On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions

José Yunier Bello Cruz*

November 17, 2014

Abstract

In this paper we present a variant of the proximal forward-backward splitting method for solving nonsmooth optimization problems in Hilbert spaces, when the objective function is the sum of two nondifferentiable convex functions. The proposed iteration, which will be call the Proximal Subgradient Splitting Method, extends the classical projected subgradient iteration for important classes of problems, exploiting the additive structure of the objective function. The weak convergence of the generated sequence was established using different stepsizes and under suitable assumptions. Moreover, we analyze the complexity of the iterates.

Keywords: Convex problems; Nonsmooth optimization problems; Proximal forward-backward splitting iteration; Subgradient method.

Mathematical Subject Classification (2010): 65K05, 90C25, 90C30.

1 Introduction

The purpose of this paper is to study the convergence properties of a variant of the proximal forward-backward splitting method for solving optimization problems of the following form:

$$\min\ f(x) + g(x) \ \text{ s.t. } \ x \in \mathcal{H}, \tag{1}$$

where $\mathcal{H}$ is a nontrivial real Hilbert space, and $f : \mathcal{H} \to \bar{\mathbb{R}}$ and $g : \mathcal{H} \to \bar{\mathbb{R}}$ are two proper lower semicontinuous and convex functions. We are interested in the case where both functions f and g are nondifferentiable, and the domain of f is an open subset of $\mathcal{H}$ containing the domain of g. The solution set of this problem will be denoted by S_*, which is a closed and convex subset of the domain of g. Problem (1) has recently been received much of attention, because it has broad applications to several different areas such as control, signal processing, system identification, machine learning and restoration of images; see, for instance, [17, 18, 23, 29] and the references therein.

A special case of problem (1) is the nonsmooth constrained optimization problem, taking $g = \delta_C$ where δ_C is the indicator function of a nonempty closed and convex set C in $\mathcal{H}$, defined by $\delta_C(y) :=$

*Institute of Mathematics and Statistics, Federal University of Goiás, Goiânia, G.O. 74001-970, Brazil. E-mails: yunier@impa.br & yunier@ufg.br.
This work was partially supported by CNPq grants 303492/2013-9, 474160/2013-0 and 202677/2013-3 and by projects CAPES-MES-CUBA 226/2012 and UNIVERSAL FAPEG/CNPq.

0, if $y \in C$ and $+\infty$, otherwise. Then, problem (1) becomes in the constrained minimization problem

$$\min \; f(x) \; \text{ s.t. } \; x \in C. \tag{2}$$

Another important case of problem (1), which has had much interest in signal denoising and data mining, is the following optimization problem with ℓ_1-regularization

$$\min \; f(x) + \lambda \|x\|_1 \; \text{ s.t. } \; x \in \mathcal{H}, \tag{3}$$

where $\lambda > 0$ and the norm $\|\cdot\|_1$ is used to induce the sparsity in the solutions. Moreover, problem (3) contains the important and well studied Low-Rank problem, when $\mathcal{H} = \mathbb{R}^n$ and $f(x) = \|Ax - b\|_2^2$ where $A \in \mathbb{R}^{m \times n}$, $m << n$, and $b \in \mathbb{R}^m$, which is just a convex approximation of the very famous ℓ_0 minimization problem; see [11]. Recently, this problem became popular in signal processing and statistical inference; see, for instance, [21, 40].

We focus here our attention on the so-called proximal forward-backward splitting iteration [29], which contains a forward gradient step of f (an explicit step) followed by a backward proximal step of g (an implicit step). The main idea of our approach consists of replacing, in the forward step of the proximal forward-backward splitting iteration, the gradient of f by a subgradient of f (note that here f is assumed nondifferentiable in general). In the particular case that g is the indicator function, the proposed iteration becomes to the classical projected subgradient iteration.

To describe and motivate our iteration, first we recall the definition of the so-called proximal operator as $\operatorname{prox}_g : \mathcal{H} \to \mathcal{H}$ where $\operatorname{prox}_g(x)$, $x \in \mathcal{H}$ is a unique solution of the following optimization problem

$$\min \; g(y) + \frac{1}{2}\|x - y\|^2 \; \text{ s.t. } \; y \in \mathcal{H}. \tag{4}$$

The proximal operator prox_g is well-defined and has attractive properties, e.g., it is continuous and firmly nonexpansive; for other properties and algebra rules see [3, 17, 18]. If $g = \delta_C$ is the indicator function, the orthogonal projection onto C, $P_C(x) := \{y \in C : \|x - y\| = \operatorname{dist}(x, C)\}$ is the same that $\operatorname{prox}_{\delta_C}(x)$ for all $x \in \mathcal{H}$; see, for instance, [2]. Now, let us recall the definition of the subdifferential operator $\partial g : \mathcal{H} \rightrightarrows \mathcal{H}$ by $\partial g(x) := \{w \in \mathcal{H} : g(y) \geq g(x) + \langle w, y - x \rangle, \; \forall \, y \in \mathcal{H}\}$. We also present the proximal operator $\operatorname{prox}_{\alpha g}$ through of its relation with the subdifferential operator ∂g, i.e., $\operatorname{prox}_{\alpha g} = (I + \alpha \partial g)^{-1}$ and as a direct consequence of the first optimality condition of (4), we have a useful inclusion:

$$\frac{z - \operatorname{prox}_{\alpha g}(z)}{\alpha} \in \partial g(\operatorname{prox}_{\alpha g}(z)), \tag{5}$$

for any $z \in \mathcal{H}$ and $\alpha > 0$. The iteration proposed here called the Proximal Subgradient Splitting Method for solving problem (1), is motivated by the well-known fact that $x \in S_*$ if and only if exists $u \in \partial f(x)$ such that $x = \operatorname{prox}_{\alpha g}(x - \alpha u)$. Thus, the iteration generalizes the proximal forward-backward splitting iteration for the differentiable case, as a fixed point iteration, defined as the following form: beginning with x^0 belonging to the domain of g, set

$$x^{k+1} = \operatorname{prox}_{\alpha_k g}(x^k - \alpha_k u^k), \tag{6}$$

where $u^k \in \partial f(x^k)$ and the stepsize α_k is positive for all $k \in \mathbb{N}$.

Iteration (6) covers important situations in which f is not differentiable. The nondifferentiability of the function f has a direct impact on the computational effort and the importance of such problems is underlined because they occur frequently in applications. Actually, nondifferentiability arises, for instance, in the problem of minimizing the total variation of a signal over a convex set, in the problem of minimizing the sum of two set-distance functions, in problems involving maxima of convex functions, the Dantzing selector-type problems, the non-Gaussian image denoising problem and in Tykhonov regularization problems with L_1 norms; see, for instance, [12,16,26]. The iteration of the proximal subgradient splitting method, proposed in (6), solves these important instances, extending the classical projected subgradient iteration for solving (2).

Within of problem (1), f is usually assumed to be differentiable, the convergence of the iteration (6) to a solution of (1) has been established in the literature, when the gradient of f is globally Lipschitz continuous. Moreover, the stepsizes α_k have to be chosen very small or less than some constant related with the available Lipschitz constant or throughout of a linesearch; see, for instance, [5, 9, 18, 29, 32]. It is important to mention that the forward-backward iteration finds also application for solving more general problems, like the variational inequality and inclusion problems; see, for instance, [8, 10, 13, 14, 39] and the references therein. On the other hand, the standard convergence analysis for this iteration requires at least a co-coercivity assumption and the stepsizes into a suitable interval; see, for instance, Theorem 25.8 of [3]. Note that co-coercive operators are monotone and Lipschitz continuous, but the converse does not hold in general; see [41]. Although, for gradients of lower semicontinuous, proper and convex functions, the co-coercivity is equivalent to the global Lipschitz continuity assumption. This nice and surprising fact, used strongly in the convergence analysis of the proximal forward-backward method for problem (1), when f is differentiable, is known as the Baillon-Haddad Theorem; see Corollary 18.16 of [3].

The main aim of this work is to release the differentiability on f of the forward-backward splitting method, extending the classical projected subgradient method for problem (2) and containing, as particular case, a new proximal subgradient iteration for problem (3). Note that, in general, for evaluating of the proximal operator is necessary to solve a strongly convex minimization problem. Thus, in the context of problem (1), we assume that to evaluate the proximal operator of f is very hard, leaving out the possibility to use the Douglas-Rachford splitting iteration presented in [16]. The proposed iteration here uses the proximal operator of g and the explicit subgradient iteration of f (the proximal operator of f is never evaluated), which is much easier to implement than the proximal operator of $f+g$ or f as in the standard proximal point algorithm or the Douglas-Rachford splitting iteration, respectively for nonsmooth problems, like (1); see, for instance, [14, 16].

This work is organized as follows. The next subsection provides our notations and assumptions, and some preliminaries results that will be used in the remainder of this paper. The proximal subgradient splitting method and its weak convergence are analyzed by choosing different stepsizes in Section 2. Finally, Section 3 gives some concluding remarks.

1.1 Assumptions and Preliminaries

In this section, we present our assumptions, a classical definition and some results needed for the convergence analysis of the proposed method.

We begin reminded some definitions and notation used in this paper, which is standard and follows [3, 38]. Throughout this paper, we write $p := q$ to indicate that p is defined to be equal to q. The inner product in $\mathcal{H}$ is denoted by $\langle \cdot, \cdot \rangle$, and the norm induced by this inner product, by $\| \cdot \|$, i.e., $\|x\| := \sqrt{\langle x, x \rangle}$ for all $x \in \mathcal{H}$. We write $\mathbb{N}$ for the nonnegative integers $\{0, 1, 2, \dots\}$ and the extended-real number system is $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. The closed ball centered at $x \in \mathcal{H}$ with radius $\rho > 0$ will be denoted by $\mathbb{B}[x; \rho]$, *i.e.*, $\mathbb{B}[x; \rho] := \{y \in \mathcal{H} \colon \|y - x\| \leq \rho\}$. The domain of any function $h : \mathcal{H} \to \bar{\mathbb{R}}$, denoted by $dom(h)$, is defined as $dom(h) := \{x \in \mathcal{H} : h(x) < +\infty\}$. The optimal value of problem (1) will be denoted by $s_* := \inf\{(f + g)(x) : x \in \mathcal{H}\}$. Finally, $\ell_1(\mathbb{N})$ denotes the set of summable sequences in $[0, +\infty)$.

Throughout this paper we assume the following:

A1. ∂f is bounded on bounded sets on the domain of g, i.e., $\exists\,\zeta > 0$ such that $\partial f(x) \subseteq \mathbb{B}[0;\zeta]$ for all $x \in V$, where V is any bounded subset of $dom(g)$.

A2. ∂g has bounded elements on the domain of g, i.e., $\exists\,\rho > 0$ such that $\partial g(x) \cap \mathbb{B}[0;\rho] \neq \emptyset$ for all $x \in dom(g)$.

In connection with Assumption **A1**, we recall that ∂f is locally bounded on its open domain. In finite dimension spaces, this result implies that **A1** always holds. Furthermore, the boundedness of the subgradients is crucial for the convergence analysis of many classical subgradient methods in Hilbert spaces and it has been widely considered in the literature; see, for instance, [1, 7, 8, 35]. Regarding to Assumption **A2**, we emphasize that it holds trivially for important instance of problem (1), e.g., problems (2) and (3), or when $dom(g)$ is a bounded set or when $\mathcal{H}$ is a finite dimensional space. Note that Assumption **A2** even allows instances where ∂g is an unbounded set as is the particular case when g is the indicator function. It is an existence condition, which is in general weaker than **A1**.

Now we recall the definition of the quasi-Féjer convergence.

Definition 1.1. *We say that the sequence $\left(x^k\right)_{k\in\mathbb{N}}$ is quasi-Fejér convergent to a nonempty subset S of $\mathcal{H}$ iff $\forall x \in S$, $\exists(\epsilon_k)_{k\in\mathbb{N}} \in \ell_1(\mathbb{N})$ such that $\|x^{k+1} - x\|^2 \le \|x^k - x\|^2 + \epsilon_k$, for all $k \in \mathbb{N}$.*

This definition, originated in [20], has been elaborated further in [15, 25]. In the following we present two well-known fact for quasi-Fejér convergent sequences.

Fact 1.1. *If the sequence $\left(x^k\right)_{k\in\mathbb{N}}$ is quasi-Fejér convergent to S, then*

(a) the sequence $\left(x^k\right)_{k\in\mathbb{N}}$ is bounded and

(b) $\left(x^k\right)_{k\in\mathbb{N}}$ is weakly convergent iff all weak accumulation points of $\left(x^k\right)_{k\in\mathbb{N}}$ belong to S.

Proof. Item (a) follows from Proposition 3.3(i) of [15], and Item (b) follows from Theorem 3.8 of [15]. □

2 The Proximal Subgradient Splitting Method

In this section we propose the proximal subgradient splitting method extending the classical subgradient iteration. We prove that the sequence of point generated by the proposed method converges weakly to a solution of (1) using different ways to choose the stepsizes.

The method formally is stated as follows:

Proximal **S**ubgradient **S**plitting Method (**PSS Method**)

Initialization Step. Take $x^0 \in dom(g)$.

Iterative Step. Set

$$x^{k+1} = \operatorname{prox}_{\alpha_k g}\left(x^k - \alpha_k u^k\right), \tag{6}$$

where $u^k \in \partial f(x^k)$.

Stop Criteria. If $x^{k+1} = x^k$, then stop.

If **PSS Method** stops at step k, then $x^k = \text{prox}_{\alpha_k g}\left(x^k - \alpha_k u^k\right)$ with $u^k \in \partial f(x^k)$, implying that x^k is solution of problem (1). Then, from now on we assume that **PSS Method** generates an infinite sequence $\left(x^k\right)_{k\in\mathbb{N}}$. Moreover, it follows direct from (6) that the sequence $\left(x^k\right)_{k\in\mathbb{N}}$ belongs to $dom(g)$.

In the following we prove a crucial property of the iterates generated by **PSS Method**.

Lemma 2.1. *Let $x \in dom(g)$. Then, for all $k \in \mathbb{N}$,*

$$\|x^{k+1} - x\|^2 \leq \|x^k - x\|^2 + 2\alpha_k \left[(f+g)(x) - (f+g)(x^k)\right] + \alpha_k^2 \|u^k + w^k\|^2,$$

where $w^k \in \partial g(x^k)$ is arbitrary.

Proof. Take any $x \in dom(g)$. Note that (5) and (6) imply that $w^{k+1} := \dfrac{x^k - x^{k+1}}{\alpha_k} - u^k$, with $u^k \in \partial f(x^k)$ as defined by **PSS Method**, belongs to $\partial g(x^{k+1})$. Then,

$$\begin{aligned}
&\alpha_k^2\|u^k + w^{k+1}\|^2 + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 = \|x^{k+1} - x^k\|^2 + \|x^k - x\|^2 - \|x^{k+1} - x\|^2 \\
&= 2\langle x^k - x^{k+1}, x^k - x\rangle = 2\alpha_k\langle u^k, x^k - x\rangle + 2\langle x^k - x^{k+1} - \alpha_k u^k, x^k - x\rangle \\
&= 2\alpha_k\langle u^k, x^k - x\rangle + 2\alpha_k\left\langle \frac{x^k - x^{k+1}}{\alpha_k} - u^k, x^{k+1} - x\right\rangle + 2\alpha_k\left\langle \frac{x^k - x^{k+1}}{\alpha_k} - u^k, x^k - x^{k+1}\right\rangle \\
&= 2\alpha_k\langle u^k, x^k - x\rangle + 2\alpha_k\left\langle \frac{x^k - x^{k+1}}{\alpha_k} - u^k, x^{k+1} - x\right\rangle + 2\|x^k - x^{k+1}\|^2 - 2\alpha_k\langle u^k, x^k - x^{k+1}\rangle.
\end{aligned}$$

Now using again that $\dfrac{x^k - x^{k+1}}{\alpha_k} - u^k = w^{k+1} \in \partial g(x^{k+1})$ and the convexity of f and g, we obtain

$$\begin{aligned}
&2\langle x^k - x^{k+1}, x^k - x\rangle \geq 2\alpha_k\left[f(x^k) - f(x) + g(x^{k+1}) - g(x) + \langle u^k, x^{k+1} - x^k\rangle\right] + 2\|x^k - x^{k+1}\|^2 \\
&= 2\alpha_k\left[(f+g)(x^k) - (f+g)(x) + g(x^{k+1}) - g(x^k) + \langle u^k, x^{k+1} - x^k\rangle\right] + 2\|x^k - x^{k+1}\|^2 \\
&\geq 2\alpha_k\left[(f+g)(x^k) - (f+g)(x) + \langle w^k + u^k, x^{k+1} - x^k\rangle\right] + 2\alpha_k^2\|u^k + w^{k+1}\|^2,
\end{aligned}$$

for any $w^k \in \partial g(x^k)$. We thus have shown that

$$\begin{aligned}
\|x^{k+1} - x\|^2 \leq\ & \|x^k - x\|^2 + 2\alpha_k\left[(f+g)(x) - (f+g)(x^k)\right] \\
& + 2\alpha_k^2\langle w^k + u^k, u^k + w^{k+1}\rangle - \alpha_k^2\|u^k + w^{k+1}\|^2 \\
=\ & \|x^k - x\|^2 + 2\alpha_k\left[(f+g)(x) - (f+g)(x^k)\right] + \alpha_k^2\|u^k + w^k\|^2.
\end{aligned}$$

Note that $w^k \in \partial g(x^k)$ is arbitrary and the result follows. ☐

Since the subgradient methods are not a descent methods, like the proposed method here, it is common to keep track of the best point found so far, i.e., the one with smallest function value of the iterates. At each step, we set it recursively as $(f+g)^0_{\text{best}} := (f+g)(x^0)$ and

$$(f+g)^k_{\text{best}} := \min\left\{(f+g)^{k-1}_{\text{best}}, (f+g)(x^k)\right\}, \tag{7}$$

for all k. Since $\left((f+g)^k_{\text{best}}\right)_{k\in\mathbb{N}}$ is a decreasing sequence, it has a limit (which can be $-\infty$). When the function f is differentiable and its gradient Lipschitz continuous, it is possible to prove the complexity of the iterates generated by **PSS Method**; see [32]. In our instance (f is not necessarily differentiable) we expect, of course, slower convergence.

Next we present a convergence rate result for the sequence of the best functional values $\big((f+g)^k_{\text{best}}\big)_{k\in\mathbb{N}}$ to $s_* = \inf\{(f+g)(x) : x \in \mathcal{H}\}$.

Lemma 2.2. *Let $\big((f+g)^k_{\text{best}}\big)_{k\in\mathbb{N}}$ be the sequence defined by (7). If $S_* \neq \emptyset$, then, for all $k \in \mathbb{N}$,*

$$(f+g)^k_{\text{best}} - s_* \le \frac{[\operatorname{dist}(x^0, S_*)]^2 + C_k \sum_{i=0}^k \alpha_i^2}{2\sum_{i=0}^k \alpha_i},$$

where $C_k := \max\left\{\|u^i + w^i\|^2 : 0 \le i \le k\right\}$ with $w^i \in \partial g(x^i)$ $(i = 0, \ldots, k)$ are arbitrary.

Proof. Define $x_* := P_{S_*}(x^0)$. Note that x_* exists because S_* is a nonempty closed and convex set of $\mathcal{H}$. Using Lemma 2.5, with $x_* \in S_*$, we get

$$\begin{aligned}
\|x^{k+1} - x_*\|^2 &\le \|x^k - x_*\|^2 + 2\alpha_k\left[s_* - (f+g)(x^k)\right] + \alpha_k^2\|u^k + w^k\|^2 \\
&\le \|x^0 - x_*\|^2 + 2\sum_{i=0}^k \alpha_i\left[s_* - (f+g)(x^i)\right] + C_k\sum_{i=0}^k \alpha_i^2 \qquad (8)\\
&\le [\operatorname{dist}(x^0, S_*)]^2 + 2\left[s_* - (f+g)^k_{\text{best}}\right]\sum_{i=0}^k \alpha_i + C_k\sum_{i=0}^k \alpha_i^2,
\end{aligned}$$

where $(f+g)^k_{\text{best}}$ is defined by (7) and the result follows after simple algebra. $\square$

Next we establish the rate of convergence of the ergodic sequence $\big(\bar{x}^k\big)_{k\in\mathbb{N}}$ of $\big(x^k\big)_{k\in\mathbb{N}}$, which is defined recursively as $\bar{x}^0 = x^0$ and given $\sigma_0 = \alpha_0$ and $\sigma_k = \sigma_{k-1} + \alpha_k$, we define

$$\bar{x}^k = \left(1 - \frac{\alpha_k}{\sigma_k}\right)\bar{x}^{k-1} + \frac{\alpha_k}{\sigma_k}x^k.$$

After easy induction, we have $\sigma_k = \sum_{i=0}^k \alpha_i$ and

$$\bar{x}^k = \frac{1}{\sigma_k}\sum_{i=0}^k \alpha_i\, x^i, \qquad (9)$$

for all $k \in \mathbb{N}$.

The following result is very similar to Lemma 2.2, now over the ergodic sequence.

Lemma 2.3. *Let $\big(\bar{x}^k\big)_{k\in\mathbb{N}}$ be the ergodic sequence defined by (9). If $S_* \neq \emptyset$, then*

$$(f+g)(\bar{x}^k) - s_* \le \frac{[\operatorname{dist}(x^0, S_*)]^2 + C_k \sum_{i=0}^k \alpha_i^2}{2\sum_{i=0}^k \alpha_i},$$

where $C_k = \max\left\{\|u^i + w^i\|^2 : 0 \le i \le k\right\}$ with $w^i \in \partial g(x^i)$ $(i = 0, \ldots, k)$ are arbitrary.

Proof. Repeating the proof of Lemma 2.2 until Equation (8) and after dividing by $\sigma_k := \sum_{i=0}^{k} \alpha_i$, we get

$$\begin{aligned}\sum_{i=0}^{k} \frac{\alpha_i}{\sigma_k}\left[(f+g)(x^i) - s_*\right] &\leq \frac{1}{2\sigma_k}\left([\operatorname{dist}(x^0, S_*)]^2 - \|x^{k+1} - x_*\|^2\right) + \frac{C_k}{2\sigma_k}\sum_{i=0}^{k}\alpha_i^2 \\ &\leq \frac{1}{2\sigma_k}\left([\operatorname{dist}(x^0, S_*)]^2 + C_k\sum_{i=0}^{k}\alpha_i^2\right). \qquad (10)\end{aligned}$$

Using the convexity of $f+g$ and (9) in the above inequality (10), the result follows. $\square$

If we consider constant stepsizes, i.e., $\alpha_k = \alpha$ for all $k \in \mathbb{N}$, then the optimal rate is obtained when $\alpha = \frac{\operatorname{dist}(x^0, S_*)}{\sqrt{C_k}} \cdot \frac{1}{\sqrt{k+1}}$ from minimizing the right part of Lemmas 2.2 and 2.3. Our focus on constant stepsizes is motivated by the fact that we are interested in quantifying the progress of the proposed method in finite number of iterations to archiving an approximate solution.

Corollary 2.4. *Let $\left(x^k\right)_{k\in\mathbb{N}}$ be the sequence generated by* **PSS Method** *with the stepsizes α_k constant equal to α, $\left((f+g)_{\text{best}}^k\right)_{k\in\mathbb{N}}$ be the sequence defined by (7) and $\bar{x}^k$ be the ergodic sequence as (9). Then, the iteration attains the optimal rate at $\alpha = \alpha_* := \frac{\operatorname{dist}(x^0, S_*)}{\sqrt{C_k}} \cdot \frac{1}{\sqrt{k+1}}$, i.e., for all $k \in \mathbb{N}$,*

$$(f+g)_{\text{best}}^k - s_* \leq \frac{[\operatorname{dist}(x^0, S_*)]^2 + \alpha^2(k+1)C_k}{2(k+1)\alpha} \leq \frac{\operatorname{dist}(x^0, S_*) \cdot \sqrt{C_k}}{\sqrt{k+1}}$$

and

$$(f+g)(\bar{x}^k) - s_* \leq \frac{[\operatorname{dist}(x^0, S_*)]^2 + \alpha^2(k+1)C_k}{2(k+1)\alpha} \leq \frac{\operatorname{dist}(x^0, S_*) \cdot \sqrt{C_k}}{\sqrt{k+1}},$$

where $C_k = \max\left\{\|u^i + w^i\|^2 : 0 \leq i \leq k\right\}$ with $w^i \in \partial g(x^i)$ $(i = 0, \ldots, k)$ are arbitrary.

Our analysis showed that the expected error of the iterates generated by **PSS Method** with constant stepsizes after k iterations is $\mathcal{O}\left((k+1)^{-1/2}\right)$. Hence, we can search an ε-solution of problem (1) with $\mathcal{O}\left(\varepsilon^{-2}\right)$ iterations. Of course, this is worse than the rate $\mathcal{O}(k^{-1})$ and $\mathcal{O}\left(\varepsilon^{-1}\right)$ iterations of the proximal forward-backward iteration for the differentiable and convex f with Lipschitz continuous gradient; see, for instance, [32].

2.1 Exogenous stepsizes

In this subsection we analyze the convergence of **PSS Method** using exogenous stepsizes, i.e. the positive exogenous sequence of stepsizes $(\alpha_k)_{k\in\mathbb{N}}$ satisfies that $\alpha_k = \dfrac{\beta_k}{\eta_k}$ where $\eta_k := \max\{1, \|u^k\|\}$ for all k, and

$$\sum_{k=0}^{\infty}\beta_k^2 < +\infty \qquad \text{and} \qquad \sum_{k=0}^{\infty}\beta_k = +\infty. \qquad (11)$$

We begin with a useful consequence of Lemma 2.1.

Corollary 2.5. *Let $x \in dom(g)$. Then, for all $k \in \mathbb{N}$,*

$$\|x^{k+1} - x\|^2 \leq \|x^k - x\|^2 + 2\frac{\beta_k}{\eta_k}\left[(f+g)(x) - (f+g)(x^k)\right] + \left(1 + 2\rho + \rho^2\right)\beta_k^2,$$

where $\rho > 0$ as Assumption **A2**.

Proof. The result follows after note that $\eta_k \geq \|u^k\|$, $\eta_k \geq 1$ for all $k \in \mathbb{N}$ and as consequence

$$\frac{\|u^k + w^k\|^2}{\eta_k^2} = \frac{\|u^k\|^2}{\eta_k^2} + 2\frac{\|u^k\|\|w^k\|}{\eta_k^2} + \frac{\|w^k\|^2}{\eta_k^2} \leq 1 + 2\rho + \rho^2,$$

since $w^k \in \partial g(x^k)$ is arbitrary, in view of Assumption **A2** we can assume that $\|w^k\| \leq \rho$ for all $k \in \mathbb{N}$. $\square$

Now we define the auxiliary set

$$S_{\text{lev}} := \left\{ x \in dom(g) : \quad (f+g)(x) \leq (f+g)(x^k), \ \forall k \in \mathbb{N} \right\}. \tag{12}$$

When the solution set of problem (1) is nonempty, $S_{\text{lev}} \neq \emptyset$ because $S_* \subseteq S_{\text{lev}}$. Now, we prove the main result of this subsection in the following theorem.

Theorem 2.6. *Let $\left(x^k\right)_{k\in\mathbb{N}}$ be the sequence generated by* **PSS Method** *with exogenous stepesizes.*

(a) *If exists $\bar{x} \in S_{\text{lev}}$, then:*

  (i) *The sequence $\left(x^k\right)_{k\in\mathbb{N}}$ is quasi-Fejér convergent to*

$$\mathcal{L}_{f+g}(\bar{x}) := \{x \in dom(g) \ : \ (f+g)(x) \leq (f+g)(\bar{x})\}.$$

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

  (iii) *The sequence $\left(x^k\right)_{k\in\mathbb{N}}$ is weakly convergent to some $\tilde{x} \in \mathcal{L}_{f+g}(\bar{x})$.*

(b) $\liminf_{k\to\infty}(f+g)(x^k) = \inf_{x\in\mathcal{H}}(f+g)(x) = s_*$ *(possibly $s_* = -\infty$).*

(c) *If $S_* \neq \emptyset$, then the sequence $\left(x^k\right)_{k\in\mathbb{N}}$ converges weakly to some $\bar{x} \in S_*$.*

(d) *If $S_* = \emptyset$, then $\left(x^k\right)_{k\in\mathbb{N}}$ is unbounded.*

Proof.

(a) By assumption there exists $\bar{x} \in S_{\text{lev}}$, i.e., $(f+g)(\bar{x}) \leq (f+g)(x^k)$, for all $k \in \mathbb{N}$.

(i) To show that $\left(x^k\right)_{k\in\mathbb{N}}$ is quasi-Fejér convergent to $\mathcal{L}_{f+g}(\bar{x})$ (which is nonempty because $\bar{x} \in \mathcal{L}_{f+g}(\bar{x})$), we use Corollary 2.5, for any $x \in \mathcal{L}_{f+g}(\bar{x}) \subseteq dom(g)$, establishing that $\|x^{k+1} - x\|^2 \leq \|x^k - x\|^2 + (1 + 2\rho + \rho^2)\beta_k^2$, for all $k \in \mathbb{N}$. Thus, $\left(x^k\right)_{k\in\mathbb{N}}$ is quasi-Fejér convergent to $\mathcal{L}_{f+g}(\bar{x})$.

(ii) The sequence $\left(x^k\right)_{k\in\mathbb{N}}$ is bounded from Fact 1.1(a), and hence it has accumulation points. To prove that

$$\lim_{k\to\infty} (f+g)(x^k) = (f+g)(\bar{x}), \tag{13}$$

we use Corollary 2.5, with $x = \bar{x} \in \mathcal{L}_{f+g}(\bar{x}) \subseteq dom(g)$, getting

$$\beta_k \left[(f+g)(x^k) - (f+g)(\bar{x})\right] \leq \frac{1}{2}(\|x^k - \bar{x}\|^2 - \|x^{k+1} - \bar{x}\|^2) + 2(1 + 2\rho + \rho^2)\beta_k^2.$$

Summing, from $k = 0$ to m, the above inequality, we have

$$\sum_{k=0}^{m} \beta_k \left[(f+g)(x^k) - (f+g)(\bar{x})\right] \leq \frac{1}{2}(\|x^0 - \bar{x}\|^2 - \|x^{m+1} - \bar{x}\|^2) + 2(1 + 2\rho + \rho^2)\sum_{k=0}^{m} \beta_k^2,$$

and taking limit, when m goes to ∞,

$$\sum_{k=0}^{\infty}\beta_k\left[(f+g)(x^k)-(f+g)(\bar{x})\right]<+\infty. \tag{14}$$

Then, (14) together with (11) implies that there exists a subsequence $\big((f+g)(x^{i_k})\big)_{k\in\mathbb{N}}$ of $\big((f+g)(x^k)\big)_{k\in\mathbb{N}}$ such that

$$\liminf_{k\to\infty}\left[(f+g)(x^{i_k})-(f+g)(\bar{x})\right]=0. \tag{15}$$

Indeed, if (15) does not hold, then there exists $\sigma>0$ and $k\geq\tilde{k}$, such that $(f+g)(x^k)-(f+g)(\bar{x})\geq\sigma$ and using (14), we get

$$+\infty>\sum_{k=\tilde{k}}^{\infty}\beta_k\left[(f+g)(x^k)-(f+g)(\bar{x})\right]\geq\sigma\sum_{k=\tilde{k}}^{\infty}\beta_k,$$

in contradiction with (11). Also, define $\varphi_k:=(f+g)(x^k)-(f+g)(\bar{x})$, which is positive for all k because $\bar{x}\in S_{\text{lev}}$. Then, for any $u^k\in\partial g(x^k)$ and $w^k\in\partial g(x^k)$, we get

$$\begin{aligned}\varphi_k-\varphi_{k+1}&=(f+g)(x^k)-(f+g)(x^{k+1})\leq\langle u^k+w^k,x^k-x^{k+1}\rangle\\&\leq\|u^k+w^k\|\|x^k-x^{k+1}\|\leq(\zeta+\rho)\|x^k-x^{k+1}\|,\end{aligned} \tag{16}$$

where $\zeta>0$ such that $\|u^k\|\leq\zeta$, for all $k\in\mathbb{N}$ (ζ exists in virtue of the boundedness of $\left(x^k\right)_{k\in\mathbb{N}}$ and Assumption **A1**) and $\|w^k\|\leq\rho$, for all $k\in\mathbb{N}$ (ρ exists because $w^k\in\partial g(x^k)$ are arbitrary and Assumption **A2**). Using Corollary 2.5, with $x=x^k$, we have $\|x^k-x^{k+1}\|\leq\sqrt{1+2\rho+\rho^2}\cdot\beta_k$, which together with (16) implies that

$$\varphi_k-\varphi_{k+1}\leq\sqrt{1+2\rho+\rho^2}\cdot(\zeta+\rho)\beta_k:=\bar{\rho}\beta_k \tag{17}$$

for all $k\in\mathbb{N}$. From (15), there exists a subsequence $(\varphi_{i_k})_{k\in\mathbb{N}}$ of $(\varphi_k)_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}\varphi_{i_k}=0$. If the claim given in (13) does not hold, then there exists some $\delta>0$ and a subsequence $(\varphi_{\ell_k})_{k\in\mathbb{N}}$ of $(\varphi_k)_{k\in\mathbb{N}}$, such that $\varphi_{\ell_k}\geq\delta$ for all $k\in\mathbb{N}$. Thus, we can construct a third subsequence $(\varphi_{j_k})_{k\in\mathbb{N}}$ of $(\varphi_k)_{k\in\mathbb{N}}$, where the indices j_k are chosen in the following way:

$$j_0:=\min\{m\geq 0\mid\varphi_m\geq\delta\},$$

$$j_{2k+1}:=\min\{m\geq j_{2k}\mid\varphi_m\leq\delta/2\},$$
$$j_{2k+2}:=\min\{m\geq j_{2k+1}\mid\varphi_m\geq\delta\},$$

for each k. The existence of the subsequences $(\varphi_{i_k})_{k\in\mathbb{N}}$, $(\varphi_{\ell_k})_{k\in\mathbb{N}}$ of $(\varphi_k)_{k\in\mathbb{N}}$, guarantees that the subsequence $(\varphi_{j_k})_{k\in\mathbb{N}}$ of $(\varphi_k)_{k\in\mathbb{N}}$ is well-defined for all $k\geq 0$. It follows from the definition of j_k that

$$\begin{aligned}&\varphi_m\geq\delta\quad\text{for}\quad j_{2k}\leq m\leq j_{2k+1}-1\\&\varphi_m\leq\frac{\delta}{2}\quad\text{for}\quad j_{2k+1}\leq m\leq j_{2k+2}-1\end{aligned} \tag{18}$$

for all k, and hence

$$\varphi_{j_{2k}}-\varphi_{j_{2k+1}}\geq\frac{\delta}{2}, \tag{19}$$

for all $k \in \mathbb{N}$. In view of (14) and recall that $\varphi_k = (f+g)(x^k) - (f+g)(\bar{x}) \geq 0$ for all $k \in \mathbb{N}$,

$$
\begin{aligned}
+\infty &> \sum_{k=0}^{\infty} \beta_k \varphi_k \geq \sum_{k=0}^{\infty} \sum_{m=j_{2k}}^{j_{2k+1}-1} \beta_m \varphi_m \geq \frac{\delta}{2} \sum_{k=0}^{\infty} \sum_{m=j_{2k}}^{j_{2k+1}-1} \beta_m \\
&= \frac{\delta}{2\bar{\rho}} \sum_{k=0}^{\infty} \sum_{m=j_{2k}}^{j_{2k+1}-1} \bar{\rho} \beta_m \geq \frac{\delta}{2\bar{\rho}} \sum_{k=0}^{\infty} \sum_{m=j_{2k}}^{j_{2k+1}-1} (\varphi_m - \varphi_{m+1}) = \frac{\delta}{2\bar{\rho}} \sum_{k=0}^{\infty} (\varphi_{j_{2k}} - \varphi_{j_{2k+1}}) \\
&\geq \frac{\delta}{2\bar{\rho}} \sum_{k=0}^{\infty} \frac{\delta}{2} = +\infty,
\end{aligned}
$$

where we have used (18) in the second inequality and (17) in the third inequality and (19) in the last one. Thus, $\lim_{k\to\infty} (f+g)(x^k) = (f+g)(\bar{x})$, establishing (ii).

(iii) Let $\tilde{x}$ a weak accumulation point of $\left(x^k\right)_{k\in\mathbb{N}}$, which exists by Item (a)(i) and Fact 1.1(i). From now on, we denote $\left(x^{i_k}\right)_{k\in\mathbb{N}}$ any subsequence of $\left(x^k\right)_{k\in\mathbb{N}}$ converging weakly to $\tilde{x}$. Since $f+g$ is weakly lower semicontinuous and using (13), we get

$$(f+g)(\tilde{x}) \leq \liminf_{k\to\infty} (f+g)(x^{i_k}) = \lim_{k\to\infty} (f+g)(x^k) = (f+g)(\bar{x}),$$

implying that $(f+g)(\tilde{x}) \leq (f+g)(\bar{x})$ and thus $\tilde{x} \in \mathcal{L}_{f+g}(\bar{x})$. As consequence, all accumulation points of $\left(x^k\right)_{k\in\mathbb{N}}$ belong to $\mathcal{L}_{f+g}(\bar{x})$ and since $\left(x^k\right)_{k\in\mathbb{N}}$ is quasi-Fejér convergent to $\mathcal{L}_{f+g}(\bar{x})$, we get that $\left(x^k\right)_{k\in\mathbb{N}}$ converges to $\tilde{x} \in \mathcal{L}_{f+g}(\bar{x})$ from Fact 1.1(b).

(b) Since $\left(x^k\right)_{k\in\mathbb{N}} \subset dom(g)$, we get $s_* \leq \liminf_{k\to\infty}(f+g)(x^k)$. Suppose that $s_* < \liminf_{k\to\infty}(f+g)(x^k)$. Hence, there exists $\hat{x}$ such that

$$(f+g)(\hat{x}) < \liminf_{k\to\infty} (f+g)(x^k). \tag{20}$$

It follows from (20) that there exists $\bar{k} \in \mathbb{N}$ such that $(f+g)(\hat{x}) \leq (f+g)(x^k)$ for all $k \geq \bar{k}$. Since $\bar{k}$ is finite we can assume without loss of generality that $(f+g)(\hat{x}) \leq (f+g)(x^k)$ for all $k \in \mathbb{N}$. Using the definition of S_{lev}, given in (12), we have that $\hat{x} \in S_{\text{lev}}$. By Item (a)(ii) $\lim_{k\to\infty}(f+g)(x^k) = (f+g)(\hat{x})$, in contradiction with (20). Establishing the result.

(c) Since $S_* \neq \emptyset$, take $x_* \in S_*$ and as consequence, $\mathcal{L}_{f+g}(x_*) = S_*$. Since $\left(x^k\right)_{k\in\mathbb{N}} \subset dom(g)$, we get $(f+g)(x_*) \leq (f+g)(x^k)$ for all $k \in \mathbb{N}$ implying that $x_* \in S_{\text{lev}}$. Applying Item (a)(iii), with $\bar{x} = x_*$, we get that $\left(x^k\right)_{k\in\mathbb{N}}$ converges weakly to some $\tilde{x} \in S_*$.

(d) Assume that S_* is empty but $\left(x^k\right)_{k\in\mathbb{N}}$ is bounded. Let $\left(x^{\ell_k}\right)_{k\in\mathbb{N}}$ be a subsequence of $\left(x^k\right)_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}(f+g)(x^{\ell_k}) = \liminf_{k\to\infty}(f+g)(x^k)$. Since $\left(x^{\ell_k}\right)_{k\in\mathbb{N}}$ is bounded, without loss of generality (i.e. refining $\left(x^{\ell_k}\right)_{k\in\mathbb{N}}$ if necessary), we may assume that $\left(x^{\ell_k}\right)_{k\in\mathbb{N}}$ converges weakly to some $\bar{x} \in dom(g)$. By the weak lower semicontinuity of $f+g$ on $dom(g)$,

$$(f+g)(\bar{x}) \leq \liminf_{k\to\infty}(f+g)(x^{\ell_k}) = \lim_{k\to\infty}(f+g)(x^{\ell_k}) = \liminf_{k\to\infty}(f+g)(x^k) = s_*, \tag{21}$$

using Item (b) in the last equality. By (21), $\bar{x} \in S_*$, in contradiction with the hypothesis and the result follows. $\square$

For exogenous stepsizes, Theorem 2.6 guarantees the convergence of $\big((f+g)(x^k)\big)_{k\in\mathbb{N}}$ to the optimal value of problem (1), i.e., $\liminf_{k\to\infty}(f+g)(x^k)=s_*$, implying the convergence of $\big((f+g)^k_{\text{best}}\big)_{k\in\mathbb{N}}$, defined in (7), to s_*. It is important to say that in the proof of the above crucial result, we have used a similar idea recently presented in [6] for a different instance.

In the following we present a direct consequence of Lemmas 2.2 and 2.3, when the stepsizes are given by (11).

Corollary 2.7. *Let $\left(\bar{x}^k\right)_{k\in\mathbb{N}}$ be the ergodic sequence defined by* (9) *and $(\beta_k)_{k\in\mathbb{N}}$ as* (11)*. If $S_*\neq\emptyset$, then, for all $k\in\mathbb{N}$,*

$$(f+g)^k_{\text{best}}-s_*\leq\zeta\frac{[\operatorname{dist}(x^0,S_*)]^2+(1+2\rho+\rho^2)\sum_{i=0}^k\beta_i^2}{2\sum_{i=0}^k\beta_i}$$

and

$$(f+g)(\bar{x}^k)-s_*\leq\zeta\frac{[\operatorname{dist}(x^0,S_*)]^2+(1+2\rho+\rho^2)\sum_{i=0}^k\beta_i^2}{2\sum_{i=0}^k\beta_i},$$

where $\zeta>0$ and $\rho>0$ are as Assumptions **A1** *and* **A2***, respectively.*

The above corollary shows that if we assume existence of solutions, the expected error of the iterates generated by **PSS Method** with the exogenous stepsizes (11) after k iterations is $\mathcal{O}\left((\sum_{i=0}^k\beta_i)^{-1}\right)$. Since $(\beta_k)_{k\in\mathbb{N}}$ satisfies (11) the best performance of the iteration (in term of functional values) is archived for example taking $\beta_k\cong 1/k^r$ with r bigger than $1/2$, but near of this value, for all k.

2.2 Polyak stepsizes

In this subsection we analyze the convergence of **PSS Method** using Polyak stepsizes, i.e. the positive exogenous sequence of stepsizes $(\alpha_k)_{k\in\mathbb{N}}$ satisfies that having chose any $w^k\in\partial g(x^k)$ and denoted $\rho_k:=\|w^k\|$ for all $k\in\mathbb{N}$. Then define, for all $k\in\mathbb{N}$,

$$\alpha_k=\gamma_k\frac{(f+g)(x^k)-s_k}{\|u^k\|^2+2\rho_k\|u^k\|+\rho_k^2},\tag{22}$$

where $0<\gamma\leq\gamma_k\leq 2-\gamma$ and assume that s_k a monotone decreasing variable target value approximating $s_*:=\inf\{(f+g)(x):x\in\mathcal{H}\}$ is available, satisfying that $s_k\leq(f+g)(x^k)$ for all $k\in\mathbb{N}$. When s_* is known, the simplest variant of the stepsizes proposed in (22) is obtained selecting the stepsizes

$$\alpha_k=\gamma_k\frac{(f+g)(x^k)-s_*}{\|u^k\|^2+2\rho_k\|u^k\|+\rho_k^2},\tag{23}$$

for all $k\in\mathbb{N}$. Unfortunately, for finding an optimal solution, scheme (23) requires prior knowledge of the optimal objective function value s_*. As s_* is usually unknown, we prefer to do our analysis over (22), replacing it by the variable target value s_k. When g is the indicator function of a closed and convex set a further discussion about how to choose s_k has been presented in the literature for problems where a good upper or lower bound of the optimal objective function value is available; see, for instance, [24,27,36,38].

Now we present a direct consequence of Lemma 2.1.

Corollary 2.8. *Suppose that* $\lim_{k\to\infty} s_k = \tilde{s} \geq s_*$ *and let any* $x \in \mathcal{L}_{f+g}(\tilde{s})$. *Then,*

$$\|x^{k+1} - x\|^2 \leq \|x^k - x\|^2 - \gamma(2-\gamma)\frac{\left[s_k - (f+g)(x^k)\right]^2}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2},$$

for all $k \in \mathbb{N}$.

Proof. Take $x \in \mathcal{L}_{f+g}(\tilde{s}) := \{x \in dom(g) \;:\; (f+g)(x) \leq \tilde{s}\}$. Since $(s_k)_{k\in\mathbb{N}}$ is a monotone decreasing sequence convergent to $\tilde{s}$, which is below to the functional values iterates,

$$(f+g)(x^k) \geq s_k \geq \tilde{s} \geq (f+g)(x), \quad \forall x \in \mathcal{L}_{f+g}(\tilde{s}), \tag{24}$$

for all $k \in \mathbb{N}$. Then, applying Lemma 2.1 and using (24), we get, for all $k \in \mathbb{N}$,

$$\begin{aligned}
\|x^{k+1} - x\|^2 &\leq \|x^k - x\|^2 - 2\gamma_k \frac{\left[s_k - (f+g)(x^k)\right]\left[(f+g)(x) - (f+g)(x^k)\right]}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2} \\
&\quad + \gamma_k^2 \frac{\left[s_k - (f+g)(x^k)\right]^2}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2} \\
&\leq \|x^k - x\|^2 - \gamma_k(2-\gamma_k)\frac{\left[s_k - (f+g)(x^k)\right]^2}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2} \\
&\leq \|x^k - x\|^2 - \gamma(2-\gamma)\frac{\left[s_k - (f+g)(x^k)\right]^2}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2},
\end{aligned} \tag{25}$$

where we used that $x \in \mathcal{L}_{f+g}(\tilde{s})$ and (24) in the second inequality. The result follows from (25). $\square$

Now, we prove the first main result of this subsection in the following theorem.

Theorem 2.9. *Let* $\left(x^k\right)_{k\in\mathbb{N}}$ *the sequence generated by* **PSS Method** *with* α_k *given by* (22). *If* $\lim_{k\to\infty} s_k = \tilde{s} \geq s_*$ *and* $\mathcal{L}_{f+g}(\tilde{s}) \neq \emptyset$, *then*

(*a*) $\left(x^k\right)_{k\in\mathbb{N}}$ *is quasi-Fejér convergent to* $\mathcal{L}_{f+g}(\tilde{s})$.

(*b*) $\lim_{k\to\infty} \left[(f+g)(x^k) - s_k\right] = 0$.

(*c*) $\left(x^k\right)_{k\in\mathbb{N}}$ *is weakly convergent to some* $\tilde{x} \in \mathcal{L}_{f+g}(\tilde{s})$.

Proof.

(a) It is direct consequence of Corollary 2.8.

(b) By Item (a), $\left(x^k\right)_{k\in\mathbb{N}}$ is bounded. Using Corollary 2.8, we get, for any $x \in \mathcal{L}_{f+g}(\tilde{s})$,

$$\begin{aligned}
\gamma(2-\gamma)\left[s_k - (f+g)(x^k)\right]^2 &\leq \left(\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2\right)\left[\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right] \\
&\leq \left(\zeta^2 + 2\rho\zeta + \rho^2\right)\left[\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right] \\
&:= \hat{\rho}\left[\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right],
\end{aligned} \tag{26}$$

where the last inequality following from Assumptions **A1** and **A2** ($\|u^k\| \le \zeta$ and $\rho_k = \|w^k\| \le \rho$ for all $k \in \mathbb{N}$). Summing (26), over $k = 0$ to m, we obtain

$$\gamma(2-\gamma)\sum_{k=0}^{m}\Big[s_k - (f+g)(x^k)\Big]^2 \le \hat{\rho}\left[\|x^0 - x\|^2 - \|x^{m+1} - x\|^2\right] \le \hat{\rho}\|x^0 - x\|^2. \tag{27}$$

Taking limit when m goes to ∞, we get the desired result.

(c) From Item (b) if $\tilde{s} = \lim_{k\to\infty} s_k$, then $\lim_{k\to\infty}(f+g)(x^k) = \tilde{s}$. Let $\tilde{x}$ a weak accumulation point of $(x^k)_{k\in\mathbb{N}}$, which exists by the boundedness of $(x^k)_{k\in\mathbb{N}}$ direct consequence of Item (a). From now on, we denote $(x^{\ell_k})_{k\in\mathbb{N}}$ any subsequence of $(x^k)_{k\in\mathbb{N}}$ converging weakly to $\tilde{x}$. Since $f+g$ is weakly lower semicontinuous, we get $(f+g)(\tilde{x}) \le \liminf_{k\to\infty}(f+g)(x^{\ell_k}) = \lim_{k\to\infty}(f+g)(x^k) = \tilde{s}$, implying that $(f+g)(\tilde{x}) \le \tilde{s}$ and thus $\tilde{x} \in \mathcal{L}_{f+g}(\tilde{s})$. The result follows from Fact 1.1(b) and Item (a). □

Before the analysis of the inconsistent case when $\tilde{s} = \lim_{k\to\infty} s_k$ is strictly less than $s_* = \inf\{(f+g)(x) : x \in \mathcal{H}\}$, we present a useful corollary direct consequence of Theorem 2.9, which will be used for the analysis of this case, $\tilde{s} < s_*$. In the next corollary, we show the special case when the optimal value s_* is known and finite and the stepsize α_k is defined by (23), i.e., for all $k \in \mathbb{N}$,

$$\alpha_k = \gamma_k \frac{(f+g)(x^k) - s_*}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2},$$

where $0 < \gamma \le \gamma_k \le 2 - \gamma$.

Corollary 2.10. *Let $(x^k)_{k\in\mathbb{N}}$ the sequence generated by* **PSS Method** *with α_k given by (23), and $S_* \neq \emptyset$. Then,*

(*a*) *$(x^k)_{k\in\mathbb{N}}$ is quasi-Fejér convergent to S_*.*

(*b*) $\lim_{k\to\infty}(f+g)(x^k) = s_*$.

(*c*) *$(x^k)_{k\in\mathbb{N}}$ is weakly convergent to some $\tilde{x} \in S_*$.*

(*d*) $\lim_{k\to\infty}\sqrt{k+1}\cdot\left[(f+g)(x^k) - s_*\right] = 0$.

Proof. The items (a) to (c) are direct consequence of Theorem 2.9. The proof of Item (d) is by contradiction. Assume that $\lim_{k\to\infty}\sqrt{k+1}\cdot\left[(f+g)(x^k) - s_*\right] = 2\delta$, for some $\delta > 0$. Then, for $\bar{k}$ large enough, we have $(f+g)(x^k) - s_* \ge \frac{\delta}{\sqrt{k+1}}$ for all $k \ge \bar{k}$. Thus,

$$\sum_{k=\bar{k}}^{\infty}\Big[(f+g)(x^k) - s_*\Big]^2 \ge \delta^2\sum_{k=\bar{k}}^{\infty}\frac{1}{k+1} = +\infty. \tag{28}$$

On the other hand, by substituting the expression for the stepsize α_k given by (23), in (26) ($s_k = s_*$ for all $k \in \mathbb{N}$), we get, for all $k \ge \bar{k}$,

$$\sum_{k=\bar{k}}^{\infty}\Big[(f+g)(x^k) - s_*\Big]^2 < +\infty,$$

which contradicts (28). Establishing the result. □

Next a result on the complexity of the iterates

Lemma 2.11. *Let $\left(x^k\right)_{k\in\mathbb{N}}$ the sequence generated by* **PSS Method** *with α_k, given by (22). If $\lim_{k\to\infty} s_k = \tilde{s} \geq s_*$ and $\mathcal{L}_{f+g}(\tilde{s}) \neq \emptyset$, then, for all $k \in \mathbb{N}$,*

$$(f+g)^k_{\text{best}} - \tilde{s} \leq \sqrt{\frac{D_k}{\gamma(2-\gamma)}} \cdot \frac{\operatorname{dist}(x^0, \mathcal{L}_{f+g}(\tilde{s}))}{\sqrt{k+1}},$$

where $D_k := \max\left\{\|u^i\|^2 + 2\rho_i\|u^i\| + \rho_i^2 : 1 \leq i \leq k\right\}$ with $\rho_i := \|w^i\|$ and $w^i \in \partial g(x^i)$ $(i = 0, \ldots, k)$ are arbitrary. Moreover,

$$\lim_{k\to\infty} (f+g)^k_{\text{best}} = \tilde{s}.$$

Proof. Using the proof of Theorem 2.9, with any $x \in \mathcal{L}_{f+g}(\tilde{s})$, until (26), we obtain

$$(k+1)\left[(f+g)^k_{\text{best}} - \tilde{s}\right]^2 \leq \sum_{i=0}^{k}\left[(f+g)(x^i) - s_k\right]^2 \leq \frac{D_k}{\gamma(2-\gamma)}\left[\operatorname{dist}(x^0, \mathcal{L}_{f+g}(\tilde{s}))\right]^2,$$

where $D_k := \max\left\{\|u^i\|^2 + 2\rho_i\|u^i\| + \rho_i^2 : 1 \leq i \leq k\right\}$ with $\rho_i = \|w^i\|$ and $w^i \in \partial g(x^i)$ $(i = 0, \ldots, k)$ are arbitrary. After simple algebra the result follows. □

Our analysis proved that the expected error of the iterates generated by **PSS Method** with the Polyak stepsizes (22) after k iterations is $\mathcal{O}\left((k+1)^{-1/2}\right)$ if we assume $s_k \geq s_*$ for all $k \in \mathbb{N}$. Of course, the result is weaker than Corollary 2.10(d), when $s_k = s_*$ for all $k \in \mathbb{N}$.

Now we are ready to prove the last main result of this subsection.

Theorem 2.12. *Let $\left(x^k\right)_{k\in\mathbb{N}}$ the sequence generated by* **PSS Method** *with α_k, given by (22). If $S_* \neq \emptyset$ and $\lim_{k\to\infty} s_k = \tilde{s} < s_*$, then*

$$\lim_{k\to\infty} (f+g)^k_{\text{best}} = \lim_{k\to\infty} \min_{0\leq i\leq k} (f+g)(x^i) \leq s_* + \frac{2-\gamma}{\gamma}(s_* - \tilde{s}).$$

Proof. Suppose that $(f+g)(x^k) > s_*$, otherwise the result holds trivially. It is clear that, for all $k \in \mathbb{N}$,

$$\alpha_k = \gamma_k \frac{(f+g)(x^k) - s_k}{(f+g)(x^k) - s_*} \frac{(f+g)(x^k) - s_*}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2} := \tilde{\gamma}_k \frac{(f+g)(x^k) - s_*}{\|u^k\|^2 + 2\rho_k\|u^k\| + \rho_k^2},$$

where

$$\gamma \leq \tilde{\gamma}_k = \gamma_k \frac{(f+g)(x^k) - s_k}{(f+g)(x^k) - s_*},$$

which must be bigger than $2-\gamma$ for some $\bar{k} \in \mathbb{N}$. Otherwise, if

$$\tilde{\gamma}_k \leq 2 - \gamma \tag{29}$$

for all $k \in \mathbb{N}$, we can apply Corollary 2.10(b) for getting of $\lim_{k\to\infty} (f+g)(x^k) = s_*$, implying that $\tilde{\gamma}_k$ goes to $+\infty$ (note that for all sufficiently large k, $s_k < s_* \leq (f+g)(x^k)$, because $\tilde{s} < s_*$), which is a contradiction with (29). Thus, there exist $\bar{k}$ and $\delta > 0$ arbitrary such that

$$\gamma_{\bar{k}} \frac{(f+g)(x^{\bar{k}}) - s_{\bar{k}}}{(f+g)(x^{\bar{k}}) - s_*} = \tilde{\gamma}_{\bar{k}} > 2 - \delta.$$

After simple algebra and using that $s_{\bar{k}} \geq \tilde{s}$, we get that

$$(f+g)(x^{\bar{k}}) < s_* + \frac{\gamma_{\bar{k}}}{2-\delta-\gamma_{\bar{k}}}(s_* - \tilde{s}) \leq s_* + \frac{2-\gamma}{\gamma-\delta}(s_* - \tilde{s}),$$

since $\delta > 0$ was arbitrary and the result follows. □

Finally in the following corollary we summarize the behaviour of the limit of the sequence of $\left((f+g)^k_{\text{best}}\right)_{k\in\mathbb{N}}$ depending of the limit of $\tilde{s} = \lim_{k\to\infty} s_k$, which is direct consequence of Theorem 2.12 and Lemma 2.11.

Corollary 2.13. *Let* $\left(x^k\right)_{k\in\mathbb{N}}$ *the sequence generated by* **PSS Method** *with* α_k*, given by* (22)*. If* $S_* \neq \emptyset$ *and* $\lim_{k\to\infty} s_k = \tilde{s}$*, then*

$$\lim_{k\to\infty} (f+g)^k_{\text{best}} \begin{cases} = \tilde{s}, & if \quad \tilde{s} \geq s_* \\ \leq s_* + \dfrac{2-\gamma}{\gamma}(s_* - \tilde{s}), & if \quad \tilde{s} < s_*. \end{cases}$$

3 Final Remarks

In this work we dealt with the weak convergence of the new approach called the Proximal Subgradient Splitting (PSS) Method for minimizing the sum of two nonsmooth and convex functions. In the iteration of this method, neither the functions need be differentiable or finite in all $\mathcal{H}$ and, therefote, a broad class of problems can be solved. **PSS Method** is very useful when the proximal operator of f is complex to evaluate and its (sub)gradient is simple to compute.

As future research, we will investigate variations of our scheme for solving structured convex optimization problems with the aim of finding new methods, like the coordinate gradient method, which have been proposed, for instance, in [28, 33, 37] only for the differentiable case. We also are looking to the incremental subgradient method [30] for problem (1), when f is the sum of a large number of nonsmooth convex functions. The idea is to perform the subgradient iteration incrementally, by sequentially taking steps along the subgradients of the component functions, before the proximal step. This incremental approach has been very successful in solving large least squares problems, and it has resulted in a much better practical rate of convergence with application in the training of neural networks; see, for instance, [22].

On the other hand, it is important to say that the main drawback of subgradient iterations is their rather slow convergence. However, subgradient methods are distinguished by their applicability, simplicity and efficient use of memory, which is very important for large scale problems; especially if the required accuracy for the solution is not too high; see, for instance, [31] and the references therein. We also intend to study fast and variable metric versions of the proximal subgradient splitting method proposed here to achieve better performance, like in the differentiable case; see [19, 34].

Finally, we hope that this study serves as a basis for future research on other more efficient variants on the proximal subgradient iteration, like cutting plane method, ϵ-subgradients and bundle variants, conjugate gradient method and nonsmooth quasi-Newton and Newton methods for solving problem (1) and its variations.

ACKNOWLEDGMENTS

This work was completed while the author was visiting the University of British Columbia. The author is very grateful for the warm hospitality of the Irving K. Barber School of Arts and Sciences at the University of British Columbia.

References

[1] Alber, Ya.I., Iusem, A.N., Solodov, M.V. On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming* **81** (1998) 23–37.

[2] Bauschke, H.H., Borwein, J.M. On projection algorithms for solving convex feasibility problems. *SIAM Review* **38** (1996) 367–426.

[3] Bauschke, H.H., Combettes, P.L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer, New York (2011).

[4] Beck, A., Teboulle, M. A fast iterative shrinkage–thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** (2009) 183–202.

[5] Beck, A., Teboulle, M. *Gradient-Based Algorithms with Applications to Signal Recovery Problems.* in *Convex Optimization in Signal Processing and Communications*, (D. Palomar and Y. Eldar, eds.) 42–88, University Press, Cambribge (2010).

[6] Bello Cruz, J.Y. A subgradient method for vector optimization problems. *SIAM Journal on Optimization* **23** (2013) 2169–2182.

[7] Bello Cruz, J.Y., Iusem, A.N. A strongly convergent method for nonsmooth convex minimization in Hilbert spaces. *Numerical Functional Analysis and Optimization* **32** (2011) 1009–1018.

[8] Bello Cruz, J.Y., Iusem, A.N. Convergence of direct methods for paramonotone variational inequalities. *Computational Optimization and Application* **46** (2010) 247–263.

[9] Bello Cruz, J.Y., Nghia, T.T.A. On the convergence of the proximal forward-backward splitting method with linesearches. Technical report, (2014). Available in http://arxiv.org.

[10] Bot, R.I, Csetnek, E.R. Forward-Backward and Tseng's type penalty schemes for monotone inclusion problems. *Set-Valued and Variational Analysis* **22** (2014) 313–331.

[11] Candes, E.J., Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory* **51** (2005) 4203–4215.

[12] Chavent, G., Kunisch, K. Convergence of Tikhonov regularization for constrained ill-posed inverse problems. *Inverse Problems* **10** (1994) 63–76.

[13] Chen, G.H.-G., Rockafellar, R.T. Convergence rates in forward-backward splitting. *SIAM Journal on Optimization* **7** (1997) 421–444.

[14] Combettes, P.L. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization* **53** (2004) 475–504.

[15] Combettes, P.L. Quasi-Fejérian analysis of some optimization algorithms. *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications. Studies in Computational Mathematics* **8** 115–152 North-Holland, Amsterdam (2001).

[16] Combettes, P.L., Pesquet, J.-C. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of selected topics in signal precessing* **1** (2007) 564–574.

[17] Combettes, P.L., Pesquet, J.-C. Proximal splitting methods in signal processing. in *Fixed-Point Algorithms for Inverse Problems. Science and Engineering. Springer Optimization and Its Applications* **49** 185–212 Springer, New York (2011).

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

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

[20] Ermoliev, Yu.M. On the method of generalized stochastic gradients and quasi-Fejér sequences. *Cybernetics* **5** (1969) 208–220.

[21] Figueiredo, M., Novak, R., Wright, S.J. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing* **1** (2007) 586–597.

[22] Gaivoronski, A.A. Convergence analysis of parallel backpropagation algorithm for neural networks. *Optimization Methods and Software* **4** (1994) 117–134.

[23] Geman, S., Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (1984) 721–741.

[24] Held, M., Wolfe, P., Crowder, H. Validation of subgradient optimization. *Mathematical Programming* **6** (1974) 66–68.

[25] Iusem, A.N., Svaiter, B.F., Teboulle, M. Entropy-like proximal methods in convex programming. *Mathematics of Operations Research* **19** (1994) 790–814.

[26] James, G.M., Radchenko, P., Lv, J. DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **71** (2009) 127–142.

[27] Kim, S., Ahn, H., Cho, S.-C. Variable target value subgradient method. *Mathematical Programming* **49** (1991) 359–369.

[28] Lu, Z., Xiao, L. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming* DOI: 10.1007/s10107-014-0800-2 (2014).

[29] Neal, P., Boyd, S. Proximal Algorithms. *Foundations and Trends in Optimization* **1** (2014) 127–239.

[30] Nedic, A., Bertsekas, D.P. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization* **12** (2001) 109–138.

[31] Nesterov, Yu. Subgradient methods for huge-scale optimization problems. *Mathematical Programming* **146** (2014) 275–297.

[32] Nesterov, Yu. Gradient methods for minimizing composite functions. *Mathematical Programming* **140** (2013) 125-161.

[33] Nesterov, Yu. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization* **22** (2012) 341–362.

[34] Nesterov, Yu. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady* **27** (1983) 372–376.

[35] Polyak, B.T. Minimization of unsmooth functionals. *U.S.S.R. Computational Mathematics and Mathematical Physics* **9** (1969) 14–29.

[36] Polyak, B.T. A general method for solving extremal problems. *Soviet Mathematics Doklady* **8** (1967) 593-597.

[37] Richtárik, P., Takác, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming* **144** (2014) 1–38.

[38] Shor, N.Z. *Minimization methods for nondifferentiable functions*. Springer, Berlin (1985).

[39] Svaiter, B.F. A class of Fejér convergent algorithms, approximate resolvents and the hybrid Proximal-Extragradient method. *Journal of Optimization Theory and Application* **162** (2014) 133–153.

[40] Tropp, J. Just relax: convex programming methods for identifying sparse signals. *IEEE Transactions on Information Theory* **51** (2006) 1030–1051.

[41] Zhu, D.L., Marcotte, P. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization* **6** (1996) 714–726.