - z^{(k-1)}_{\mathcal{I}_{k-1}}\| = \frac{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T s^{(k-1)}_{\mathcal{I}_{k-1}}}{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}}. \tag{24}
$$

On the other hand, the use of the BB2 rule defined in (7) applied to problem (9) may not guarantee the sweeping of the spectrum of the submatrix $A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}}$, due to the presence of an additional term that depends on the components of the vector $z^{(k-1)}$ that are indexed in $\mathcal{J}_{k-1}$. Indeed, from (7) we obtain

$$
\begin{aligned}
\alpha_k^{\mathrm{BB2}} = \frac{{s^{(k-1)}}^T z^{(k-1)}}{{z^{(k-1)}}^T z^{(k-1)}} &= \frac{{s^{(k-1)}_{\mathcal{J}_{k-1}}}^T z^{(k-1)}_{\mathcal{J}_{k-1}} + {s^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}}{{z^{(k-1)}_{\mathcal{J}_{k-1}}}^T z^{(k-1)}_{\mathcal{J}_{k-1}} + {z^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}} = \\
&= \frac{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}}{{z^{(k-1)}_{\mathcal{J}_{k-1}}}^T z^{(k-1)}_{\mathcal{J}_{k-1}} + {z^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}}.
\end{aligned}
$$

Following the idea proposed in [19], we introduce a modified BB2 rule, with the aim of imitating the spectral behaviour of BB1. The new rule, which we refer to as BB2$-\ell_1$, is defined as follows:

$$
\alpha_k^{\mathrm{BB2}-\ell_1} = \frac{{s^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}}{{z^{(k-1)}_{\mathcal{I}_{k-1}}}^T z^{(k-1)}_{\mathcal{I}_{k-1}}}. \tag{25}
$$

Now, for the BB2 version above defined, a similar result to Theorem 3.1 can be proved:

Theorem 3.2. *Given problem* (9)*, if the matrix A is symmetric positive definite, then*

$$
\frac{1}{\gamma_{\max}(A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}})} \le \alpha_k^{\mathrm{BB2}-\ell_1} \le \frac{1}{\gamma_{\min}(A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}})} \tag{26}
$$

where $\gamma_{\max}(A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}})$ and $\gamma_{\min}(A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}})$ are the maximum and the minimum eigenvalue of the matrix $A_{\mathcal{I}_{k-1}\mathcal{I}_{k-1}}$, respectively.

In addition, from the Cauchy-Schwarz inequality, we obtain

$$
\alpha_k^{\mathrm{BB2}-\ell_1} \le \alpha_k^{\mathrm{BB1}}. \tag{27}
$$

Remark 3.1. *This analysis is similar to the one developed in [19]. More precisely, the set $\mathcal{J}_{k-1}$ is defined according to the same principle followed for the gradient projection scheme applied to a box-constrained minimization problem in [19, Section 2.1]. Indeed, by denoting with $\Omega = \{x \in \mathbb{R}^n : \ell_i \leq x_i \leq u_i,\ i = 1, \dots, n\}$ a box-constraint, a point x^* solves the box-constrained problem if and only if*

$$\begin{aligned} x^* \in \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\ f(x) + \iota_\Omega(x) \quad &\Leftrightarrow \quad -\nabla f(x^*) \in \partial\iota_\Omega(x^*) \\ &\Leftrightarrow \quad -(\nabla f(x^*))_i \in N_{\ell_i\leq x\leq u_i}(x_i^*), \quad i = 1, \dots, n, \end{aligned}$$

where $N_{\ell_i\leq x\leq u_i}(x_i^*)$ is the normal cone of the interval $[\ell_i, u_i]$ at x_i^*. Then, in this case,

$$\begin{aligned} \mathcal{J}_{k-1} &= \left\{i :\ x_i^{(k-1)} \in \{\ell_i, u_i\} \ \wedge\ -g_i^{(k-1)} \in N_{\ell_i\leq x\leq u_i}(x_i^{(k-1)})\right\} \\ &= \left\{i :\ (x_i^{(k-1)} = \ell_i \ \wedge\ g_i^{(k-1)} \geq 0) \ \vee\ (x_i^{(k-1)} = u_i \ \wedge\ g_i^{(k-1)} \leq 0)\right\}, \end{aligned}$$

which is the binding set at $x_i^{(k-1)}$. Given $\mathcal{I}_{k-1} = \{1, \dots, n\} \setminus \mathcal{J}_{k-1}$, the proposed BB2-like for box-constrained problems is then defined as [19, eq. (28)]

$$\alpha_k^{BoxBB2} = \frac{{s_{\mathcal{I}_{k-1}}^{(k-1)}}^T z_{\mathcal{I}_{k-1}}^{(k-1)}}{{z_{\mathcal{I}_{k-1}}^{(k-1)}}^T z_{\mathcal{I}_{k-1}}^{(k-1)}},$$

specularly to our proposed BB2-like rule (25) *for ℓ_1-regularized problems.*

3.2. The non-quadratic case

Let us now consider the general ℓ_1-regularized problem (1). We want to recover the spectral properties of BB1 and BB2-ℓ_1 for the general non-quadratic case. To this aim, we assume that the curvature condition

$$ {s^{(k-1)}}^T z^{(k-1)} > 0 \tag{28}$$

holds. We remark that when f_0 is the strictly convex quadratic function defined in the previous subsection, condition (28) is satisfied for any $x^{(k)}$ and $x^{(k-1)}$ [43]. Let us define the average Hessian as [43, eq. (8.11)]

$$\tilde{H}^{(k-1)} = \int_0^1 \nabla^2 f_0(x^{(k-1)} + ts^{(k-1)})dt.$$

From the multidimensional variant of the Taylor's theorem [43, Thm. 11.1], we have

$$z^{(k-1)} = \nabla f_0(x^{(k)}) - \nabla f_0(x^{(k-1)}) = \int_0^1 \nabla^2 f_0(x^{(k-1)} + ts^{(k-1)})s^{(k-1)}dt = \tilde{H}^{(k-1)}s^{(k-1)}.$$

Since $s_{\mathcal{J}_{k-1}}^{(k-1)} = 0$, which holds by repeating the same passages employed to get (16) in Section 3, the previous equation yields

$$\begin{aligned} \frac{1}{\alpha_k^{\text{BB1}}} &= \int_0^1 \frac{{s^{(k-1)}}^T \nabla^2 f_0(x^{(k-1)} + ts^{(k-1)})s^{(k-1)}dt}{\|s^{(k-1)}\|^2} \\ &= \int_0^1 \frac{{s_{\mathcal{I}_{k-1}}^{(k-1)}}^T \nabla^2 f_0(x^{(k-1)} + ts^{(k-1)})_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}} s_{\mathcal{I}_{k-1}}^{(k-1)}dt}{\|s_{\mathcal{I}_{k-1}}^{(k-1)}\|^2} \\ &= \frac{{s_{\mathcal{I}_{k-1}}^{(k-1)}}^T \tilde{H}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}^{(k-1)} s_{\mathcal{I}_{k-1}}^{(k-1)}}{\|s_{\mathcal{I}_{k-1}}^{(k-1)}\|^2}. \end{aligned} \tag{29}$$

The above equation expresses the fact that the inverse of the steplength defined by the BB1 rule can be interpreted as the Rayleigh quotient related to the average matrix $\tilde{H}^{(k-1)}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}$. Assuming that the average Hessian is positive definite, there exists a square root $(\tilde{H}^{(k-1)})^{1/2}$, and we may write

$$\alpha_k^{\mathrm{BB2}-\ell_1} = \frac{\left((\tilde{H}^{(k-1)}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}})^{1/2} s^{(k-1)}_{\mathcal{I}_{k-1}}\right)^T (\tilde{H}^{(k-1)}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}})^{1/2} s^{(k-1)}_{\mathcal{I}_{k-1}}}{\left((\tilde{H}^{(k-1)}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}})^{1/2} s^{(k-1)}_{\mathcal{I}_{k-1}}\right)^T \tilde{H}^{(k-1)}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}} (\tilde{H}^{(k-1)}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}})^{1/2} s^{(k-1)}_{\mathcal{I}_{k-1}}}.$$

Again, the above formula tells us that the reciprocal of the BB2-ℓ_1 steplength approximates one of the eigenvalues of the average matrix $\tilde{H}^{(k-1)}_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}$.

As a final observation, we expect that the simple correction introduced to the BB2 rule, in both single and alternating strategies, can affect the practical performance of the proximal gradient scheme (3), fostering the acceleration of the method. Section 5 is dedicated to highlight this fact by evaluating the proposed rules on different ℓ_1-regularized test problems, both quadratic and non-quadratic.

4. A non monotone line search based proximal gradient method

We are now interested in defining a line search procedure for computing the parameter ν_k in the proximal gradient method (3)-(4) that fits properly with the BB2-ℓ_1 steplength selection rule (25). The proposed line search aims at generalizing the non monotone strategies adopted in combination with the BB rules for gradient projection methods, see e.g. [4, 10, 19, 18, 27]. The resulting algorithm and related convergence analysis is valid for the general composite problem (2).

Several existing line search strategies proposed for (3)-(4) are *monotone*, i.e., they enforce the sequence $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ to be monotone decreasing. Here, we recall the generalized monotone Armijo-type line search first proposed in [49], and then further employed and studied in [6, 8, 9, 40]. Given $\sigma, \beta \in (0, 1)$, $\gamma \in [0, 1]$, and the function $h^{(k)} \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ defined as [6, Section 3, eq. (9)]

$$h^{(k)}_\gamma(z, x) = \nabla f_0(x)^T(z - x) + \frac{\gamma}{2\alpha_k}\|z - x\|^2 + f_1(z) - f_1(x), \qquad z, x \in \mathbb{R}^n, \tag{30}$$

the aforementioned line search sets the parameter as $\nu_k = \beta^{m_k}$, where m_k is the first nonnegative integer such that

$$f(x^{(k)} + \beta^{m_k}(y^{(k)} - x^{(k)})) \le f(x^{(k)}) + \sigma\beta^{m_k} h^{(k)}_\gamma(y^{(k)}, x^{(k)}). \tag{31}$$

Note that, due to the fact that $y^{(k)}$ is the minimum point of $h^{(k)}_1(\cdot, x^{(k)})$, there holds

$$h^{(k)}_\gamma(y^{(k)}, x^{(k)}) \le h^{(k)}_1(y^{(k)}, x^{(k)}) \le h^{(k)}_1(x^{(k)}, x^{(k)}) \le 0,$$

hence (31) enforces the sequence $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ to be monotone non increasing. The generalized Armijo-type line search based on (31) is well-defined and terminates in a finite number of steps, see [6, Proposition 3.1]. Condition (31) naturally reduces to the classical Armijo condition for constrained differentiable problems when we set $\gamma = 0$ and $f_1(x) = \iota_\Omega(x)$, being $\Omega \subseteq \mathbb{R}^n$ a non empty, closed and convex set.

Since the rules defined in Section 3 aim at providing larger steplengths α_k than the ones defined by the inverse of the Lipschitz constant of ∇f_0, it might happen that the point $y^{(k)}$ yields an increase of the objective function, rather than a decrease. As a consequence, a monotone line search procedure might require several backtracking iterations before the Armijo-like condition (31) is met, which in turn might dissipate the acceleration effect of the Barzilai-Borwein-like strategy. Thus, it is important to equip (3)-(4) with a line search procedure that preserves the intrinsic non monotonicity of the BB-like rules (24)-(25).

In this light, we propose to compute the line search parameter ν_k in (3)-(4) as $\nu_k = \beta^{m_k}$, where m_k is the first nonnegative integer such that the following nonmonotone Armijo-like condition holds

$$f(x^{(k)} + \beta^{m_k}(y^{(k)} - x^{(k)})) \le \bar{f}_k + \sigma\beta^{m_k} h^{(k)}_\gamma(y^{(k)}, x^{(k)}), \tag{32}$$

where $\bar{f}_k$ is a reference value with respect to which the objective function must decrease. By introducing a nonnegative sequence $\{\xi_k\}_{k\in\mathbb{N}}$, we can specify the reference value as follows

$$\bar{f}_k = f(x^{(k)}) + \xi_k, \quad \xi_k \geq 0. \tag{33}$$

Note that, when $\xi_k \equiv 0$, (32) reduces to the monotone Armijo-like condition (31). We also remark that the line search based on (32) is performed along the feasible direction $d^{(k)} = y^{(k)} - x^{(k)}$, and thus we do not require the recalculation of the proximal operator at each backtracking iteration. In this sense, our proposal differs substantially from other non monotone line searches adopted in the non differentiable setting, which perform the backtracking directly on the steplength parameter α_k rather than on ν_k, see e.g. [2, 48, 51].

We report the proximal gradient method (3)-(4) combined with the non monotone Armijo-like condition (32) in Algorithm 1.

Algorithm 1 Proximal gradient method with non monotone line search

```
Choose $0 < \alpha_{\min} \leq \alpha_{\max}$, $\gamma \in [0,1]$, $\sigma, \beta \in (0,1)$, $x^{(0)} \in \mathrm{dom}(f_1)$.
For $k = 0, 1, 2, ...$

STEP 1  Choose $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$.
STEP 2  Compute $y^{(k)} = \mathrm{prox}_{\alpha_k f_1}(x^{(k)} - \alpha_k \nabla f_0(x^{(k)}))$.
STEP 3  Set $d^{(k)} = y^{(k)} - x^{(k)}$ and $\bar{f}_k = f(x^{(k)}) + \xi_k$ with $\xi_k \geq 0$.
STEP 4  Compute the smallest nonnegative integer $m_k$ such that
            $f(x^{(k)} + \beta^{m_k} d^{(k)}) \leq \bar{f}_k + \sigma \beta^{m_k} h_\gamma^{(k)}(y^{(k)}, x^{(k)})$        (34)
        and set $\nu_k = \beta^{m_k}$.
STEP 5  Compute $x^{(k+1)} = x^{(k)} + \nu_k d^{(k)}$.
```

In the following, we present the convergence analysis of Algorithm 1 for problem (2). The analysis will combine elements from the works [6, 27]. From now on, $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{y^{(k)}\}_{k\in\mathbb{N}}$, $\{\nu_k\}_{k\in\mathbb{N}}$ will denote the sequences generated by Algorithm 1. Furthermore, we assume that $f_0 : \Omega \to \mathbb{R}$ is continuously differentiable on an open set Ω containing $\overline{\mathrm{dom}(f_1)}$, and $f_1 : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ is proper, convex, and lower semicontinuous throughout the entire analysis. Further assumptions on the involved functions will be specified where needed.

In the following result, we state that the line search performed at Step 4 of Algorithm 1 terminates in a finite number of steps.

Theorem 4.1. *Suppose that*

$$h_\gamma(y^{(k)}, x^{(k)}) < 0, \quad \forall\ k \geq 0. \tag{35}$$

Then, for all $k \geq 0$, the generalized Armijo line search at Step 4 *of Algorithm 1 is well-defined, i.e., there exists $m_k < \infty$ such that* (34) *holds.*

*Proof.*See the Appendix. □

The theorem below states that each limit point of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ is stationary, under appropriate conditions on the parameters $\{\xi_k\}_{k\in\mathbb{N}}$. As the result is an easy generalization of [6, Theorem 3.1] to the case where the line search is non monotone, we postpone the proof to the Appendix.

Theorem 4.2. *Suppose that the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ admits a limit point, and let $K \subseteq \mathbb{N}$ be a subset of indices such that $\lim_{k\in K,k\to\infty} x^{(k)} = \bar{x}$. Assume that the sequence $\{\xi_k\}_{k\in\mathbb{N}}$ satisfies the following conditions*

$$\lim_{k\to\infty} \xi_k = 0, \tag{36}$$

$$0 \leq \xi_{k+1} \leq (1-\delta_{k+1})(f(x^{(k)}) + \xi_k - f(x^{(k+1)})), \quad \text{where } 0 \leq \delta_{\min} \leq \delta_{k+1} \leq 1. \tag{37}$$

Then $\bar{x}$ is a stationary point for problem (2).

*Proof.*See the Appendix. □

In the experiments of Section 5, we will perform Step 4 of Algorithm 1 with the following reference value

$$\bar{f}_k = \max_{0\leq j\leq \min\{k,M-1\}} f(x^{(k-j)}), \tag{38}$$

i.e., at each backtracking iteration i, we compare the trial function value $f(x^{(k)} + \beta^i d^{(k)})$ with the maximum among the M past function values, where M is a positive integer. The same max-type strategy has been frequently used in the differentiable setting, see e.g. [34, 10, 19, 27]. The following corollary states that the stationarity of the limit points holds when Algorithm 1 is equipped with the choice (38) for the reference value.

Corollary 4.1. *Suppose that the parameter $\nu_k \in (0,1]$ is computed according to* (34) *with the following choice of the reference value ξ_k:*

$$\xi_k = \max_{0\leq j\leq \min\{k,M-1\}} f(x^{(k-j)}) - f(x^{(k)}) \geq 0, \quad \forall\ k \geq 0, \tag{39}$$

where $M \geq 1$ is an integer parameter. Assume that $\mathrm{dom}(f_1)$ *is closed, f_1 is continuous on* $\mathrm{dom}(f_1)$, *and the level set $\Omega_0 = \{x \in \mathrm{dom}(f_1) : \ f(x) \leq f(x^{(0)})\}$ is bounded. Then, any limit point $\bar{x}$ of the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ is stationary for problem* (2).

Proof. The proof follows by extending the arguments employed in [27, Proposition 14] from the constrained differentiable setting to a more general nondifferentiable setting.

Without loss of generality, we assume that $h_\gamma(y^{(k)}, x^{(k)}) < 0$ for all $k \geq 0$; if this is not true, then one of the iterate is stationary and the sequence remains indefinitely stuck to that iterate (see proof of Theorem 4.2).

Our goal is to show that the sequence $\{\xi_k\}_{k\in\mathbb{N}}$ defined in (39) satisfies the hypotheses (36)-(37) of Theorem 4.2, from which the stationarity of the limit points follows. For the sake of brevity, we let $f(x^{(\ell(k))}) = \max_{0\leq j\leq\min\{k,M-1\}} f(x^{(k-j)})$.

First, we note that condition (37) holds by setting

$$\xi_k = f(x^{(\ell(k))}) - f(x^{(k)}), \quad 0 = \delta_{\min} \leq \delta_{k+1} \leq \frac{f(x^{(\ell(k))}) - f(x^{(\ell(k+1))})}{f(x^{(\ell(k))}) - f(x^{(k+1)})}. \tag{40}$$

Indeed, since $h_\gamma(y^{(k)}, x^{(k)}) < 0$, the nonmonotone Armijo-like condition (32) with the reference value (39) implies that $f(x^{(\ell(k))}) - f(x^{(k+1)}) > 0$. Since $\ell(k+1) \leq \ell(k)+1$, we deduce that $f(x^{(\ell(k))}) - f(x^{(\ell(k+1))}) \geq 0$. Therefore, since $f(x^{(k+1)}) \leq f(x^{(\ell(k+1))})$ by definition of $\ell(k+1)$, we conclude that $\delta_{k+1} \in [0,1]$. Moreover, the right-hand side inequality is equivalent to

$$\delta_{k+1}(f(x^{(\ell(k))}) - f(x^{(k+1)})) \leq f(x^{(\ell(k))}) - f(x^{(\ell(k+1))}),$$

or, recalling the definition of ξ_k in (39), also as

$$\delta_{k+1}(f(x^{(k)}) + \xi_k - f(x^{(k+1)})) \leq (f(x^{(k)}) + \xi_k) - (f(x^{(k+1)}) + \xi_{k+1}).$$

By rearranging terms in the previous inequality, we obtain (37).

Second, we show that condition (36) is satisfied. To this aim, we start by recalling that the function $h_1(\cdot; x^{(k)})$ is strongly convex with modulus $m = 1/\alpha_{\max}$, i.e,

$$h_1^{(k)}(x, x^{(k)}) \geq h_1^{(k)}(y, x^{(k)}) + w^T(x - y) + \frac{m}{2}\|x - y\|^2, \quad \forall\ x, y \in \mathbb{R}^n,\ \forall\ w \in \partial_x h_1^{(k)}(y, x^{(k)}). \tag{41}$$

Applying inequality (41) with $x = x^{(k)}$, $y = y^{(k)}$, $w = 0 \in \partial h_1^{(k)}(y^{(k)}, x^{(k)})$, and $m = 1/\alpha_{\max}$, yields

$$\|y^{(k)} - x^{(k)}\|^2 \leq -2\alpha_{\max} h_1^{(k)}(y^{(k)}, x^{(k)}). \tag{42}$$

Since (37) holds, we know from (A.1) that the sequence $\{f(x^{(k)}) + \xi_k\}_{k\in\mathbb{N}} = \{f(x^{(\ell(k))})\}_{k\in\mathbb{N}}$ is monotone nonincreasing. Consequently, we can write $f(x^{(k+1)}) \leq f(x^{(k+1)}) + \xi_{k+1} \leq f(x^{(k)}) + \xi_k \leq f(x^{(0)})$, where the last inequality comes from the fact that $\xi_0 = 0$. Thus, we have that the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ is contained in the compact set Ω_0, which implies (being f continuous) that the monotone nonincreasing sequence $\{f(x^{(\ell(k))})\}_{k\in\mathbb{N}}$ is bounded from below, hence convergent. Note also that $\ell(k)$ is an integer such that

$$k - \min\{M - 1, k\} \leq \ell(k) \leq k. \tag{43}$$

Since $x^{(\ell(k))} = x^{(\ell(k)-1)} + \nu_{\ell(k)-1}(y^{(\ell(k)-1)} - x^{(\ell(k)-1)})$, condition (32) equipped with (39) implies that

$$f(x^{(\ell(k))}) \leq f(x^{(\ell(\ell(k)-1))}) + \sigma\nu_{\ell(k)-1} h_\gamma(y^{(\ell(k)-1)}, x^{(\ell(k)-1)}), \quad \forall\ k > M.$$

Observing that $h_\gamma(z, x^{(\ell(k)-1)}) \leq h_1(z, x^{(\ell(k)-1)})$ for all $z \in \mathbb{R}^n$, and combining the previous inequality with (42), we get

$$f(x^{(\ell(k))}) \leq f(x^{\ell(\ell(k)-1)}) - \frac{\sigma\nu_{\ell(k)-1}}{2\alpha_{\max}}\|y^{(\ell(k)-1)} - x^{(\ell(k)-1)}\|^2, \quad \forall\ k > M.$$

Since $\{f(x^{(\ell(k))})\}_{k\in\mathbb{N}}$ is a convergent sequence, and $\{\|y^{(\ell(k)-1)} - x^{(\ell(k)-1)}\|\}_{k\in\mathbb{N}}$ is a bounded sequence (due to the boundedness of $\{x^{(k)}\}_{k\in\mathbb{N}}$, $\{\alpha_k\}_{k\in\mathbb{N}}$, and the continuity of the proximal operator) the previous inequality implies that

$$\lim_{k\to\infty} \nu_{\ell(k)-1}\|y^{(\ell(k)-1)} - x^{(\ell(k)-1)}\| = 0. \tag{44}$$

Our aim is now to show that $\lim_{k\to\infty} f(x^{(\ell(k))}) = \lim_{k\to\infty} f(x^{(k)})$, where the latter limit exists finite due to the monotonicity of $\{f(x^{(k)}) + \xi_k\}_{k\in\mathbb{N}}$ and the continuity of f (see proof of Theorem 4.2). To this aim, we set $\hat{\ell}(k) = \ell(k + M + 1)$, and prove by induction that

$$\lim_{k\to\infty} \nu_{\hat{\ell}(k)-j}\|y^{(\hat{\ell}(k)-j)} - x^{(\hat{\ell}(k)-j)}\| = 0, \quad \lim_{k\to\infty} f(x^{(\hat{\ell}(k)-j)}) = \lim_{k\to\infty} f(x^{\ell(k)}), \tag{45}$$

for all $j \geq 1$ and $k \geq j - 1$. For $j = 1$, the first equality in (45) follows directly from (44), as $\{\hat{\ell}(k) :\ k \geq 0\} \subseteq \{\ell(k) :\ k \geq 0\}$. Consequently, from (4), there holds $\lim_{k\to\infty} \|x^{(\hat{\ell}(k))} - x^{(\hat{\ell}(k)-1)}\| = 0$. Since Ω_0 is compact and f is uniformly continuous on Ω_0, the previous limit yields $\lim_{k\to\infty} f(x^{(\hat{\ell}(k)-1)}) = \lim_{k\to\infty} f(x^{\hat{\ell}(k)}) = 0$, which implies the second equality in (45), due again to $\{\hat{\ell}(k) :\ k \geq 0\} \subseteq \{\ell(k) :\ k \geq 0\}$. Now, assume that (45) holds for a given $j \geq 1$. From (4), we can write

$$x^{(\hat{\ell}(k)-j)} = x^{(\hat{\ell}(k)-j-1)} + \nu_{\hat{\ell}(k)-j-1}(y^{(\hat{\ell}(k)-j-1)} - x^{(\hat{\ell}(k)-j-1)}).$$

Hence, condition (32) equipped with (39) implies again that

$$f(x^{(\hat{\ell}(k)-j)}) \leq f(x^{(\ell(\hat{\ell}(k)-j-1))}) + \sigma\nu_{\hat{\ell}(k)-j-1} h_\gamma^{(k)}(y^{(\hat{\ell}(k)-j-1)}, x^{(\hat{\ell}(k)-j-1)}).$$

Analogous arguments employed to obtain (44) show that

$$\lim_{k\to\infty} \nu_{\hat{\ell}(k)-(j+1)}\|y^{(\hat{\ell}(k)-(j+1))} - x^{(\hat{\ell}(k)-(j+1))}\| = 0,$$

namely, the first limit in (45) holds for $j+1$, from which we obtain $\lim_{k\to\infty} \|x^{(\hat{\ell}(k)-j)} - x^{(\hat{\ell}(k)-(j+1))}\| = 0$. By using again the fact that f is uniformly continuous on Ω_0, we get

$$\lim_{k\to\infty} f(x^{(\hat{\ell}(k)-(j+1))}) = \lim_{k\to\infty} f(x^{(\hat{\ell}(k)-j)}),$$

from which the second limit in (45) follows thanks to the induction hypothesis. Thus, the proof of (45) is complete.

Finally, note that from(43) and $\hat{\ell}(k) = \ell(k+M+1)$, it follows that $\hat{\ell}(k) - k - 1 \leq M$. Then, since from (4) we can write the following relation between the iterates $x^{(k+1)}$ and $x^{\hat{\ell}(k)}$

$$x^{(k+1)} = x^{\hat{\ell}(k)} - \sum_{j=1}^{\hat{\ell}(k)-k-1} \nu_{\hat{\ell}(k)-j}(y^{(\hat{\ell}(k)-j)} - x^{(\hat{\ell}(k)-j)}),$$

it follows from the first limit in (45) that

$$\lim_{k\to+\infty} \|x^{(k+1)} - x^{\hat{\ell}(k)}\| = 0.$$

Therefore, the uniform continuity of f on Ω_0 and the convergence of the the sequence $\{f(x^{\ell(k)})\}_{k\in\mathbb{N}}$ allow us to conclude that

$$\lim_{k\to\infty} f(x^{(k)}) = \lim_{k\to\infty} f(x^{\hat{\ell}(k)}) = \lim_{k\to\infty} f(x^{\ell(k)}),$$

which is equivalent to say that (36) holds with the choice of ξ_k specified in (39). Now, the stationarity of the limit points follows from Theorem 4.2. □

5. Numerical experiments

In this section, we provide numerical experiments on both quadratic and non-quadratic ℓ_1-regularized problems, with the aim of evaluating the numerical performances obtained by Algorithm 1 equipped with different BB-based steplengths. In particular, we compare the behaviour of the modified rule introduced in (25) with the standard BB rules employed in both single and alternating strategies. The experiments were performed in the Matlab R2022b environment on the *magicbox* server operating at the Dept. of Mathematics and Physics of the University of Campania "L. Vanvitelli", equipped with 8 Intel Xeon Platinum 8168 CPUs, 1536 GB of RAM and Linux CentOS 7.5 operating system. A single Intel Xeon CPU with 192 GB of RAM was used in the experiments.

5.1. Numerical results on quadratic ℓ_1-regularized problems

In this subsection, we inspect the spectral behaviour in practical implementations of the BB rules introduced in Section 3 on some special problems of type (9), which were built by considering different eigenvalues distributions for the matrix A. In these numerical tests, we equipped Algorithm 1 with the steplength rules BB1, BB2, BB2-ℓ_1, ABB$_{\min}$ [32] and ABB$_{\min}$-ℓ_1, the latter of which refers to the modified adaptive strategy obtained from ABB$_{\min}$ by replacing the standard BB2 rule with the modified version BB2-ℓ_1 (25).

All the methods were stopped when the relative difference between two consecutive iterations was sufficiently small, i.e.,

$$\|x^{(k)} - x^{(k-1)}\|_\infty \leq \epsilon \cdot \|x^{(k)}\|_\infty \tag{46}$$

with $\epsilon = 10^{-15}$ or a maximum number of 2000 iterations was reached. The parameter setting used is the following: $M = 10$, $\beta = 0.5$, $\gamma = 1$, $\sigma = 10^{-4}$, $\alpha_0 = 1$, $\alpha_{min} = 10^{-10}$, $\alpha_{max} = 10^6$; in ABB$_{\min}$, and ABB$_{\min}$-ℓ_1 the parameter τ that controls the switching between the two BB rules was set equal to 0.6. The values chosen for the Armijo linesearch parameters M, β, γ, σ are quite common in the literature, see e.g. [4, 10, 6, 40, 27], where similar or identical values are adopted. It has been noted that the performance of Armijo linesearches is not particularly sensitive with respect to the choice of these parameters [10, 6]. Concerning $\alpha_{min}, \alpha_{max}$, we opted for the strategy of using a tiny value for α_{min} and a huge value for α_{max}, employing the same values adopted in [19, 18], whereas α_0 was set to 1 for simplicity. Finally, the switching parameter τ was selected as an intermediate value in the interval $(0, 1)$, as a value too close to 1 might favour the selection of the BB2 rule for most iterations, whereas a value too close to 0 might impose a bias towards the selection of the BB1 rule.

We considered four test problems of size $n = 1000$, denoted by LQP1, LQP2, LQP3, LQP4. For each problem, the solution x^* is defined in order that half of the entries are zero and the remaining ones are uniformly randomly chosen within an interval around 0. Furthermore, we report the distributions of the eigenvalues of the matrix A and the starting vectors for each test problem:

- LQP1: random uniform distribution in $[1, 10^3]$ such that $\gamma_{\min} = 1$ and $\gamma_{\max} = 10^3$, $x^{(0)} = 2 \cdot r$ where $r \in \mathbb{R}^n$ has random entries from uniform distribution in $[1, 10^3]$;
- LQP2: for $i = 1, \ldots, n$,

  $$\gamma_i = \frac{\gamma_{\min} + \gamma_{\max}}{2} + \frac{\gamma_{\min} - \gamma_{\max}}{2} \cos\left(\frac{\pi(i-1)}{n-1}\right),$$

  where $\gamma_{\min} = 1$ and $\gamma_{\max} = 10^3$, $x^{(0)} = 2 \cdot r$ where $r \in \mathbb{R}^n$ has random entries from uniform distribution in $[1, 10^3]$;
- LQP3: for $i = 1, \ldots, n$,

  $$\gamma_i = \frac{(\underline{\gamma} b - \overline{\gamma} a)}{(b-a)} + \frac{(\underline{\gamma} - \overline{\gamma})}{(b-a)} \omega_i,$$

  where $\underline{\gamma} = 1$, $\overline{\gamma} = 10^4$, $a = (1-c)^2$, $b = (1+c)^2$, $c = 1/2$ and the values ω_i are distributed in accordance with the Marčenko-Pastur density $p_c(x) = \dfrac{\sqrt{(b-a)(x-a)}}{2\pi x c^2}$, $a < x < b$ [41], such that $\gamma_{\min} = 15$, $\gamma_{\max} = 9863$, $x^{(0)} = 10 \cdot r$ where $r \in \mathbb{R}^n$ has random entries from uniform distribution in $[1, 10^3]$;
- LQP4: logarithmic distribution in $[1, 10^3]$ generated through the MATLAB function `logspace`, such that $\gamma_{\min} = 1$, $\gamma_{\max} = 10^3$, $x^{(0)} = 10 \cdot r$ where $r \in \mathbb{R}^n$ has random entries from uniform distribution in $[1, 10^3]$.

Figures 1-4 show how the spectral properties of the considered proximal gradient scheme can be affected by different choices of the steplength, resulting in different effects on the acceleration. In the top panels of the figures, for each version of the algorithms the errors $\frac{\|x^{(k)} - x^*\|}{\|x^*\|}$ (left panel) and $f(x^{(k)}) - f(x^*)$ (right panel) are compared; the remaining panels of the figures show how the sequence of the inverse of steplengths $\{\frac{1}{\alpha_k}\}$ distributes with respect to the spectra of the submatrices $A_{\mathcal{I}_{k-1}, \mathcal{I}_{k-1}}$: in particular, at the k-th iteration, the black dots denote 20 eigenvalues of $A_{\mathcal{I}_{k-1}, \mathcal{I}_{k-1}}$ with linearly spaced indices (included the maximum and the minimum eigenvalues), the red crosses represent the quantities $\frac{1}{\alpha_k}$, and the blue lines correspond to the maximum and the minimum eigenvalues of A respectively. For all the test problems, we may observe that the standard BB2 steplength reveals unsatisfactory results, since it is not able to properly exploit the correct information deriving from the spectra of the matrices $A_{\mathcal{I}_{k-1}, \mathcal{I}_{k-1}}$ along the iterative procedure, causing also damaging effects on the $\text{ABB}_{\min}$ strategy. On the other hand, the BB1 rule confirms its natural ability of capturing the second order information correctly; a similar behaviour is realized by the modifications introduced in the BB2-ℓ_1 rule, which enable an earlier stabilization of the nonzero components with respect to the original BB2 rule. Finally, the benefits of the alternating strategy are preserved when the modified rule BB2-ℓ_1 is employed, as we can observe from the plots related to $\text{ABB}_{\min}$-ℓ_1.

To better assess the performance of the proposed strategies, we prepared a set of l1-regularized QP problems inspired by the synthetic problems used in [25]. As done for the previous test, we first fix a

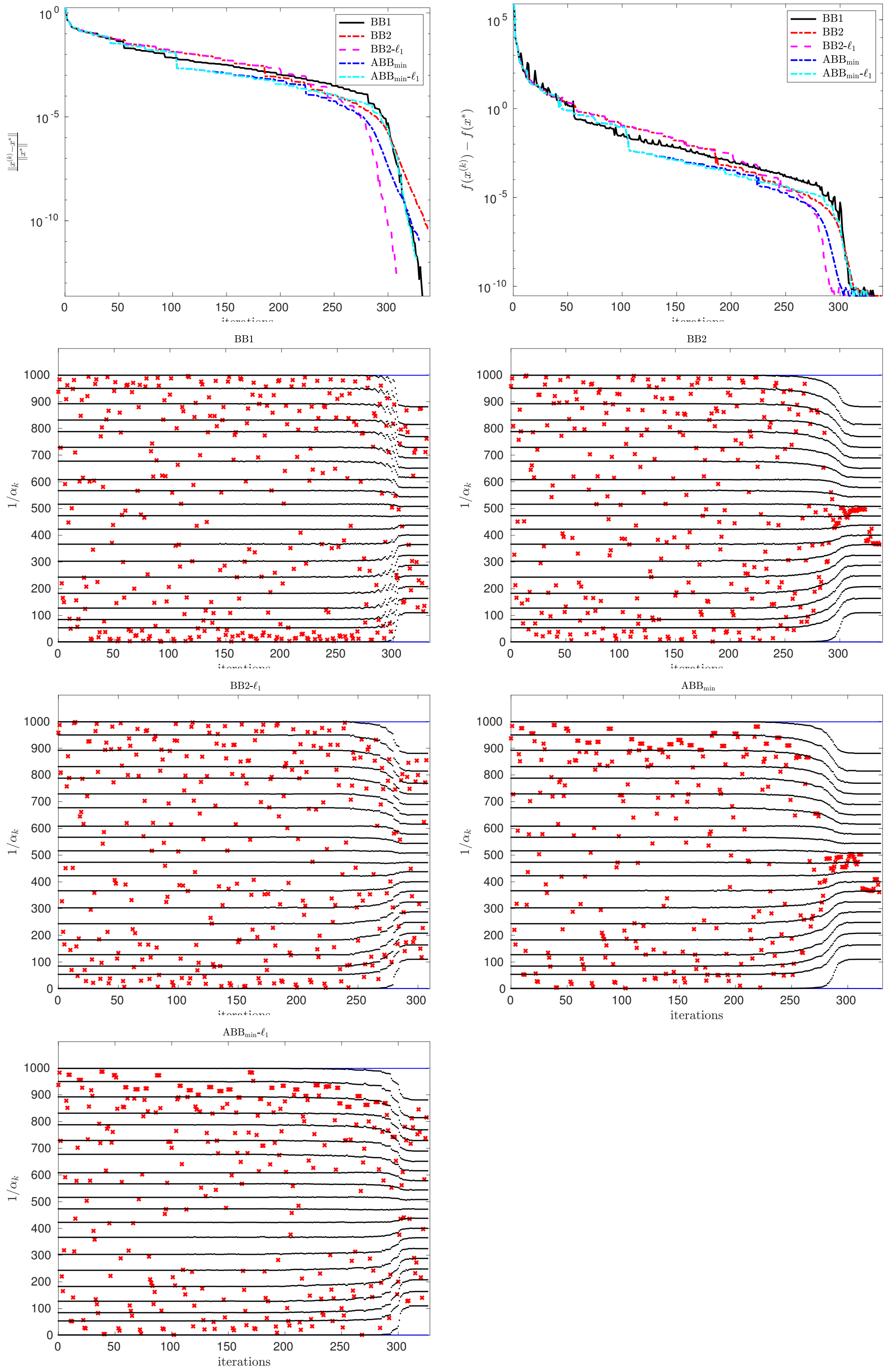

Figure 1: Results on LQP1. Errors on the computed solution and on the objective function for the different rules (first row). Distribution of $\frac{1}{\alpha_k}$ with respect to the iterations for BB1, BB2 (second row), BB2-ℓ_1, ABB$_{\min}$ (third row), ABB$_{\min}$-ℓ_1 (fourth row).

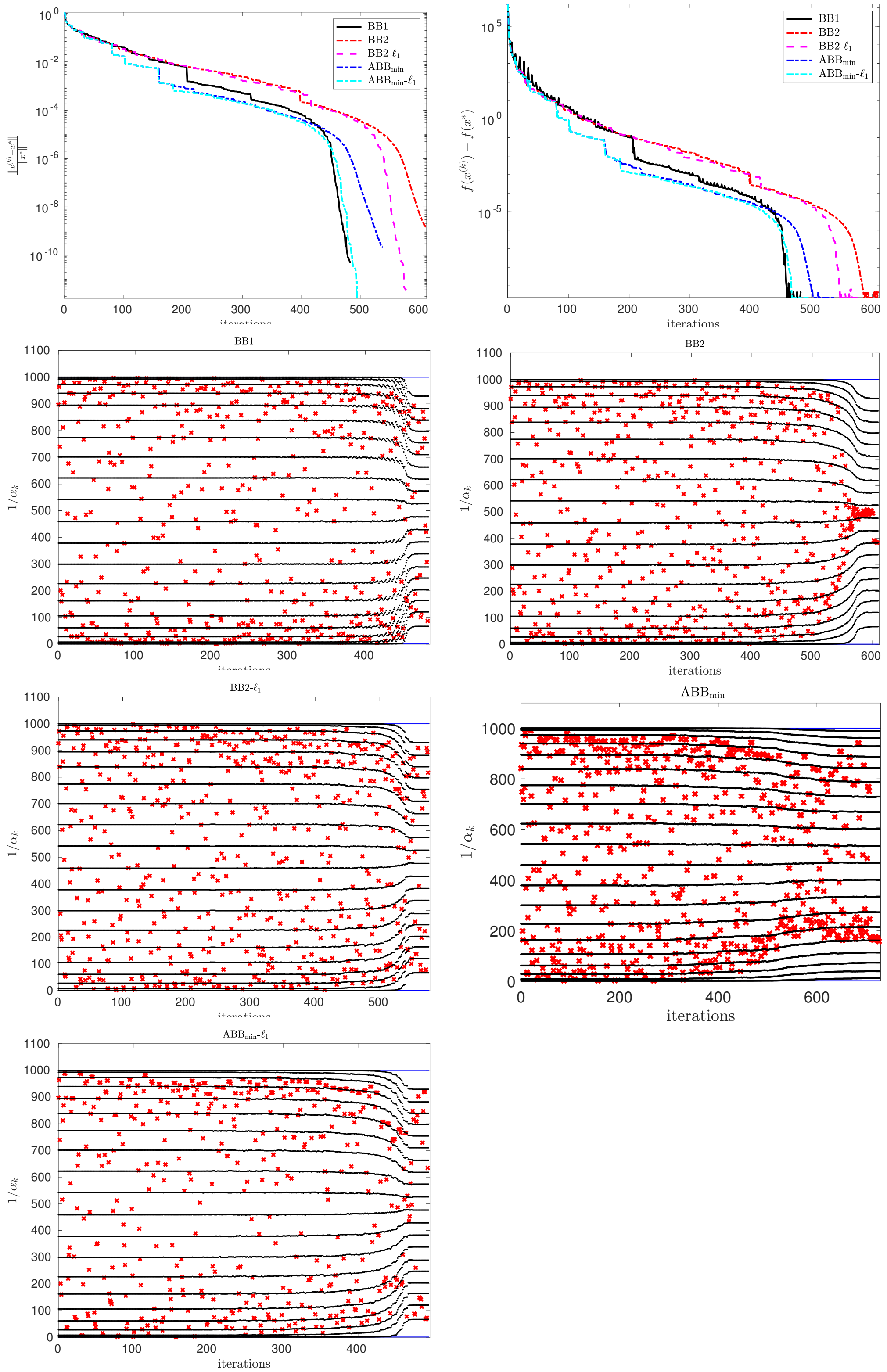

Figure 2: Results on LQP2. Errors on the computed solution and on the objective function for the different rules (first row). Distribution of $\frac{1}{\alpha_k}$ with respect to the iterations for BB1, BB2 (second row), BB2-ℓ_1, ABB$_{\min}$ (third row), ABB$_{\min}$-ℓ_1 (fourth row).

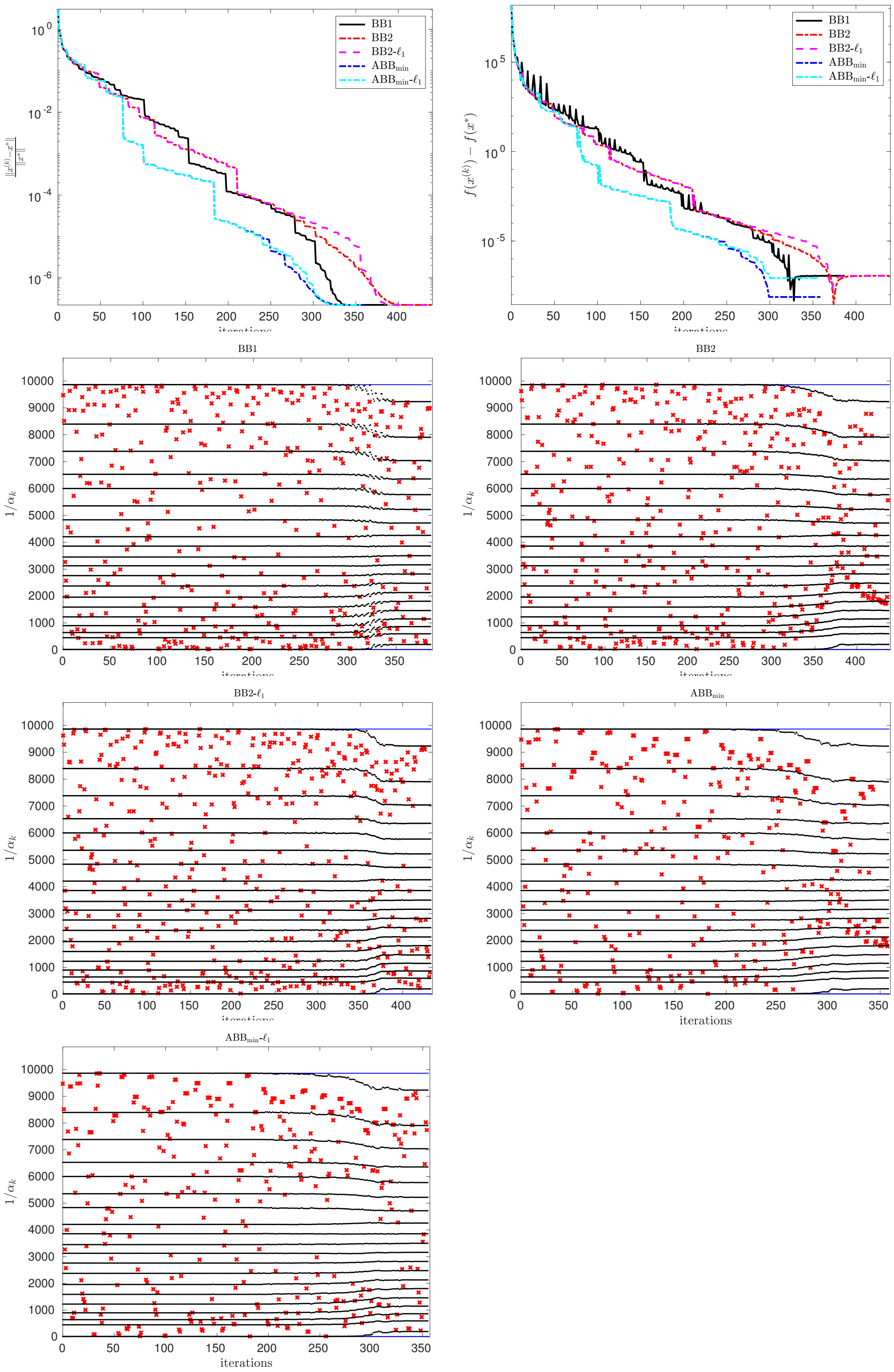

Figure 3: Results on LQP3. Errors on the computed solution and on the objective function for the different rules (first row). Distribution of $\frac{1}{\alpha_k}$ with respect to the iterations for BB1, BB2 (second row), BB2-ℓ_1, ABB$_{\min}$ (third row), ABB$_{\min}$-ℓ_1 (fourth row).

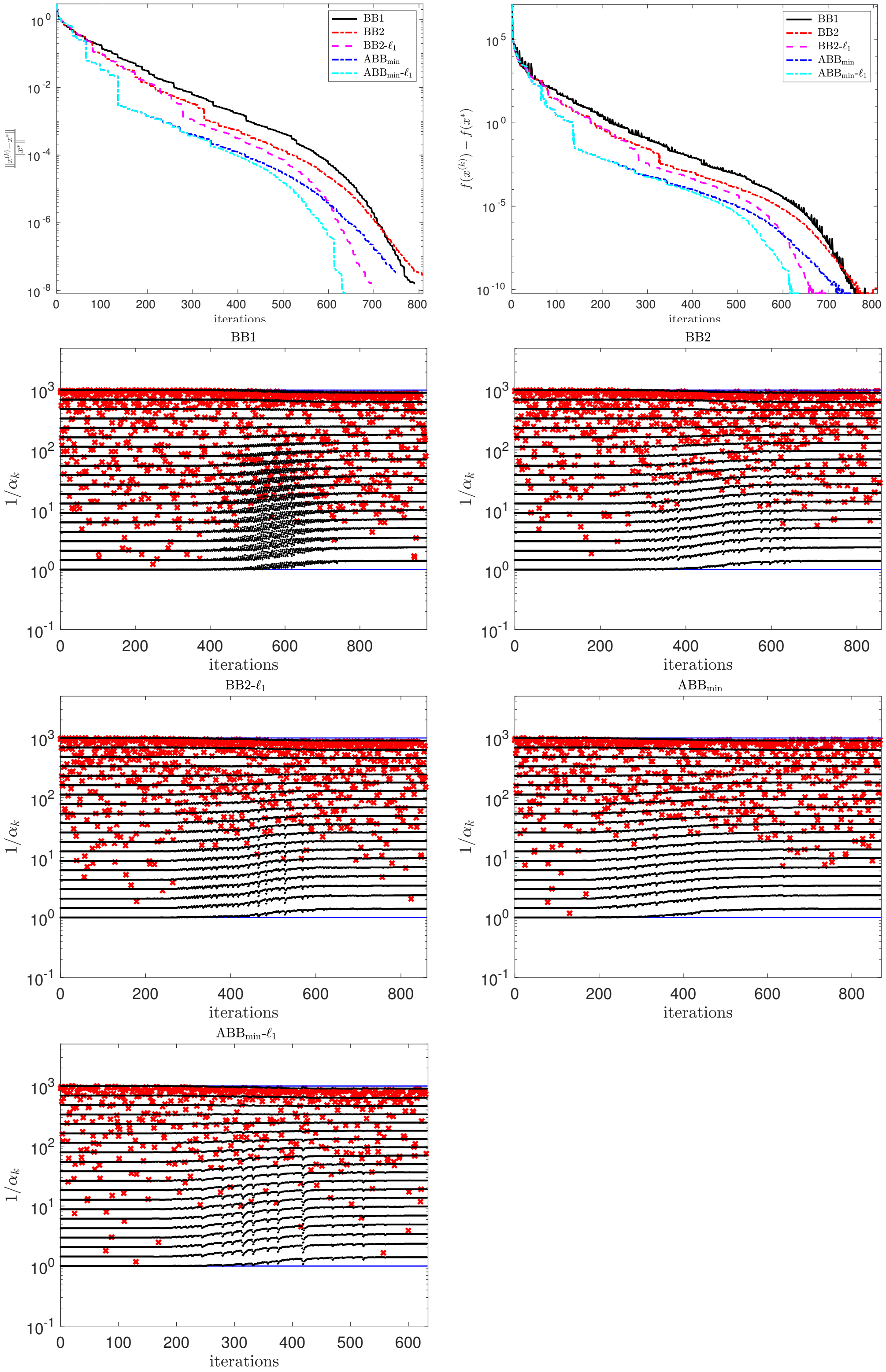

Figure 4: Results on LQP4. Errors on the computed solution and on the objective function for the different rules (first row). Distribution of $\frac{1}{\alpha_k}$ with respect to the iterations for BB1, BB2 (second row), BB2-ℓ_1, ABB$_{\min}$ (third row), ABB$_{\min}$-ℓ_1 (fourth row).

point x^* and set the regularization parameter to $\lambda = 10^{-2}$, and then build a problem having solution x^*. We built a set of problems with the following parameters:

- `n`, number of variables, in $\{10000, 15000, 20000\}$;
- `ncond`, $\log_{10}$ of the Hessian condition number, in $\{4, 5, 6\}$;
- `zerosol`, fraction of zero variables at x^*, in $\{0.25, 0.5, 0.75\}$.

For all the problems we generated a set of four random starting point with fraction of zero entries respectively equal to 0, 0.25, 0.5, 0.75. This process results in a total number of 108 instances onto which we compared the various steplength selection strategies. The results are presented by using the performance profiles proposed by Dolan and Moré [26].

We first run a test on the whole set of instances with the set of parameters described above and the stopping criterion (46), where $\epsilon = 10^{-5}$, together with a maximum number of 2000 iterations.

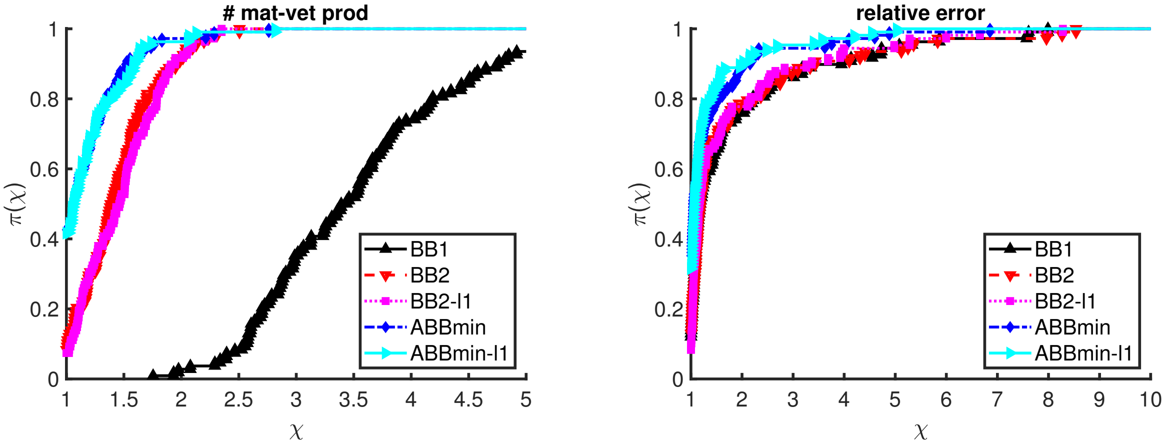

Figure 5: Performance profiles corresponding to matrix-vector products (left panel) and relative error (right panel) on a set of 108 synthetic quadratic ℓ_1-regularized test problems

From the results shown in Figure 5, it is clear how the proposed modified strategies are able to perform comparably in terms of computational cost but have an advantage in terms of relative error with respect to the solution x^*. This suggests that the methods equipped with the proposed ℓ_1-specialized steplengths converge faster towards the solution. To better visualize this, we decided to perform a second test, in which we restricted our focus on the problems with `ncond` in $\{4, 5\}$. We tested the ability of each algorithm to reach a certain neighborhood of the solution, and compared them in terms of matrix-vector products performed. In detail, we run each method until they found a point satisfying the condition

$$\|x^{(k)} - x^*\|_\infty \leq 10^{-2} \cdot \|x^*\|_\infty. \tag{47}$$

or a maximum number of 7500 iterations was reached.

From the results of this test, reported in Figure 6, one can see how the novel strategies are able to outperform the original ones, both in terms of efficiency and robustness, especially when the conditioning of the problems increases. It is worth noting that in the case of problems with condition number equal to 10^5 (right panel), BB1 always failed to reach the desired tolerance in 7500 iterations.

5.2. Comparisons with SpaRSA and GPSR-BB on ℓ_2-ℓ_1 problems

In this subsection, we show a comparison among the versions of Algorithm 1 equipped with BB2-ℓ_1 and ABB$_{\min}$-ℓ_1 and some well-known algorithms emplyoing standard versions of the BB rules, namely SpaRSA [51] and GPSR-BB [28], for which we considered the Matlab implementations freely available at `http://www.lx.it.pt/~mtf/SpaRSA/` and `http://www.lx.it.pt/~mtf/GPSR/`, respectively.

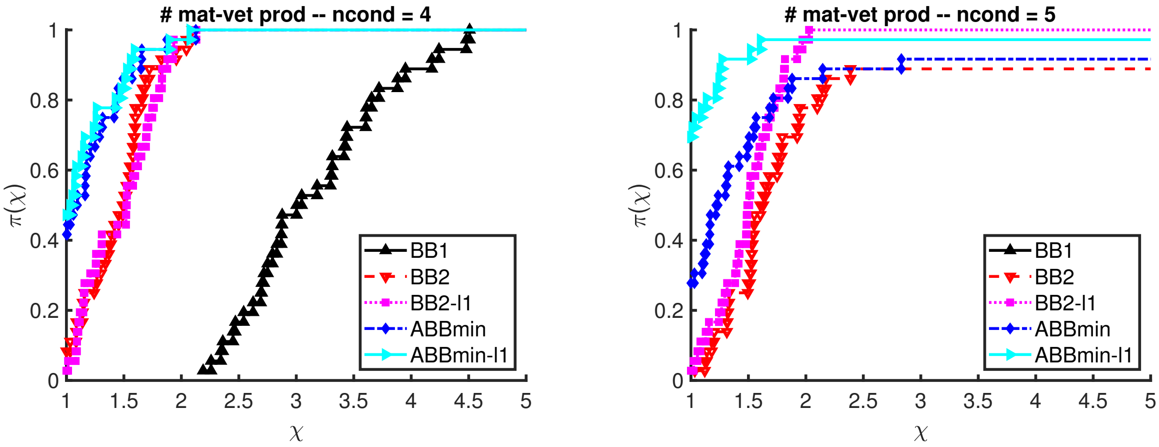

Figure 6: Performance profiles corresponding to matrix-vector products on a subset of synthetic quadratic ℓ_1-regularized test problems generated by setting Hessian condition number equal to 10^4 (left panel) and 10^5 (right panel)

For these experiments, we solve the problem

$$\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{2}\|Ax-y\|^2 + \lambda\|x\|_1,$$

generated as in [51] Section IV-A. In particular, the main features of the test problem are the following:

- $A \in \mathbb{R}^{k\times n}$ is a random matrix with Gaussian i.i.d. entries of zero mean and variance $1/2n$, where $k = 2^{10}$ and $n = 2^{12}$;
- the solution $x^* \in \mathbb{R}^n$ is a vector with 160 randomly placed ±1 spikes, with zeros in the other components, such that its density (fraction of non-zero elements is $\rho = 0.039$;
- $y = Ax^* + e$, where e is a Gaussian white vector with variance 10^{-4};
- regularization parameter: $\lambda = 0.1\|A^\top y\|_\infty$.

Coherently with the test performed in [51], the stopping rule shared by all the methods is

$$f(x^{(k)}) \leq f_b \tag{48}$$

where f_b is a benchmark objective value obtained by running Fixed Point Continuation (FPC) algorithm [35]. The parameters settings for BB2-ℓ_1 and ABB$_{\min}$-ℓ_1 are the same as those considered in the previous experiments, whereas for the parameters of SpaRSA and GPSR-BB we considered the values suggested in the original papers [51] and [28] respectively, except for the line search parameter M in SpaRSA that was set equal to 10 for consistency with BB2-ℓ_1 and ABB$_{\min}$-ℓ_1. The maximum number of iterations was set equal to 10000. The numerical performance of each solver are reported in Table 1, for two different choices of the starting point, i.e. when $x^{(0)}$ is the zero vector and when $x^{(0)}$ is a random vector whose entries are extracted from a standard normal distribution. In particular, we compare the average values over 10 runs of the elapsed time and the mean squared error (MSE); the number of iterations needed to satisfy the stopping criterion and the fraction of non-zero elements of the computed solutions (denoted by ρ) are also reported. The final value of the objective function is approximately equal to 3.52 for all the algorithms. Although the considered methods show comparable performances in terms of computational time, it is interesting to note that the method employing ABB$_{\min}$-ℓ_1 shows slightly better results in terms of MSE values.

Table 1: Elapsed times, iterations, mean squared error (MSE) values and density of the estimate (average over 10 runs)

	times (secs).	iterations	MSE	ρ
		$x^{(0)} = \mathbf{0}$		
SpaRSA	0.02	23	3.044e-03	0.065
GPSR-BB	0.04	25	2.995e-03	0.064
BB2-ℓ_1	0.02	25	2.943e-03	0.069
ABB$_{\min}$-ℓ_1	0.03	28	2.975e-03	0.066
		$x^{(0)}$ random		
SpaRSA	0.03	39	3.017e-03	0.064
GPSR-BB	0.03	37	2.938e-03	0.064
BB2-ℓ_1	0.03	46	3.000e-03	0.064
ABB$_{\min}$-ℓ_1	0.04	50	2.837e-03	0.062

5.3. Total Variation based image deblurring with Gaussian noise

Next, we consider an image deblurring test problem inspired by the Helsinki Deblur Challenge (HDC) 2021 [29]. Our aim is to recover a binary image of some text string from a noisy blurred acquisition. The ground truth image has been generated by randomly picking an image from the HDC dataset (the ones acquired by Camera 1), and resizing and binarizing it. The resulting image has size 83 × 295 and its pixels are zero (black) or one (white). We generate the noisy blurred image by convolving the ground truth with an out-of-focus PSF of radius 7, and then adding Gaussian noise with zero mean and standard deviation 0.01. Following a Maximum A Posteriori approach, it is then feasible to restore the acquired image by addressing the following regularized least squares problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Hx-g\|^2 + \rho TV_\delta(x) + \lambda\|x\|_1, \tag{49}$$

where $H \in \mathbb{R}^{n\times n}$ represents the blurring matrix corresponding to the out-of-focus PSF, $g \in \mathbb{R}^n$ is the noisy blurred image, $\rho > 0$ and $\lambda > 0$ are the regularization parameters, and TV_δ is a smooth approximation of the Total Variation term given by [3]

$$TV_\delta(x) = \sum_{i=1}^{n} \sqrt{(\nabla_i u)_1^2 + (\nabla_i u)_2^2 + \delta^2},$$

where $\delta > 0$ is the smoothing parameter, and $\nabla_i \in \mathbb{R}^{2\times n}$ is the discrete gradient operator at pixel i. Note that the ℓ_1-norm is introduced to enforce sparsity in the image, whereas the TV_δ term is used to preserve the sharp edges of the text string. This test problem is quite interesting for testing the behavior of the proposed BB2-like rule (25), as we expect the solution of problem (49) to possess several active variables. The regularization parameters are set as $\rho = 5\cdot 10^{-5}$, $\delta = 10^{-2}$, $\lambda = 10^{-3}$. A solution of problem (49), denoted by x^*, is approximated by running the FISTA algorithm with constant steplength [2] for 10000 iterations, and the corresponding function value is denoted by f^*.

We apply Algorithm 1 to problem (49) with $f_0(x) = \frac{1}{2}\|Hx-g\|^2 + \rho TV_\delta(x)$ and $f_1(x) = \lambda\|x\|_1$. For this test problem, we compare the performances of Algorithm 1 equipped with the steplength selection strategies BB1, BB2, BB2-ℓ_1, ABB$_{\min}$, and ABB$_{\min}$-ℓ_1. The parameter setting of the resulting algorithms is the same as the previous numerical tests, except for the switching parameter τ in the alternating strategies, which is set equal to 0.6 for ABB$_{\min}$ and equal to 0.8 for ABB$_{\min}$-ℓ_1. This choice is motivated by the fact that the standard BB2 rule may be less reliable than BB2-ℓ_1, as the theoretical spectral analysis prescribed as well as the previous experiments showed, and hence it seems reasonable to differentiate the selected value for the switching parameter, in order to reduce or promote, respectively, the effect of the corresponding rule. We also compared the results with those obtained by using SpaRSA with its default

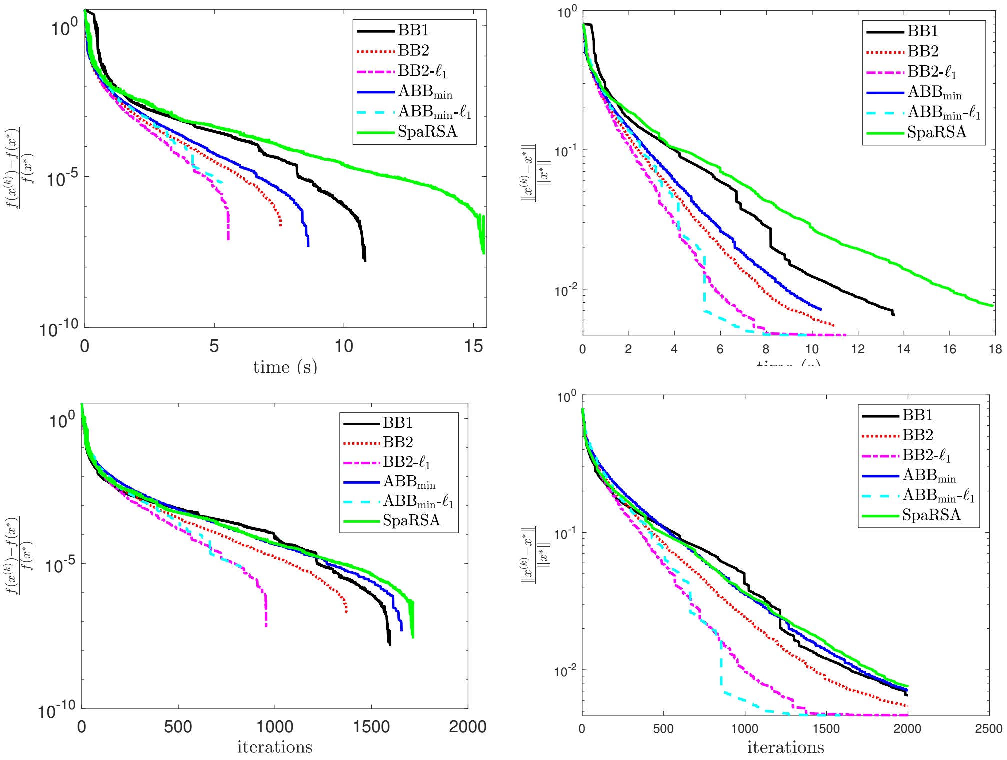

Figure 7: Histories of the relative function errors and the relative minimization errors vs time (top row). Histories of the relative function errors and the relative minimization errors vs iterations (bottom row)

parameters settings and $M = 10$. All the methods, including SpARSA, were stopped when either the stopping criterion (46) was met with $\epsilon = 10^{-8}$, or the maximum number of 2000 iterations was reached.

In Figure 7, we plotted the relative error on the objective function $\frac{f(x^{(k)})-f^*}{f^*}$ and the relative minimization error $\frac{\|x^{(k)}-x^*\|}{\|x^*\|}$ both versus computational time (top row) and iterations (bottom row). Table 2 shows the number of iterations, the relative minimization error and the computational time (in seconds), required by each algorithm to reduce the relative error on the objective function below a prefixed threshold *tol*. Finally, in Figure 8 we reported the original object, its corrupted version, and the reconstructed images provided by the considered methods. These results clearly highlight the accelerating effect achieved by employing the new steplength BB2-ℓ_1 compared to the other steplength selection strategies and SpaRSA. Interestingly, we can observe that the use of the single modified rule BB2-ℓ_1 turned out to be competitive with the alternating strategy ABB$_{\min}$-ℓ_1, which however is able to gain efficiency at the later iterations, improving the accuracy on the reconstruction, due to the combined effect of the two BB rules when the active components start to stabilize. Table 2 shows that the gap between the modified and the original BB steplengths in terms of number of iterations becomes more evident as the threshold *tol* decreases. As a final remark, we can observe that SpaRSA and Algorithm 1 equipped with BB1 turned out to be less efficient compared to the other solvers in terms of per-iteration cost. This is due to the higher numbers of overall backtracking steps performed by the two methods, respectively 888 and 914, compared to the ones needed by BB2, BB2-ℓ_1, ABB$_{\min}$ and ABB$_{\min}$-ℓ_1, which were respectively, 2, 27, 0, and 31.

Table 2: Number of iterations and execution times required to reduce the relative function error below a prefixed tolerance *tol*. The corresponding relative minimization error obtained is reported.

	$tol = 10^{-2}$			$tol = 10^{-4}$			$tol = 10^{-6}$		
	It.	$\frac{\|x^{(k)}-x^*\|}{\|x^*\|}$	Time(s)	It.	$\frac{\|x^{(k)}-x^*\|}{\|x^*\|}$	Time(s)	It.	$\frac{\|x^{(k)}-x^*\|}{\|x^*\|}$	Time(s)
BB1	98	0.260	1.093	996	0.042	6.699	1484	0.012	10.036
BB2	128	0.262	0.072	717	0.048	4.020	1323	0.012	7.308
BB2-ℓ_1	116	0.253	0.734	568	0.048	3.327	935	0.012	5.427
ABB$_{\min}$	161	0.260	0.871	870	0.048	4.567	1581	0.012	8.238
ABB$_{\min}$-ℓ_1	142	0.259	0.899	614	0.048	3.842	854	0.007	5.302
SpaRSA	121	0.260	1.001	859	0.048	7.570	1649	0.012	14.787

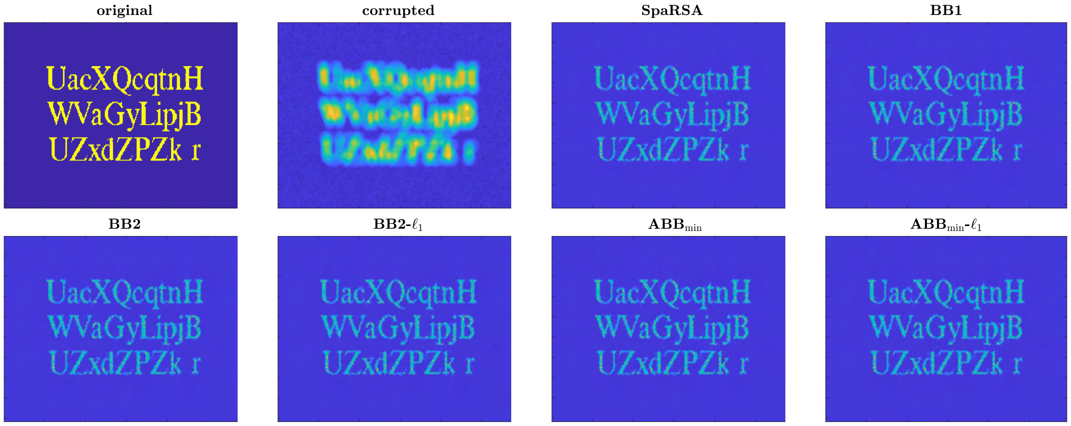

Figure 8: Original object (top-left panel) followed by its corrupted version, the reconstruction obtained by FISTA and the reconstructions obtained by Algorithm 1 equipped with the different steplength selection strategies specified by image title.

6. Conclusions

In this paper, we investigated the spectral properties of the Barzilai-Borwein rules within a proximal gradient scheme for solving ℓ_1-regularized minimization problems. We proposed a modification to the second Barzilai-Borwein rule in order to exploit the optimality conditions along the iterative procedure in a suitable manner, thus obtaining an acceleration of the method. We also provided theoretical results related to the global convergence to a stationary point of a proximal gradient method equipped with a nonmonotone line search. Numerical results on quadratic ℓ_1-regularized test problems confirmed the theoretical analysis on the spectral behavior of the Barzilai-Borwein rules, and provided evidences of the gain in terms of efficiency obtained by employing the modified rule. These advantages are also confirmed by preliminary numerical results on an image restoration problem, where the performance obtained using the single modified rule appears competitive with those of the adaptive alternating strategies. Comparisons with two state-of-the-art solvers for ℓ_1 regularized problems on a synthetic test and on a real application are presented, showing comparable or superior behaviour of the proposed methods. We plan to further investigate the numerical impact of the proposed steplengths on large-scale minimization problems arising in signal and image processing, as well as extend our spectral analysis to more general regularizers.

Acknowledgements

This work has been partially supported by the INDAM research group GNCS. S. Crisci aknowledges the support from the EU-FESR PON Ricerca e Innovazione 2014-2020, art. 24, comma 3, lett. a) L. 240/2010 e s.m.i., D.M. 1062/2021.

Disclosure statement

The authors report there are no competing interests to declare.

References

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

[2] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

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

[4] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optim.*, 10:1196–1211, 2000.

[5] E. G. Birgin, J. M. Martínez, and M. Raydan. Spectral projected gradient methods: review and perspectives. *J. Stat. Softw.*, 60:1–21, 2014.

[6] S. Bonettini, I. Loris, F. Porta, and M. Prato. Variable metric inexact line-search-based methods for nonsmooth optimization. *SIAM J. Optim.*, 26(2):891–921, 2016.

[7] S. Bonettini, I. Loris, F. Porta, M. Prato, and S. Rebegoldi. On the convergence of a linesearch based proximal-gradient method for nonconvex optimization. *Inverse Probl.*, 33(5):055005, 2017.

[8] S. Bonettini, F. Porta, M. Prato, S. Rebegoldi, V. Ruggiero, and L. Zanni. *Recent Advances in Variable Metric First-Order Methods*, pages 1–31. Springer International Publishing, Cham, 2019.

[9] S. Bonettini, M. Prato, and S. Rebegoldi. New convergence results for the inexact variable metric forward–backward method. *Appl. Math. Comput.*, 392:125719, 2021.

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

[11] L. Bottou, F. C. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.*, 60(2):223–311, 2018.

[12] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[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(1):107–132, July 2014.

[14] P. L. Combettes and V. R. Wajs. Signal Recovery by Proximal Forward-Backward Splitting. *Multiscale Model. Simul.*, 4(4):1168–1200, 2005.

[15] S. Corsaro, V. De Simone, Z. Marino, and F. Perla. ℓ_1-regularization for multi-period portfolio selection. *Ann. Oper. Res.*, 294(1):75–86, 2020.

[16] S. Crisci, F. Porta, V. Ruggiero, and L. Zanni. Spectral properties of Barzilai-Borwein rules in solving singly linearly constrained optimization problems subject to lower and upper bounds. *SIAM J. Optim.*, 30(2):1300–1326, 2020.

[17] S. Crisci, F. Porta, V. Ruggiero, and L. Zanni. Hybrid limited memory gradient projection methods for box-constrained optimization problems. *Comput. Optim. Appl.*, pages 1–39, 2022.

[18] S. Crisci, F. Porta, V. Ruggiero, and L. Zanni. On the convergence properties of scaled gradient projection methods with non-monotone armijo–like line searches. *Ann. Univ. Ferrara*, 68(2):521–554, 2022.

[19] S. Crisci, V. Ruggiero, and L. Zanni. Steplength selection in gradient projection methods for box-constrained quadratic programs. *Appl. Math. Comput.*, 356:312–327, 2019.

[20] Y. Dai and Y.-X. Yuan. Analysis of monotone gradient methods. *J. Ind. Manag. Optim.*, 1(2):181, 2005.

[21] Y. H. Dai and R. Fletcher. Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming. *Numer. Math.*, 100:21–47, 2005.

[22] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57:1413–1457, 2004.

[23] A. De Asmundis, D. di Serafino, H. Hager, G. Toraldo, and H. Zhang. An efficient gradient method using the Yuan steplength. *Comput. Optim. Appl.*, 59(3):541–563, 2014.

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

[25] D. di Serafino, G. Toraldo, M. Viola, and J. Barlow. A two-phase gradient method for quadratic programming problems with a single linear constraint and bounds on the variables. *SIAM J. Optim.*, 28(4):2809–2838, 2018.

[26] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Math. Program. Ser. B*, 91(2):201–213, 2002.

[27] O. P. Ferreira, M. Lemes, and L. F. Prudente. On the inexact scaled gradient projection method. *Comput. Optim. Appl.*, 81:91–125, 2022.

[28] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top.*, 1(4):586–597, 2007.

[29] Finnish Inverse Problems Society (FIPS). Helsinki Deblur Challenge 2021 (HDC2021). `https://fips.fi/HDC2021.php`, 2021.

[30] R Fletcher. Low storage methods for unconstrained optimization. *Lectures in Appl. Math.*, 26:165–179, 1990.

[31] R. Fletcher. On the Barzilai-Borwein method. In L. Qi, K. Teo, X. Yang, P. M. Pardalos, and D. Hearn, editors, *Optimization and Control with Applications*, volume 96 of *Applied Optimization*, pages 235–256. Springer, US, 2005.

[32] G. Frassoldati, G. Zanghirati, and L. Zanni. New adaptive stepsize selections in gradient methods. *J. Ind. Manag. Optim.*, 4(2):299–312, 2008.

[33] C. C. Gonzaga and R. M. Schneider. On the steepest descent algorithm for quadratic functions. *Comput. Optim. Appl.*, 63(2):523–542, 2016.

[34] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for newton's method. *SIAM journal on Numerical Analysis*, 23(4):707–716, 1986.

[35] Elaine T Hale, Wotao Yin, and Yin Zhang. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. *CAAM TR07-07, Rice University*, 43:44, 2007.

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

[37] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, New York, NY, USA, 2nd edition, 2012.

[38] Y. Huang, Y.-H. Dai, X.-W. Liu, and H. Zhang. On the acceleration of the Barzilai–Borwein method. *Comput. Optim. Appl.*, 81(3):717–740, 2022.

[39] Y. Huang and H. Liu. A Barzilai-Borwein type method for minimizing composite functions. *Numer. Algor.*, 69:819–838, 2015.

[40] C.-P. Lee and S. J. Wright. Inexact successive quadratic approximation for regularized optimization. *Comput. Optim. Appl.*, 72(3):641–674, 2019.

[41] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb.*, 1(4):457, 1967.

[42] J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *C. R. Acad. Sci. Ser. A*, 255:2897–2899, 1962.

[43] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer, 1999.

[44] L. Pospíšil and Z. Dostál. The projected Barzilai–Borwein method with fall-back for strictly convex qcqp problems with separable constraints. *Math. Comput. Simul.*, 145:79–89, 2018.

[45] M. Raydan. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM J. Optim.*, 7(1):26–33, 1997.

[46] R. T. Rockafellar, R. J.-B. Wets, and M. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften.* Springer, Berlin, 1998.

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

[48] S. Solntsev, J. Nocedal, and R. H. Byrd. An algorithm for quadratic ℓ_1-regularized optimization with a flexible active-set strategy. *Optim. Methods Softw.*, 30(6):1213–1237, 2015.

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

[50] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM J. Sci. Comput.*, 32(4):1832–1857, 2010.

[51] S. Wright, R. Nowak, and M. Figueiredo. Sparse Reconstruction by Separable Approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

[52] Y.-X. Yuan. Step-sizes for the gradient method. *AMS/IP Stud. Adv. Math.*, 42(2):785, 2008.

[53] A. Zalinescu. *Convex analysis in general vector spaces.* World Scientific Publishing Co. Inc., River Edge, NJ, 2002.

[54] B. Zhou, L. Gao, and Y. H. Dai. Gradient methods with adaptive step-sizes. *Comput. Optim. Appl.*, 35(1):69–86, 2006.

Appendix A. Proofs of Lemma 4.1 and Theorem 4.2

Proof. [Proof of Lemma 4.1] The proof runs similarly to [6, Proposition 3.1]. By contradiction, assume that there exists $k \geq 0$ such that the line search performs an infinite number of reductions. Therefore, for all $j \geq 0$, we have

$$\begin{aligned}
\sigma h_\gamma^{(k)}(y^{(k)}, x^{(k)}) &< \frac{f(x^{(k)} + \beta^j d^{(k)}) - \bar{f}_k}{\beta^j} \\
&= \frac{f_0(x^{(k)} + \beta^j d^{(k)}) - f_0(x^{(k)})}{\beta^j} + \frac{f_1(x^{(k)} + \beta^j d^{(k)}) - f_1(x^{(k)})}{\beta^j} - \frac{\xi_k}{\beta^j} \\
&\leq \frac{f_0(x^{(k)} + \beta^j d^{(k)}) - f_0(x^{(k)})}{\beta^j} + \frac{\beta^j f_1(y^{(k)}) + (1-\beta^j) f_1(x^{(k)}) - f_1(x^{(k)})}{\beta^j} \\
&= \frac{f_0(x^{(k)} + \beta^j d^{(k)}) - f_0(x^{(k)})}{\beta^j} + f_1(y^{(k)}) - f_1(x^{(k)}),
\end{aligned}$$

where the second inequality follows from an application of Jensen's inequality to the convex function f_1, and the fact that $\xi_k \geq 0$ for all $k \geq 0$. Taking the limit on the right-hand side for $j \to \infty$ yields

$$\begin{aligned}
\sigma h_\gamma(y^{(k)}, x^{(k)}) &\leq \nabla f_0(x^{(k)})^T d^{(k)} + f_1(y^{(k)}) - f_1(x^{(k)}) \\
&\leq \nabla f_0(x^{(k)})^T d^{(k)} + f_1(y^{(k)}) - f_1(x^{(k)}) + \frac{\gamma}{2\alpha_k} \|y^{(k)} - x^{(k)}\|^2 \\
&= h_\gamma(y^{(k)}, x^{(k)}) < 0,
\end{aligned}$$

where the last equality is due to the definition of $h_\gamma(\cdot; x^{(k)})$ in (30), and the last inequality is assumed from (35). Since $\sigma \in (0, 1)$, this is absurd. □

Before providing the proof of Theorem 4.2, we make the following useful remark. Suppose that $f = f_0 + f_1$ with f_0, f_1 satisfying the assumptions of Section 4. By Definition 2.2, a point $\bar{x}$ is stationary for f if and only if $0 \in \partial f(\bar{x})$. Then, we can apply Lemma 2.2 and Definition 2.3 to write the following relations

$$\begin{aligned} \bar{x} \text{ is stationary for } f \quad & \Leftrightarrow \quad 0 \in \partial f(\bar{x}) \\ & \Leftrightarrow \quad -\alpha \nabla f_0(\bar{x}) \in \partial(\alpha f_1)(\bar{x}), \quad \forall\ \alpha > 0 \\ & \Leftrightarrow \quad \bar{x} = \operatorname{prox}_{\alpha f_1}(\bar{x} - \alpha \nabla f_0(\bar{x})), \quad \forall\ \alpha > 0. \end{aligned}$$

Proof. [Proof of Theorem 4.2] The proof is obtained by slightly modifying the arguments in [6, Theorem 3.1].

If there exists $\bar{k} \geq 0$ such that $h_\gamma^{(k)}(y^{(\bar{k})}, x^{(\bar{k})}) = 0$, then $x^{(\bar{k})}$ is stationary [6, Proposition 2.3] and $x^{(\bar{k}+j)} = x^{(\bar{k})}$ for all $j \geq 0$, hence the thesis holds.

Otherwise, assume that $h_\gamma(y^{(k)}, x^{(k)}) < 0$ for all $k \in \mathbb{N}$. Then the sequence $\{x^{(k)}\}_{k \in \mathbb{N}}$ is well-defined (due to Lemma 4.1) and infinite. Since $\{x^{(k)}\}_{k \in K}$ converges, it is a bounded sequence. Due to the continuity of the operator $p(x, \alpha) = \operatorname{prox}_{\alpha f_1}(x - \alpha \nabla f_0(x))$ with respect to its arguments, together with the boundedness of the sequences $\{x^{(k)}\}_{k \in K}$ and $\{\alpha_k\}_{k \in K}$, it follows that also $\{y^{(k)}\}_{k \in K}$ is bounded, hence it admits a limit point $\bar{y}$. Let $K' \subseteq K$ be such that $\lim_{k \in K', k \to \infty} y^{(k)} = \bar{y}$ and $\lim_{k \to \infty} \alpha_k = \bar{\alpha} > 0$. By continuity of $p(x, \alpha)$, we deduce that $\bar{y} = \operatorname{prox}_{\bar{\alpha} f_1}(\bar{x} - \bar{\alpha} \nabla f_0(\bar{x}))$.

Next, we note that condition (37) can be equivalently as

$$f(x^{(k+1)}) + \xi_{k+1} + \delta_{k+1}(f(x^{(k)}) + \xi_k - f(x^{(k+1)})) \leq f(x^{(k)}) + \xi_k, \tag{A.1}$$

which implies that the sequence $\{f(x^{(k)}) + \xi_k\}_{k \in \mathbb{N}}$ is monotone nonincreasing. Given that (36) holds, this means that there exists $\bar{f} \in \bar{\mathbb{R}}$ such that $\lim_{k \to \infty} f(x^{(k)}) + \xi_k = \lim_{k \to \infty} f(x^{(k)}) = \bar{f}$. Since f is lower semicontinuous and $\bar{x}$ is a limit point of $\{x^{(k)}\}_{k \in \mathbb{N}}$, we have

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

From the previous inequality, it follows that $\bar{f} \in \mathbb{R}$, which in turn gives

$$\lim_{k \to \infty} f(x^{(k)}) - f(x^{(k+1)}) = 0. \tag{A.2}$$

Note that the Armijo-like condition (32) can be rewritten as

$$0 \leq -\nu_k h_\gamma(y^{(k)}, x^{(k)}) \leq \frac{f(x^{(k)}) + \xi_k - f(x^{(k+1)})}{\sigma}.$$

Then, taking the limit for $k \to \infty$ and employing (A.2) and (36) yields

$$\lim_{k \to \infty} \nu_k h_\gamma(y^{(k)}, x^{(k)}) = 0. \tag{A.3}$$

We are now ready to show that

$$\lim_{k \in K', k \to \infty} h_\gamma(y^{(k)}, x^{(k)}) = 0. \tag{A.4}$$

To this aim, we first note that $\{h_\gamma(y^{(k)}, x^{(k)})\}_{k \in K'}$ is bounded from below. Indeed, we can write

$$\begin{aligned} h_\gamma(y^{(k)}, x^{(k)}) &= \nabla f_0(x^{(k)})^T (y^{(k)} - x^{(k)}) + \frac{\gamma}{2\alpha_k} \|y^{(k)} - x^{(k)}\|^2 + f_1(y^{(k)}) - f_1(x^{(k)}) \\ &\geq \nabla f_0(x^{(k)})^T (y^{(k)} - x^{(k)}) + f_1(y^{(k)}) - f_1(x^{(k)}) \\ &= \nabla f_0(x^{(k)})^T (y^{(k)} - x^{(k)}) + f_1(y^{(k)}) - f(x^{(k)}) + f_0(x^{(k)}) \\ &\geq \nabla f_0(x^{(k)})^T (y^{(k)} - x^{(k)}) + f_1(y^{(k)}) - f(x^{(k)}) - \xi_k + f_0(x^{(k)}) \\ &\geq \nabla f_0(x^{(k)})^T (y^{(k)} - x^{(k)}) + f_1(y^{(k)}) - f(x^{(0)}) - \xi_0 + f_0(x^{(k)}), \end{aligned}$$

where the last inequality is due to the fact that $\{f(x^{(k)}) + \xi_k\}_{k\in\mathbb{N}}$ is monotone nonincreasing. Since f_1 is proper and convex, it admits a supporting hyperplane, namely, there exist $a, b \in \mathbb{R}^n$ such that $f_1(u) \geq a^T u + b$ for all $u \in \mathbb{R}^n$. Hence

$$h_\gamma(y^{(k)}, x^{(k)}) \geq \nabla f_0(x^{(k)})^T(y^{(k)} - x^{(k)}) + a^T y^{(k)} + b - f(x^{(0)}) - \xi_0 + f_0(x^{(k)}).$$

Since the function $\varphi(x, y) = \nabla f_0(x)^T(y - x) + a^T y + b - f(x^{(0)}) - \xi_0 + f_0(x)$ is a continuous function with respect to its variables x, y, and the sequences $\{x^{(k)}\}_{k\in K'}$ and $\{y^{(k)}\}_{k\in K'}$ are closed and bounded, we conclude that the sequence $\{h_\gamma(y^{(k)}, x^{(k)})\}_{k\in K'}$ is bounded from below.

Assume by contradiction that (A.4) does not hold. Then, since $\{h_\gamma(y^{(k)}, x^{(k)})\}_{k\in K'}$ is a bounded sequence, there must exist $K'' \subseteq K'$ such that $\lim_{k\in K'', k\to\infty} h_\gamma(y^{(k)}, x^{(k)}) = \bar{h} < 0$. By (A.3), this implies that

$$\lim_{k\in K'', k\to\infty} \nu_k = 0. \tag{A.5}$$

This means that, for all sufficiently large $k \in K''$, the linesearch based on (32) performs at least one reduction. In other words, for all sufficiently large $k \in K''$, we have

$$\sigma(\nu_k/\delta) h_\gamma(y^{(k)}, x^{(k)}) < f(x^{(k)} + (\nu_k/\delta) d^{(k)}) - f(x^{(k)}) - \xi_k.$$

Repeating the same arguments as in the proof of Lemma 4.1, we get

$$\begin{aligned}\sigma h_\gamma(y^{(k)}, x^{(k)}) &< \frac{f_0(x^{(k)} + (\nu_k/\delta) d^{(k)}) - f_0(x^{(k)})}{\nu_k/\delta} + f_1(y^{(k)}) - f_1(x^{(k)}) \\ &\leq \frac{f_0(x^{(k)} + (\nu_k/\delta) d^{(k)}) - f_0(x^{(k)})}{\nu_k/\delta} + f_1(y^{(k)}) - f_1(x^{(k)}) + \frac{\gamma}{2\alpha_k}\|y^{(k)} - x^{(k)}\|^2.\end{aligned}$$

Taking the limit on both sides for $k \in K'', k \to \infty$, since $\{d^{(k)}\}_{k\in K''}$ is bounded and by (A.5), we obtain $\sigma\bar{h} \leq \bar{h} < 0$, which is absurd, being $\sigma \in (0, 1)$. Thus, the limit (A.4) holds.

Finally, setting $x = x^{(k)}$, $y = y^{(k)}$, $w = 0 \in \partial h_1^{(k)}(y^{(k)}, x^{(k)})$ inside inequality (41) leads to

$$\frac{m}{2}\|y^{(k)} - x^{(k)}\|^2 \leq -h_1^{(k)}(y^{(k)}, x^{(k)}) + h_1^{(k)}(x^{(k)}, x^{(k)}) = -h_1^{(k)}(y^{(k)}, x^{(k)}) \underset{k\in K'', k\to\infty}{\longrightarrow} 0.$$

Therefore, we have proved that $\bar{x} = \bar{y} = \mathrm{prox}_{\bar{\alpha} f_1}(\bar{x} - \bar{\alpha}\nabla f_0(\bar{x}))$, which implies that $\bar{x}$ is stationary. □