Open access • Journal Article • DOI:10.1007/S10957-017-1200-6

Extragradient Method in Optimization: Convergence and Complexity

Trong Phong Nguyen, Trong Phong Nguyen, Edouard Pauwels, Emile Richard ...+1 more authors

Institutions: University of Toulouse, University of Chile, Air Force Research Laboratory

Published on: 01 Jan 2018 - Journal of Optimization Theory and Applications (Springer US)

Topics: Rate of convergence and Line search

Official URL

DOI : https://doi.org/10.1007/s10957-017-1200-6

Extragradient Method in Optimization: Convergence and Complexity

Trong Phong Nguyen[*] Edouard Pauwels [†] Émile Richard [‡] Bruce W. Suter[§]

December 14, 2017

Abstract

We consider the extragradient method to minimize the sum of two functions, the first one being smooth and the second being convex. Under the Kurdyka-Łojasiewicz assumption, we prove that the sequence produced by the extragradient method converges to a critical point of the problem and has finite length. The analysis is extended to the case when both functions are convex. We provide, in this case, a sublinear convergence rate, as for gradient-based methods. Furthermore, we show that the recent *small-prox* complexity result can be applied to this method. Considering the extragradient method is an occasion to describe an exact line search scheme for proximal decomposition methods. We provide details for the implementation of this scheme for the one norm regularized least squares problem and demonstrate numerical results which suggest that combining nonaccelerated methods with exact line search can be a competitive choice.

Keywords Extragradient, descent method, forward-backward splitting method, Kurdyka-Łojasiewicz inequality, complexity, first order method, ℓ^1-regularized least squares.

1 Introduction

We introduce a new optimization method for approximating a global minimum of a composite objective function, i.e., a function formed as the sum of a smooth function and a simple nonsmooth convex function.

This class of problems is rich enough to encompass many smooth/nonsmooth, convex/nonconvex optimization problems considered in practice. Applications can be found in various fields throughout science and engineering, including signal/image processing [1] and machine learning [2]. Successful algorithms for these types of problems include for example fast iterative shrinkage-thresholding algorithm (FISTA) method [3] and forward-backward splitting method [4]. The goal of this paper is to investigate to which extent extragradient method can be used to tackle similar problems.

The extragradient method was initially proposed by Korpelevich [5] and it has become a classical method for solving variational inequality problems. For optimization problems, this method generates a sequence of estimates based on two projected gradient steps at each iteration.

After Korpelevich's work, a number of authors extended his extragradient method for variational inequality problems (for example, see [6,7]). In the context of convex constrained optimization, [8] considered

[*] TSE (GREMAQ, Université Toulouse I Capitole), Manufacture des Tabacs, 21 allée de Brienne, 31015 Toulouse, Cedex 06, France. Email: trong-phong.nguyen@ut-capitole.fr and Centro de Modelamiento Matemtico (UMI 2807, CNRS), Universidad de Chile, Beauchef 851, Casilla 170-3, Santiago 3, Chile

[†] IRIT-UPS, 118 route de Narbonne, 31062 Toulouse, France. Email: edouard.pauwels@irit.fr (Edouard Pauwels)

[‡] Amazon, ricemile@amazon.com

[§] Air Force. Research Laboratory / RITB, Rome, NY, United States of America. E-mail: bruce.suter@us.af.mil

the performances of the extragradient method under error bounds assumptions. In this setting, Luo and Tseng have described asymptotic linear convergence of the extragradient method applied to constrained problems. To our knowledge, this is the only attempt to analyse the method in an optimization setting.

A distinguishing feature of the extragradient method is its use of an additional projected gradient step, which can be seen as a guide during the optimization process. Intuitively, this additional iteration allows us to *examine* the geometry of the problem and take into account its curvature information, one of the most important bottlenecks for first order methods. Motivated by this observation, our goal is to extend and understand further the extragradient method in the specific setting of optimization. Apart from the work of Luo and Tseng, the literature on this topic is quite scarce. Moreover, the nonconvex case is not considered at all.

We combined the work of [5,8] and some recent extensions for first-order descent methods, (see [9–12]), to propose the extented extragradient method (EEG for short) to tackle the problem of minimizing a composite objective function. The classical extragradient method relies on orthogonal projections. We extend it by considering more general nonsmooth functions, and using proximal gradient steps at each iteration. An important challenge in this context is to balance the magnitude of the two associated parameters to maintain desirable convergence properties. We devise conditions which allow to prove convergence of this method in the nonconvex case. In addition, we describe two different rates of convergence in the convex setting.

Following [9–12] we heavily rely on the Kurdyka-Łojasiewicz (KL for short) inequality to study the nonconvex setting. The KL inequality [13, 14] has a long history in convergence analysis and smooth optimization. Recent generalizations in the seminal works [15, 16] have shown the important versatility of this approach as the inequality holds true for the vast majority of models encountered in practice, including nonsmooth and extended valued functions. This opened the possibility to devise general and abstract convergence results for first order methods [10, 11], which constitute an important ingredient of our analysis. Based on this approach, we derive a general convergence result for the proposed EEG method.

In the convex case, we focus on global convergence rates. We first describe a sublinear convergence rate in terms of objective function. This is related to classical results from the analysis of first order methods in convex optimization, see for example the analysis of forward-backward splitting method in [3]. Furthermore, we show that the *small-prox* result of [12] also applies to EEG method which echoes the error bound framework of Luo and Tseng [8] and opens the door to more refined complexity results when further properties of the objective function are available.

As already mentioned, a distinguished aspect of the extragradient method is its use of an additional proximal gradient step at each iteration. The intuition behind this mechanism is the incorporation of curvature information into the optimization process. It is expected that one of the effects of this additional step is to allow larger step sizes. With this in mind, we describe an exact line search variant of the method. Although computing exact line search is a nonconvex problem, potentially hard in the general case, we describe an active set method to tackle it for the specific and very popular case of the one norm regularized least squares problem (also known as the least absolute shrinkage and selection operator or LASSO). In this setting the computational overhead of exact line search is approximately equal to that of a gradient computation (discarding additional logarithmic terms).

On the practical side, we compare the performance of the proposed EEG method (and its line search variant) to those of FISTA and forward-backward splitting methods on the LASSO problem. The numerical results suggest that EEG combined with exact line search, constitutes a promising alternative which does not suffer too much from ill conditioning.

Structure of the paper. Section 2 introduces the problem and our main assumptions. We also recall important definitions and notations which will be used throughout the text. Section 3 contains the main convergence results of this paper. More precisely, in subsection 3.3, we present the convergence and finite length property under the KL assumption in the nonconvex case. Subsection 3.4, contains both a proof of sublinear convergence rate and the application of the *small-prox* result for EEG method leading to improved complexity analysis under the KL assumption. Section 4 describes exact line search for proximal gradient steps in the context of one norm regularized least squares and results from numerical experiments.

2 Optimization Setting and Some Preliminaries

2.1 Optimization Setting

We are interested in solving minimization problems of the form

$$\min_{x\in\mathbb{R}^n}\{F(x) := f(x) + g(x)\}, \tag{\textbf{P}}$$

where f, g are extended value functions from $\mathbb{R}^n$ to $]-\infty,+\infty]$. We make the following standing assumptions:

- $\operatorname{argmin} F \neq \emptyset$, and we note $F^* := \min_{x\in\mathbb{R}^n} F(x)$.
- g is a lower semi-continuous, convex, proper function.
- f is differentiable with L-Lipschitz continuous gradient, where $L > 0$.

2.2 Nonsmooth Analysis

In this subsection, we recall the definitions, notations and some well-known results from nonsmooth analysis which are going to be used throughout the paper. We will use notations from [17] (see also [18]). Let $h\colon \mathbb{R}^n \to]-\infty,+\infty]$ be a proper, lower-semicontinuous function. For each $x \in \operatorname{dom} h$, the Fréchet subdifferential of h at x, written $\hat{\partial}h(x)$, is the set of vectors $u \in \mathbb{R}^n$ which satisfy

$$\liminf_{y\to x} \frac{h(y) - h(x) - \langle u, y - x\rangle}{\|x - y\|} \geq 0.$$

When $x \notin \operatorname{dom} h$, we set $\hat{\partial}h(x) := \emptyset$. We will use the following set

$$\operatorname{graph}(\hat{\partial}h) := \left\{(x,u) \in \mathbb{R}^n \times \mathbb{R}^n\colon\ u \in \hat{\partial}h(x)\right\}.$$

The (limiting) subdifferential of h at $x \in \operatorname{dom} h$ is defined by the following closure process

$$\partial h(x) := \left\{u \in \mathbb{R}^n\colon \exists\, (x_m, u_m)_{m\in\mathbb{N}} \in \operatorname{graph}(\hat{\partial}h)^{\mathbb{N}},\ x_m \underset{m\to\infty}{\to} x,\ h(x_m) \underset{m\to\infty}{\to} h(x),\ u_m \underset{m\to\infty}{\to} u\right\}.$$

$\operatorname{graph}(\partial h)$ is defined similarly as $\operatorname{graph}(\hat{\partial}h)$. When h is convex, the above definition coincides with the usual notion of subdifferential in convex analysis

$$\partial h(x) := \{u \in \mathbb{R}^n\colon\ h(y) \geq h(x) + \langle u, y - x\rangle \text{ for all } y \in \mathbb{R}^n\}.$$

Independently, from the definition, when h is smooth at x then the subdifferential is a singleton, $\partial h(x) = \{\nabla h(x)\}$.

We can deduce from its definition the following closedness property of the subdifferential: if a sequence $(x_m, u_m)_{m\in\mathbb{N}} \in \operatorname{graph}(\partial h)^{\mathbb{N}}$, converges to (x, u), and $h(x_m)$ converges to $h(x)$ then $u \in \partial h(x)$. The set $\operatorname{crit} h := \{x \in \mathbb{R}^n : 0 \in \partial h(x)\}$ is called the set of critical points of h. In this nonsmooth context, Fermat's rule remains unchanged: A necessary condition for x to be local minimizer of h is that $x \in \operatorname{crit} h$ [17, Theorem 10.1].

Under our standing assumption, f is a smooth function and we have subdifferential sum rule [17, Exercise 10.10]

$$\partial(f + h)(x) = \nabla f(x) + \partial h(x). \tag{1}$$

We recall a well known important property of smooth functions which have L-Lipschitz continuous gradient, (see [19, Lemma 1.2.3]).

Lemma 1 (Descent Lemma) *For any $x, y \in \mathbb{R}^n$, we have*

$$f(y) \leq f(x) + \langle y - x, \nabla f(x)\rangle + \frac{L}{2}\|x - y\|^2.$$

For the rest of this paragraph, we suppose that h is a convex function. Given $x \in \mathbb{R}^n$ and $t > 0$, the proximal operator associated to h, which we denote by $\operatorname{prox}_{th}(x)$, is defined as the unique minimizer of function $y \longmapsto\ h(y) + \frac{1}{2t}\|y - x\|^2$, i.e:

$$\operatorname{prox}_{th}(x) := \operatorname{argmin}_{y\in\mathbb{R}^n}\left\{h(y) + \frac{1}{2t}\|y - x\|^2\right\}.$$

Using Fermat's Rule, $\operatorname{prox}_{th}(x)$ is characterized as the unique solution of the inclusion

$$\frac{x - \operatorname{prox}_{th}(x)}{t} \in \partial h\left(\operatorname{prox}_{th}(x)\right).$$

We recall that if h is convex, then prox_{th} is nonexpansive, that is Lipschitz continuous with constant 1 (see [18, Proposition 12.27]). As an illustration, let $C \subset \mathbb{R}^n$ be a closed, convex and nonempty set, then $\operatorname{prox}_{i_C}$ is the orthogonal projection operator onto C. The following property of the prox mapping will be used in the analysis, (see [3, Lemma 1.4]).

Lemma 2 *Let $u \in \mathbb{R}^n$, $t > 0$, and $v = \operatorname{prox}_{th}(u)$, then*

$$h(w) - h(v) \geq \frac{1}{2t}\ \ \|u - v\|^2 + \|w - v\|^2 - \|u - w\|^2\big), \forall w \in \mathbb{R}^n.$$

2.3 Nonsmooth Kurdyka-Łojasiewicz Inequality

In this subsection, we present the nonsmooth Kurdyka-Łojasiewicz inequality introduced in [15] (see also [16, 20], and the fundamental works [13, 14]). We note $[h < \mu] := \{x \in \mathbb{R}^n :\ h(x) < \mu\}$ and $[\eta < h < \mu] := \{x \in \mathbb{R}^n :\ \eta < h(x) < \mu\}$. Let $r_0 > 0$ and set

$$\mathcal{K}(r_0) := \left\{\varphi \in C^0\left([0, r_0[\right) \cap C^1\left(]0, r_0[\right), \varphi(0) = 0, \varphi \text{ is concave and } \varphi' > 0\right\}.$$

Definition 1 *The function h satisfies the* Kurdyka-Łojasiewicz (KL) inequality *(or has the KL* property*) locally at $\bar{x} \in \operatorname{dom} h$ if there exist $r_0 > 0$, $\varphi \in \mathcal{K}(r_0)$ and a neighborhood $U(\bar{x})$ such that*

$$\varphi'\left(h(x) - h(\bar{x})\right) \operatorname{dist}\left(0, \partial h(x)\right) \geq 1 \tag{2}$$

for all $x \in U(\bar{x}) \cap [h(\bar{x}) < h(x) < h(\bar{x}) + r_0]$. We say that φ is a desingularizing function *for h at $\bar{x}$. The function h has the KL property on S if it does so at each point of S.*

When h is smooth and $h(\bar{x}) = 0$ then (2) can be rewritten as

$$\|\nabla(\varphi \circ h)\| \geq 1, \, \forall x \in U(\bar{x}) \cap [0 < h(x) < r_0] \, .$$

This inequality may be interpreted as follows: The function h can be made sharp locally by a reparameterization of its values through a function $\varphi \in \mathcal{K}(r_0)$ for some $r_0 > 0$.

The KL inequality is obviously satisfied at any noncritical point $\bar{x} \in \operatorname{dom} h$ and will thus be useful only for critical points, $\bar{x} \in \operatorname{crit} h$. The *Łojasiewicz gradient inequality* corresponds to the case when $\varphi(s) = cs^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1[$. The class of functions which satisfy KL inequality is extremely vast. Typical KL functions are semi-algebraic functions, but there exists many extensions, (see [15]).

If h has the KL property and admits the same desingularizing function φ at *every point*, then we say that φ is a *global* desingularizing function for f. The following lemma is given in [11, Lemma 6].

Lemma 3 (Uniformized KL property) *Let Ω be a compact set and let $h : \mathbb{R}^n \to]-\infty, \infty]$ be a proper and lower semicontinuous function. We assume that h is constant on Ω and satisfies the KL property at each point of Ω. Then, there exist $\varepsilon > 0$, $\eta > 0$ and φ such that for all $\bar{x} \in \Omega$, one has*

$$\varphi'(h(x) - h(\bar{x})) \operatorname{dist}(0, \partial h(x)) \geq 1,$$

for all $x \in \{x \in \mathbb{R}^n \colon \operatorname{dist}(x, \Omega) < \varepsilon\} \cap [h(\bar{x}) < h(x) < h(\bar{x}) + \eta]$.

3 Extragradient Method, Convergence and Complexity

3.1 Extragradient Method

We now describe our extragradient method dedicated to the minimization of problem (**P**). The method is defined, given an initial estimate $x_0 \in \mathbb{R}^n$, by the following recursion, for $k \geq 1$,

$$(EEG) \begin{cases} y_k := \operatorname{prox}_{s_k g}\left(x_k - s_k \nabla f(x_k)\right), & (3) \\ x_{k+1} := \operatorname{prox}_{\alpha_k g}\left(x_k - \alpha_k \nabla f(y_k)\right). & (4) \end{cases}$$

where $(s_k, \alpha_k)_{k \in \mathbb{N}}$ are positive step size sequences. We introduce relevant quantities, $s_- = \inf_{k \in \mathbb{N}} s_k$, $s^+ = \sup_{k \in \mathbb{N}} s_k$, $\alpha_- = \inf_{k \in \mathbb{N}} \alpha_k$ and $\alpha^+ = \sup_{k \in \mathbb{N}} \alpha_k$. Throughout the paper, we will consider the following condition on the two step size sequence,

$$(\mathbf{C}) : 0 < \alpha_-, \, 0 < s_-, \, s^+ < \frac{1}{L} \text{ and } 0 < s_k \leq \alpha_k, \, \forall k \in \mathbb{N}.$$

Depending on the context, additional restrictions will be imposed on the step size sequences.

3.2 Basic Properties

We introduce in this subsection two technical properties of sequences produced by EEG method. These two technical properties, as abstract conditions introduced in tame nonconvex settings [9–11], allow us to prove the convergence of the sequences. We begin with a technical lemma.

Lemma 4 *Let $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$, $t > 0$, and $p = \operatorname{prox}_{tg}(x - t\nabla f(y))$, then, for any $z \in \mathbb{R}^n$, we have*

(i) $F(z) - F(p) \geq \left(\frac{1}{2t} - \frac{L}{2}\right)\|p - z\|^2 + \frac{1}{2t}\ \|x - p\|^2 - \|x - z\|^2) + \langle p - z, \nabla f(y) - \nabla f(z)\rangle .$

(ii) $F(z) - F(p) \geq \frac{1}{2t}\ \|x - p\|^2 + \|z - p\|^2 - \|x - z\|^2) + \langle y - z, \nabla f(y)\rangle + f(z) - f(y) - \frac{L}{2}\|p - y\|^2.$

In addition, when f is convex, we get $\langle y - z, \nabla f(y)\rangle + f(z) - f(y) \geq 0$. Therefore, inequality (ii) implies that

$$F(z) - F(p) \geq \frac{1}{2t}\ \|x - p\|^2 + \|z - p\|^2 - \|x - z\|^2) - \frac{L}{2}\|p - y\|^2.$$

Proof : We apply Lemma 2 with $u = x - t\nabla f(y)$, $v = \operatorname{prox}_{tg}(u)$ and $z = w$ which leads to

$$\begin{aligned} g(z) - g(p) &\geq \frac{1}{2t}\left(\|x - t\nabla f(y) - p\|^2 + \|z - p\|^2 - \|x - t\nabla f(y) - z\|^2\right) \\ &= \frac{1}{2t}\left(\|x - p\|^2 + \|z - p\|^2 - \|x - z\|^2\right) + \langle p - z, \nabla f(y)\rangle . \end{aligned} \tag{5}$$

Now using the descent Lemma 1, we have that

$$f(z) - f(p) \geq -\frac{L}{2}\|p - z\|^2 - \langle \nabla f(z), p - z\rangle . \tag{6}$$

The first claimed inequality results from summation of (5) and (6). Now using the descent Lemma 1 again, we have that

$$\langle \nabla f(y), p - y\rangle \geq f(p) - f(y) - \frac{L}{2}\|p - y\|^2. \tag{7}$$

Combining (5) and (7), we obtain

$$g(z) - g(p) \geq \frac{1}{2t}\ \|x - p\|^2 + \|z - p\|^2 - \|x - z\|^2) + \langle y - z, \nabla f(y)\rangle + f(p) - f(y) - \frac{L}{2}\|p - y\|^2. \tag{8}$$

The second claimed inequality follows by adding $f(z) - f(p)$ to (8). This concludes the proof. □

We are now ready to describe a descent property for EEG method.

Proposition 5 (Descent condition) *For any $k \in \mathbb{N}$, we have*

$$F(x_k) - F(x_{k+1}) \geq \frac{1}{2\alpha_k}\|x_k - x_{k+1}\|^2 + \left(\frac{1}{s_k} - \frac{L}{2} - \frac{1}{2\alpha_k}\right)\|x_k - y_k\|^2 + \left(\frac{1}{2\alpha_k} - \frac{L}{2}\right)\|y_k - x_{k+1}\|^2.$$

Proof : We fix an arbitrary $k \in \mathbb{N}$. Applying inequality (i) of Lemma 4, with $x = x_k$, $y = x_k$, $t = s_k$, $p = y_k$ and $z = x_k$, we obtain

$$F(x_k) - F(y_k) \geq \left(\frac{1}{s_k} - \frac{L}{2}\right) \|x_k - y_k\|^2. \tag{9}$$

Similarly, applying inequality (i) of Lemma 4, with $x = x_k$, $y = y_k$, $t = \alpha_k$, $p = x_{k+1}$ and $z = y_k$, we obtain

$$F(y_k) - F(x_{k+1}) \geq \frac{1}{2\alpha_k}\left(\|x_k - x_{k+1}\|^2 - \|y_k - x_k\|^2\right) + \left(\frac{1}{2\alpha_k} - \frac{L}{2}\right) \|y_k - x_{k+1}\|^2. \tag{10}$$

Combining inequalities (9) and (10), we obtain

$$F(x_k) - F(x_{k+1}) \geq \frac{1}{2\alpha_k}\|x_k - x_{k+1}\|^2 + \left(\frac{1}{s_k} - \frac{L}{2} - \frac{1}{2\alpha_k}\right) \|x_k - y_k\|^2 + \left(\frac{1}{2\alpha_k} - \frac{L}{2}\right) \|y_k - x_{k+1}\|^2,$$

which concludes the proof □

Remark 6 *If we combine the constraint that $0 < \alpha_k \leq \frac{1}{L}$ for all $k \in \mathbb{N}$ with condition (**C**), we deduce from Proposition 5 that, for all $k \in \mathbb{N}$, $\frac{1}{s_k} - \frac{L}{2} - \frac{1}{2\alpha_k} \geq 0$, and*

$$F(x_k) - F(x_{k+1}) \geq \frac{1}{2\alpha_k}\|x_k - x_{k+1}\|^2.$$

Under this condition, we have that EEG is a descent method in the sense that it will produce a decreasing sequence of objective value.

We now establish a second property of sequences produced by EEG method which is interpreted as a subgradient step property. We begin with a technical Lemma.

Lemma 7 *Assume that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition (**C**). Then, for any $k \in \mathbb{N}$, it holds that*

$$\|x_{k+1} - y_k\| \leq \left(\frac{1}{1 - Ls_k} - \frac{s_k}{\alpha_k}\right) \|x_k - x_{k+1}\|. \tag{11}$$

Proof : Denote $z_{k+1} = \mathrm{prox}_{s_k g}(x_k - s_k \nabla f(y_k))$, since $\mathrm{prox}_{s_k g}$ is 1-Lipschitz continuous, we get

$$\begin{aligned}\|y_k - z_{k+1}\| &\leq \|(x_k - s_k\nabla f(y_k)) - (x_k - s_k\nabla f(x_k))\|\\ &\leq Ls_k\|x_k - y_k\|,\end{aligned}$$

where the second inequality follows from the fact that ∇f is L–Lipschitz continuous. Therefore,

$$\|x_k - z_{k+1}\| \geq \|x_k - y_k\| - \|y_k - z_{k+1}\| \geq (1 - Ls_k)\|x_k - y_k\|. \tag{12}$$

Writing the optimality condition for (4), yields that

$$\frac{x_k - x_{k+1}}{\alpha_k} - \nabla f(y_k) \in \partial g(x_{k+1}), \tag{13}$$

and the convexity of g implies

$$\left\langle \frac{x_k - x_{k+1}}{\alpha_k} - \nabla f(y_k), z_{k+1} - x_{k+1} \right\rangle \leq g(z_{k+1}) - g(x_{k+1}).$$

Similarly, using the definition of z_{k+1}, we get

$$\left\langle \frac{x_k - z_{k+1}}{s_k} - \nabla f(y_k), x_{k+1} - z_{k+1} \right\rangle \leq g(x_{k+1}) - g(z_{k+1}).$$

Adding the last two inequalities, we obtain

$$\left\langle \frac{x_k - z_{k+1}}{s_k} - \frac{x_k - x_{k+1}}{\alpha_k}, x_{k+1} - z_{k+1} \right\rangle \leq 0,$$

or equivalently

$$\left\langle \frac{x_k - z_{k+1}}{s_k} - \frac{x_k - x_{k+1}}{\alpha_k}, (x_{k+1} - x_k) + (x_k - z_{k+1}) \right\rangle \leq 0.$$

It follows that

$$\frac{\|x_k - z_{k+1}\|^2}{s_k} + \frac{\|x_k - x_{k+1}\|^2}{\alpha_k} \leq \left(\frac{1}{s_k} + \frac{1}{\alpha_k}\right) \langle x_k - z_{k+1}, x_k - x_{k+1} \rangle .$$

Using the Cauchy-Schwarz inequality, we get

$$\frac{\|x_k - z_{k+1}\|^2}{s_k} + \frac{\|x_k - x_{k+1}\|^2}{\alpha_k} \leq \left(\frac{1}{s_k} + \frac{1}{\alpha_k}\right) \|x_k - z_{k+1}\|.\|x_k - x_{k+1}\|.$$

Since from condition (**C**), $0 < s_k$, this is equivalent to

$$(\|x_k - z_{k+1}\| - \|x_k - x_{k+1}\|) \left(\|x_k - z_{k+1}\| - \frac{s_k \|x_k - x_{k+1}\|}{\alpha_k}\right) \leq 0.$$

This inequality asserts that the product of two terms is nonpositive. Hence one of the terms must be nonpositive and the other one must be nonnegative. From condition (**C**), we have $\frac{s_k}{\alpha_k} \leq 1$, the last term is bigger than the first one and hence must be nonnegative. This yields

$$\frac{s_k}{\alpha_k} \|x_k - x_{k+1}\| \leq \|x_k - z_{k+1}\| \leq \|x_k - x_{k+1}\|.$$

By combining the latter inequality with (12), we get

$$(1 - Ls_k)\|x_k - y_k\| \leq \|x_k - z_{k+1}\| \leq \|x_k - x_{k+1}\|. \tag{14}$$

Similarly, from the definitions of y_k, x_{k+1} and the convexity of g, we obtain that

$$\left\langle \frac{x_k - y_k}{s_k} - \nabla f(x_k), x_{k+1} - y_k \right\rangle \leq g(x_{k+1}) - g(y_k),$$

and

$$\left\langle \frac{x_k - x_{k+1}}{\alpha_k} - \nabla f(y_k), y_k - x_{k+1} \right\rangle \leq g(y_k) - g(x_{k+1}).$$

Summing the last two inequalities, we have that

$$\frac{1}{s_k}\|x_{k+1} - y_k\|^2 + \left(\frac{1}{s_k} - \frac{1}{\alpha_k}\right) \langle x_{k+1} - y_k, x_k - x_{k+1} \rangle \leq \langle x_{k+1} - y_k, \nabla f(x_k) - \nabla f(y_k) \rangle.$$

Using the condition $0 < s_k \leq \alpha_k$ and the Cauchy-Schwarz inequality, we get

$$\frac{1}{s_k}\|x_{k+1} - y_k\|^2 \leq \left(\frac{1}{s_k} - \frac{1}{\alpha_k}\right)\|x_{k+1} - y_k\|\|x_k - x_{k+1}\| + \|x_{k+1} - y_k\|\|\nabla f(x_k) - \nabla f(y_k)\|.$$

Using the Lipschitz continuity of ∇f, we have that

$$\|x_{k+1} - y_k\| \leq \left(1 - \frac{s_k}{\alpha_k}\right)\|x_k - x_{k+1}\| + Ls_k\|x_k - y_k\|.$$

Combining this inequality with (14), we obtain

$$\begin{aligned}\|x_{k+1} - y_k\| &\leq \left(1 - \frac{s_k}{\alpha_k} + \frac{Ls_k}{1 - Ls_k}\right)\|x_k - x_{k+1}\| \\ &= \left(\frac{1}{1 - Ls_k} - \frac{s_k}{\alpha_k}\right)\|x_k - x_{k+1}\|, \qquad (15)\end{aligned}$$

which is the required inequality. □

We are now ready to prove the subgradient step property which is the second main element of the convergence proof.

Proposition 8 (Subgradient step) *Assume that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition (**C**). Then, for any $k \in \mathbb{N}$, there exists $u_{k+1} \in \partial g(x_{k+1})$ such that*

$$\|u_{k+1} + \nabla f(x_{k+1})\| \leq \frac{L\alpha_k + (1 - Ls_k)^2}{\alpha_k(1 - Ls_k)}\|x_k - x_{k+1}\|.$$

Proof : Thanks to (13), we deduce that there exists $u_{k+1} \in \partial g(x_{k+1})$ such that

$$\frac{x_k - x_{k+1}}{\alpha_k} + \nabla f(x_{k+1}) - \nabla f(y_k) = u_{k+1} + \nabla f(x_{k+1}).$$

This implies that

$$\|u_{k+1} + \nabla f(x_{k+1})\| \leq \frac{\|x_k - x_{k+1}\|}{\alpha_k} + \|\nabla f(x_{k+1}) - \nabla f(y_k)\|.$$

Since ∇f is L-Lipschitz continuous, it follows that

$$\|u_{k+1} + \nabla f(x_{k+1})\| \leq \frac{\|x_k - x_{k+1}\|}{\alpha_k} + L\|x_{k+1} - y_k\|. \qquad (16)$$

Combining Lemma 7 with (16), we get

$$\|u_{k+1} + \nabla f(x_{k+1})\| \leq \frac{L\alpha_k + (1 - Ls_k)^2}{\alpha_k(1 - Ls_k)}\|x_k - x_{k+1}\|, \qquad (17)$$

and the result is proved. □

Combining Remark 6 and Proposition 8 above, we have the following corollary which underlines the fact that EEG is actually an approximate gradient method in the sense of [10].

Corollary 9 *Assume that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy the following*

$$(\mathbf{C1}) : (s_k, \alpha_k)_{k\in\mathbb{N}} \text{ satisfy condition } (\mathbf{C}) \text{ and } \alpha_k \leq \frac{1}{L},\ \forall k \in \mathbb{N}.$$

Then, for all $k \in \mathbb{N}$

(i) $F(x_{k+1}) + \frac{1}{2\alpha_k}\|x_k - x_{k+1}\|^2 \leq F(x_k)$.

(ii) *There exists* $\omega_{k+1} \in \partial F(x_{k+1})$ *such that*

$$\|\omega_{k+1}\| \leq b_k \|x_k - x_{k+1}\|,$$

where,

$$0 < b_k := \frac{L\alpha_k + (1 - Ls_k)^2}{\alpha_k(1 - Ls_k)} \leq b := \frac{2}{\alpha_-(1 - s^+ L)}.$$

3.3 Convergence of EEG Method under KL Assumption

In this subsection, we analyse the convergence of EEG method in the nonconvex setting. The main result is stated in Theorem 11, which also describes the asymptotic rate of convergence. This result is based on the assumptions that F has the KL property on $\operatorname{crit} F$ and that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy conditions (**C1**) from Corollary 9. We will also assume that the sequence $(x_k)_{k\in\mathbb{N}}$ generated by EEG is bounded. This boundedness assumption is not very restrictive here, since under condition (**C1**), Corollary 9 ensures that it is satisfied for any coercive objective function. Similarly to [11, Lemma 3.5], we first give some properties of F on the set of accumulation points of $(x_k)_{k\in\mathbb{N}}$.

Lemma 10 *Assume that the sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by EEG method is bounded and that* $(s_k, \alpha_k)_{k\in\mathbb{N}}$ *satisfy condition (***C1***). Let* Ω_0 *be the set of limit points of the sequence* $(x_k)_{k\in\mathbb{N}}$*. It holds that* Ω_0 *is compact and nonempty,* $\Omega_0 \subset \operatorname{crit} F$*,* $\operatorname{dist}(x_k, \Omega_0) \to 0$ *and* $F(\bar{x}) = \lim_{k\to\infty} F(x_k)$ *for all* $\bar{x} \in \Omega_0$*.*

Proof : From the boundedness assumption, it is clear that Ω_0 is nonempty. In view of Corollary 9 **i)**, it follows that $(F(x_k))_{k\in\mathbb{N}}$ is nonincreasing. Furthermore, $F(x_k)$ is bounded from below by F^*, hence there exists $\bar{F} \in \mathbb{R}$ such that $\bar{F} = \lim_{k\to\infty} F(x_k)$. In addition, we have

$$\sum_{k=1}^{m} \|x_{k+1} - x_k\|^2 \leq 2\alpha^+ \left(F(x_1) - F(x_{m+1})\right),$$

therefore $\sum_{k=1}^{\infty} \|x_{k+1} - x_k\|^2$ converges, thus $(x_{k+1} - x_k) \to 0$. We now fix an arbitrary point $x^* \in \Omega_0$, which means that there exists a subsequence $(x_{k_q})_{q\in\mathbb{N}}$ of $(x_k)_{k\in\mathbb{N}}$ such that $\lim_{q\to\infty} x_{k_q} = x^*$, therefore, by lower semicontinuity of g and continuity of f,

$$g(x^*) \leq \liminf_{q\to\infty} g(x_{k_q}), \; f(x^*) = \lim_{q\to\infty} f(x_{k_q}). \tag{18}$$

From the definition of x_{k_q} and condition (**C1**), we get for all $q \in \mathbb{N}$,

$$\begin{aligned}
&g(x_{k_q}) + \frac{1}{2s_+}\|x_{k_q-1} - x_{k_q}\|^2 + \left\langle x_{k_q} - x_{k_q-1}, \nabla f(y_{k_q-1})\right\rangle \\
&\leq g(x_{k_q}) + \frac{1}{2s_{k_q}}\|x_{k_q-1} - x_{k_q}\|^2 + \left\langle x_{k_q} - x_{k_q-1}, \nabla f(y_{k_q-1})\right\rangle \\
&\leq g(x^*) + \frac{1}{2s_{k_q}}\|x^* - x_{k_q-1}\|^2 + \left\langle x^* - x_{k_q-1}, \nabla f(y_{k_q-1})\right\rangle. \\
&\leq g(x^*) + \frac{1}{2s_-}\|x^* - x_{k_q-1}\|^2 + \left\langle x^* - x_{k_q-1}, \nabla f(y_{k_q-1})\right\rangle.
\end{aligned}$$

Let $q \to \infty$, it follows that $\limsup_{q\to\infty} g(x_{k_q}) \leq g(x^*)$, thus, in view of (18), $\lim_{q\to\infty} g(x_{k_q}) = g(x^*)$, therefore $\lim_{q\to\infty} F(x_{k_q}) = F(x^*)$. Since $F(x_k)$ is nonincreasing, $\lim_{q\to\infty} F(x_{k_q}) = \bar{F}$, and we deduce that

$F(x^*) = \bar{F}$. Since x^* was arbitrary in Ω_0, it holds that F is constant on Ω_0.

Now, thanks to Corollary 9 **ii)**, there exist $\omega_{k+1} \in \partial F(x_{k+1})$, such that

$$\|\omega_{k+1}\| \leq b_k \|x_k - x_{k+1}\|.$$

Under condition (**C1**), it holds that b_k remains bounded. Since $\lim_{k\to\infty} x_k - x_{k+1} = 0$, it holds that $\omega_k \to 0$. Combining with the closedness of ∂F, this implies that $0 \in \partial F(x^*)$, hence $x^* \in \operatorname{crit} F$. Since x^* was taken arbitrarily in Ω_0, this means that $\Omega_0 \subset \operatorname{crit} F$. The compactness of Ω_0 is implied by [11, Lemma 5]. Combining the boundedness of $(x_k)_{k\in\mathbb{N}}$ and the compactness of Ω_0, we deduce that $\operatorname{dist}(x_k, \Omega_0) \to 0$ which concludes the proof. □

By combining Corollary 9, Lemma 3, 10 and using the methodology of [11, Theorem 1], we obtain a proof of convergence of EEG method in the non-convex case.

Theorem 11 *Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by EEG method which is assumed to be bounded. Suppose that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition (**C1**) and that F has the KL property on $\operatorname{crit} F$. Then, the sequence $(x_k)_{k\in\mathbb{N}}$ converges to $x^* \in \operatorname{crit} F$, moreover*

$$\sum_{i=1}^{\infty} \|x_k - x_{k+1}\| < \infty.$$

Proof : The proof is similar to the proof of [11, Theorem 1] and will be omitted. □

Remark 12 (Convergence rate) *When the KL desingularizing function of F is of the form $\varphi(s) = cs^{1-\theta}$, where c is a positive constant and $\theta \in (0, 1]$, then we can estimate the rate of convergence of the sequence $(x_k)_{k\in\mathbb{N}}$, as follows (see [9, Theorem 2]).*

- *$\theta = 0$ then the sequence (x_k) converges in a finite number of steps.*
- *$\theta \in \left[0, \frac{1}{2}\right]$ then there exist $C > 0$ and $\tau \in (0, 1)$ such that*

$$\|x_k - x^*\| \leq C\tau^k, \forall k \in \mathbb{N}.$$

- *$\theta \in \left]\frac{1}{2}, 1\right[$ then there exist $C > 0$ such that*

$$\|x_k - x^*\| \leq Ck^{-\frac{1-\theta}{2\theta-1}}, \forall k \in \mathbb{N}.$$

3.4 The Complexity of EEG in the Convex Case

Throughout this section, we suppose that the function f is convex and we focus on complexity and non asymptotic convergence rate analysis.

3.4.1 Sublinear Convergence Rate Analysis

We begin with a technical Lemma which introduces more restrictive step size conditions.

Lemma 13 *Assume that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy the following*

$$(\mathbf{C2})\text{: } (s_k, \alpha_k)_{k\in\mathbb{N}} \text{ satisfy condition } (\mathbf{C}) \text{ and } s_k \leq \frac{1}{2L},\ \alpha_k \leq \frac{1}{L} - s_k,\ \forall k \in \mathbb{N}.$$

Then, for all $k \in \mathbb{N}$,

$$\frac{1}{\alpha_k}\|x_k - x_{k+1}\|^2 - L\|x_{k+1} - y_k\|^2 \geq 0.$$

Proof : First, we note that if $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition (**C2**) then they also satisfy condition (**C1**) and Proposition 8 applies. Thanks to Lemma 7, we get

$$\begin{aligned}\frac{1}{\alpha_k}\|x_k - x_{k+1}\|^2 - L\|x_{k+1} - y_k\|^2 &\geq \frac{1}{\alpha_k}\|x_k - x_{k+1}\|^2 - L\left(\frac{1}{1-Ls_k} - \frac{s_k}{\alpha_k}\right)^2 \|x_k - x_{k+1}\|^2 \\ &= \frac{-L\alpha_k^2 + (1-s_k^2L^2)\alpha_k - Ls_k^2(1-Ls_k)^2}{\alpha_k^2(1-Ls_k)^2}\|x_k - x_{k+1}\|^2. \qquad (19)\end{aligned}$$

In addition, it can be checked using elementary calculation that

$$-L\alpha_k^2 + (1-s_k^2L^2)\alpha_k - Ls_k^2(1-Ls_k)^2 \geq 0,$$

is equivalent to

$$(1-Ls_k)\frac{(1+Ls_k) - \sqrt{(1+Ls_k)^2 - 4L^2s_k^2}}{2L} \leq \alpha_k \leq (1-Ls_k)\frac{(1+Ls_k) + \sqrt{(1+Ls_k)^2 - 4L^2s_k^2}}{2L}. \qquad (20)$$

Note that, for $0 \leq b \leq a$ then $a - b \leq \sqrt{a^2 - b^2}$. Using this inequality, with the condition $2Ls_k \leq 1$, we get $(1 + Ls_k) - 2Ls_k \leq \sqrt{(1+Ls_k)^2 - 4L^2s_k^2}$. Thus,

$$\begin{aligned}(1-Ls_k)\frac{(1+Ls_k) - \sqrt{(1+Ls_k)^2 - 4L^2s_k^2}}{2L} &\leq (1-Ls_k)\frac{[(1+Ls_k) - (1-Ls_k)]}{2L} \\ &= (1-Ls_k)s_k \leq s_k,\end{aligned}$$

and

$$(1-Ls_k)\frac{(1+Ls_k) + \sqrt{(1+Ls_k)^2 - 4L^2s_k^2}}{2L} \geq (1-Ls_k)\frac{(1+Ls_k) + (1-Ls_k)}{2L} = \frac{1}{L} - s_k.$$

Condition (**C2**) ensures that $s_k \leq \alpha_k \leq \frac{1}{L} - s_k$ and hence identity (20) holds and (19) implies that

$$\frac{1}{\alpha_k}\|x_k - x_{k+1}\|^2 - L\|x_{k+1} - y_k\|^2 \geq 0, \, \forall k \in \mathbb{N}.$$

□

With a similar method as in [3], we prove a sublinear convergence rate for $(F(x_k))_{k\in\mathbb{N}}$ in the convex case.

Theorem 14 (Complexity of EEG method) *Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by EEG method. Suppose that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition (**C2**) and that f is convex. Then, for any $x^* \in \operatorname{argmin} F$, we have*

$$F(x_m) - F(x^*) \leq \frac{1}{2m\alpha_-}\|x_0 - x^*\|^2, \, \forall m \in \mathbb{N}^*.$$

Proof : We first fix arbitrary $k \in \mathbb{N}$ and $x^* \in \operatorname{argmin} F$. Since f is convex, applying inequality (ii) of Lemma 4 with $x = x_k$, $y = y_k$, $t = \alpha_k$, $p = x_{k+1}$ and $z = x^*$, we obtain

$$F(x^*) - F(x_{k+1}) \geq \frac{1}{2\alpha_k} \quad \|x^* - x_{k+1}\|^2 - \|x^* - x_k\|^2) + \frac{1}{2\alpha_k}\|x_k - x_{k+1}\|^2 - \frac{L}{2}\|x_{k+1} - y_k\|^2.$$

Using the fact that $F(x_k)$ is noninreasing and bounded from bellow by $F(x^*)$, it follows from Lemma 13 that

$$0 \geq F(x^*) - F(x_{k+1}) \geq \frac{1}{2\alpha_k} \quad \|x^* - x_{k+1}\|^2 - \|x^* - x_k\|^2) \geq \frac{1}{2\alpha_-} \quad \|x^* - x_{k+1}\|^2 - \|x^* - x_k\|^2) \,.$$

Summing this inequality for $k = 0, \cdots, m-1$ gives

$$mF(x^*) - \sum_{k=1}^{m} F(x_k) \geq \frac{1}{2\alpha_-}(\|x^* - x_m\|^2 - \|x^* - x_0\|^2). \tag{21}$$

Coming back to Corollary 9, it is easy to see that the sequence $(F(x_k))_{k\in\mathbb{N}}$ is nonincreasing, then $\sum_{k=1}^{m} F(x_k) \geq mF(x_m)$. Combining with (21), we get

$$m\left(F(x^*) - F(x_m)\right) \geq \frac{1}{2\alpha_-} \ \ \|x^* - x_m\|^2 - \|x^* - x_0\|^2\right).$$

It follows that

$$F(x_m) - F(x^*) \leq \frac{1}{2m\alpha_-}\|x^* - x_0\|^2, \ \forall m \in \mathbb{N}^*.$$

□

3.4.2 Small-Prox Type Result under KL Property

We now study the complexity of EEG method when F has, in addition, the KL property on crit F. First, using the convexity of f, Proposition 5, can be improved by using the following result.

Proposition 15 *Assume that f is convex and $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition* (**C**)*, then for all $k \in \mathbb{N}$, we have*

$$F(x_k) - F(x_{k+1}) \geq c_k\|x_k - x_{k+1}\|^2,$$

where

$$c_k := \frac{1}{\alpha_k} - \frac{L}{2}\left(\frac{1}{1 - Ls_k} - \frac{s_k}{\alpha_k}\right)^2.$$

Proof : Fix an arbitrary $k \in \mathbb{N}$. Since f is convex, applying inequality (ii) of Lemma 4, with $x = x_k$, $y = y_k$, $t = \alpha_k$, $p = x_{k+1}$ and $z = x_k$, we get

$$F(x_k) - F(x_{k+1}) \geq \frac{1}{\alpha_k}\|x_k - x_{k+1}\|^2 - \frac{L}{2}\|x_{k+1} - y_k\|^2. \tag{22}$$

Combining inequality (22) with Lemma 7, we get the desired result,

$$F(x_k) - F(x_{k+1}) \geq \left[\frac{1}{\alpha_k} - \frac{L}{2}\left(\frac{1}{1 - Ls_k} - \frac{s_k}{\alpha_k}\right)^2\right]\|x_k - x_{k+1}\|^2.$$

□

We now consider another step size condition.

Lemma 16 *Suppose that s_k, α_k satisfy the following condition*

$$(\mathbf{C3}) \begin{cases} s_k, \alpha_k \textit{ satisfy condition } (\mathbf{C}) \\ s_k \leq \frac{\sqrt{5}-1}{2L}, \quad \textit{and } \alpha_k \leq \frac{2}{L} - 2s_k - (1 - Ls_k)Ls_k^2, \ \forall k \in \mathbb{N}. \end{cases}$$

Then, for all $k \in \mathbb{N}$,

$$\frac{1}{\alpha_k} - \frac{L}{2}\left(\frac{1}{1 - Ls_k} - \frac{s_k}{\alpha_k}\right)^2 \geq C := \frac{L^3 s_-^2(1 + Ls_-)}{2(2 - L^2 s_-^2)^2(1 - Ls_-)} > 0.$$

Remark 17 *Before starting the proof, we make a comment on the restriction $s_k \leq \frac{\sqrt{5}-1}{2L}$, which is only presented here to ensure consistency of condition (**C**) and condition (**C3**). By analysing a degree three polynomial, one can check that*

$$Ls_k \leq 2 - 2Ls_k - (1 - Ls_k)L^2 s_k^2$$

if and only if

$$Ls_k \in \left[-\frac{\sqrt{5}+1}{2}, \frac{\sqrt{5}-1}{2}\right] \cup [2, +\infty[\,.$$

*Hence the bound $s_k \leq \frac{\sqrt{5}-1}{2L}$ is a necessary condition to ensures that $s_k \leq \frac{2}{L} - 2s_k - (1 - Ls_k)Ls_k^2$. This upper limit on s_k could be removed from condition (**C3**), but then it would be enforced implicitly by the combination of conditions (**C**) and (**C3**) which results in $s_k \leq \alpha_k \leq \frac{2}{L} - 2s_k - (1 - Ls_k)Ls_k^2$. We preferred to write it explicitly.*

Proof : We fix an arbitrary $k \in \mathbb{N}$. Set

$$\alpha_k^+ = \frac{2}{L} - 2s_k - (1 - Ls_k)Ls_k^2 \tag{23}$$

$$Q(u) = u - \frac{1}{2}\left(\frac{1}{1 - Ls_k} - Ls_k u\right)^2, \tag{24}$$

where one can think of u satisfying $u = \frac{1}{L\alpha_k} \in \left[\frac{1}{L\alpha_k^+}, \frac{1}{Ls_k}\right]$. The maximum of $Q(u)$ is attained for $u = \frac{1}{(1-Ls_k)L^2 s_k^2} \geq \frac{1}{Ls_k}$, and the inequality stands because $Ls_k \leq 1$ and $1 - Ls_k \leq 1$. Note that conditions (**C**) and (**C3**) ensure that $\alpha_k^+ \geq s_k$, hence Q is increasing on $\left[\frac{1}{L\alpha_k^+}, \frac{1}{Ls_k}\right]$. Combining conditions (**C**) and (**C3**), we have that $s_k \leq \alpha_k \leq \alpha_k^+$ and therefore,

$$LQ\left(\frac{1}{L\alpha_k}\right) = \frac{1}{\alpha_k} - \frac{L}{2}\left(\frac{1}{1 - Ls_k} - \frac{s_k}{\alpha_k}\right)^2 \geq LQ\left(\frac{1}{L\alpha_k^+}\right). \tag{25}$$

We now turn to algebraic manipulations to compute $LQ\left(\frac{1}{L\alpha_k^+}\right)$. First we expand and reduce to common denominator.

$$\begin{aligned} & LQ\left(\frac{1}{L\alpha_k^+}\right) \\ &= L\left(\frac{1}{L\alpha_k^+} - \frac{1}{2}\left(\frac{1}{1 - Ls_k} - Ls_k\frac{1}{L\alpha_k^+}\right)^2\right) \\ &= \frac{L}{2(1 - Ls_k)^2(L\alpha_k^+)^2} \quad 2L\alpha_k^+(1 - Ls_k)^2 - (L\alpha_k^+)^2 + 2Ls_k(1 - Ls_k)L\alpha_k^+ - L^2 s_k^2(1 - Ls_k)^2\big) \\ &= \frac{L}{2(1 - Ls_k)^2(L\alpha_k^+)^2} \quad -(L\alpha_k^+)^2 + 2(1 - Ls_k)L\alpha_k^+ - L^2 s_k^2(1 - Ls_k)^2\big)\,. \end{aligned} \tag{26}$$

We now use the expression of α_k^+ given in (23) and expand the expression in (26) by using $L\alpha_k^+ = (1-Ls_k)(2-L^2s_k^2)$.

$$
\begin{aligned}
& LQ\left(\frac{1}{L\alpha_k^+}\right) \\
&= \frac{L}{2(1-Ls_k)^4(2-L^2s_k^2)^2} \quad -(1-Ls_k)^2(2-L^2s_k^2)^2 + 2(1-Ls_k)^2(2-L^2s_k^2) - L^2s_k^2(1-Ls_k)^2\big) \\
&= \frac{L}{2(1-Ls_k)^2(2-L^2s_k^2)^2}\left(-(2-L^2s_k^2)^2 + 2(2-L^2s_k^2) - L^2s_k^2\right) \\
&= \frac{L}{2(1-Ls_k)^2(2-L^2s_k^2)^2}\left(-4 + 4L^2s_k^2 - L^4s_k^4 + 4 - 2L^2s_k^2 - L^2s_k^2\right) \\
&= \frac{L}{2(1-Ls_k)^2(2-L^2s_k^2)^2}\left(L^2s_k^2 - L^4s_k^4\right) \\
&= \frac{L^3s_k^2}{2(1-Ls_k)^2(2-L^2s_k^2)^2} \quad 1 - L^2s_k^2\big) \\
&= \frac{L^3s_k^2}{2(1-Ls_k)(2-L^2s_k^2)^2}\left(1+Ls_k\right). \qquad (27)
\end{aligned}
$$

Combining (25) and (27), we obtain

$$
\begin{aligned}
\frac{1}{\alpha_k} - \frac{L}{2}\left(\frac{1}{1-Ls_k} - \frac{s_k}{\alpha_k}\right)^2 &\geq LQ\left(\frac{1}{L\alpha_k^+}\right) \\
&= \frac{L^3s_k^2(1+Ls_k)}{2(2-L^2s_k^2)^2(1-Ls_k)} \\
&\geq \frac{L^3s_-^2(1+Ls_-)}{2(2-L^2s_-^2)^2(1-Ls_-)} = C,
\end{aligned}
$$

which is the desired result. □

We can check that, when condition (**C3**) is satisfied, one has

$$
0 < b_k = \frac{L\alpha_k + (1-Ls_k)^2}{\alpha_k(1-Ls_k)} \leq \frac{2 - 2Ls_k + (1-Ls_k)^2}{\alpha_k(1-Ls_k)} = \frac{3-Ls_k}{\alpha_k} \leq B = \frac{3}{\alpha_-}.
$$

Combining this with Proposition 8, 15 and Lemma 16, we obtain the following corollary.

Corollary 18 *Suppose that $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition (**C3**) and that f is convex, then*

(i) $F(x_{k+1}) + C\|x_k - x_{k+1}\|^2 \leq F(x_k)$, $\forall k \in \mathbb{N}$.

(ii) *There exists $\omega_{k+1} \in \partial F(x_{k+1})$ such that*

$$
\|\omega_{k+1}\| \leq B\|x_k - x_{k+1}\|, \forall k \in \mathbb{N}.
$$

where C is given in Lemma 16 and $B = \frac{3}{\alpha_-}$.

We now consider the complexity for EEG method under the nonsmooth KL inequality in the form of a *small prox* result as in [12]. First, we recall some definitions from [12]. Let $0 < r_0 := F(x_0) < \bar{r}$, we assume that F has the KL property on $[0 < F < \bar{r}]$ with desingularizing function $\varphi \in \mathcal{K}(\bar{r})$.

Set $\beta_0 := \varphi(r_0)$ and consider the function $\psi := (\varphi|_{[0,r_0]})^{-1} : [0, \beta_0] \to [0, r_0]$, which is increasing and convex. We add the assumption that ψ' is Lipschitz continuous (on $[0, \beta_0]$) with constant $\ell > 0$ and $\psi'(0) = 0$.

Set

$$\zeta := \frac{\sqrt{1 + 2\ell\, C\, B^{-2}} - 1}{\ell}.$$

Starting from β_0, we define the sequence $(\beta_k)_{k\in\mathbb{N}}$ by

$$\begin{aligned}\beta_{k+1} &:= \operatorname{argmin}\left\{\psi(u) + \frac{1}{2\zeta}(u - \beta_k)^2 : u \geq 0\right\}\\ &= \operatorname{prox}_{\zeta\psi}(\beta_k).\end{aligned}$$

It is easy to prove that β_k is decreasing and converges to zero. By continuity, $\lim\limits_{k\to\infty} \psi(\beta_k) = 0$.

Now, applying the result of [12, Theorem 17], we have the complexity of EEG method in the form of a *small prox* result.

Theorem 19 (Complexity of EEG method) *Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by EEG method. Assume that f is convex and $(s_k, \alpha_k)_{k\in\mathbb{N}}$ satisfy condition (**C3**). Then, the sequence $(x_k)_{k\in\mathbb{N}}$ converges to $x^* \in \operatorname{argmin} F$, and*

$$\sum_{i=1}^{\infty} \|x_k - x_{k+1}\| < \infty,$$

moreover,

$$\begin{aligned}F(x_k) - F^* &\leq \psi(\beta_k), \quad \forall k \geq 0,\\ \|x_k - x^*\| &\leq \frac{B}{C}\beta_k + \sqrt{\frac{\psi(\beta_{k-1})}{C}}, \quad \forall k \geq 1,\end{aligned}$$

where B and C are given in Corollary 18.

4 Numerical Experiment

In this section, we compare the EEG method with standard algorithms in numerical optimization: Forward-Backward and FISTA. We describe the problem of interest, details about exact line search in this context and numerical results.

4.1 ℓ_1 Regularized Least Squares

We let $A \in \mathbb{R}^{p\times n}$ be a real matrix, $b \in \mathbb{R}^n$ be a real vector and $\lambda > 0$ be a scalar, all of them given and fixed. Following the notations of the previous section, we define $f\colon x \mapsto \frac{1}{2}\|Ax - b\|_2^2$ and $g\colon x \mapsto \lambda\|x\|_1$ (the sum of absolute values of the entries). With these notations, the optimization problem (**P**) becomes

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1. \tag{28}$$

Solutions of problem of the form of (28) (as well as many extensions) are extensively used in statistics and signal processing [2,21]. For this problem, we introduce the proximal gradient mapping, a specialization of the proximal gradient step to problem (28). This is the main building block of all the algorithms presented in the numerical experiment.

$$\begin{array}{llll} p\colon & \mathbb{R}^n \times \mathbb{R}_+ & \mapsto & \mathbb{R}^n \\ & (x,s) & \mapsto & S_{s\lambda}(x - s\nabla f(x)). \end{array} \tag{29}$$

where S_a $(a \in \mathbb{R}_+)$ is the soft-thresholding operator which acts coordinatewise and satisfies for $i = 1, 2 \ldots, n$

$$[S_a(x)]_i = \begin{cases} 0, & \text{if } |x_i| \leq a. \\ x_i - a\text{sign}(x_i), & \text{otherwise.} \end{cases}$$

4.2 Exact Line Search

One intuition behind Extragradient-Method for optimization is the use of an additional iteration as a guide or a scout to provide an estimate of the gradient that better suits the geometry of the problem. This should eventually translate to taking larger steps leading to faster convergence. In order to evaluate Extragradient-Method, we need a mechanism which would allow us to take larger steps when this is beneficial. One such mechanism is exact line search. This strategy is not widely used because of its computational overhead. In this section, we briefly describe a strategy which allows to perform exact line search efficiently in the context of ℓ_1-regularized least squares. As far as we know, this approach has not been described in the literature. Furthermore, this strategy may be extended to more general least squares problems with nonsmooth regularizers. For the rest of this section, we assume that $x \in \mathbb{R}^n$ is fixed. We heavily rely on the two simple facts:

- The mapping $s \to p(x,s)$ is continuous and piecewise affine.
- The objective function $x \mapsto f(x) + g(x)$ is continuous and piecewise quadratic.

We consider the following function

$$\begin{aligned} q_x \colon & \mathbb{R}_+ \to \mathbb{R} \\ & \qquad \alpha \to f(p(x,\alpha)) + g(p(x,\alpha)). \end{aligned}$$

It can be deduced from the properties of f, g and p that q_x is continuous and piecewise quadratic. In classical implementation of proximal splitting methods, the step-size parameter α is a well chosen constant which depends on the problem, or alternatively it is estimated using backtracking. The alternative which we propose is to choose the step-size parameter α minimizing q_x. Since q_x is a one dimensional piecewise quadratic function, then we only need to know its expression between the values of α which constitute breakpoints where the quadratic expression of the function q_x changes, *i.e.* points where q_x is not differentiable.

The nonsmooth points of q_x are given by the following set

$$\mathcal{D}_x = \left\{ \frac{x_i}{\frac{\partial f(x)}{\partial x_i} - \lambda}, \frac{x_i}{\frac{\partial f(x)}{\partial x_i} + \lambda} \right\}_{i=1}^{n} \cap \mathbb{R}_+$$

and correspond to limiting values for which coordinates of $p(x,\alpha)$ are null. We assume that the elements of $\mathcal{D}_x$ are ordered nondecrasingly (letting potential ties appear several times). The comments that we have made so far lead to the following.

- $\mathcal{D}_x$ contains no more than $2n$ elements.
- Given x and λ, computing $\mathcal{D}_x$ is as costly as computing ∇f.
- q_x is quadratic between two consecutive elements of $\mathcal{D}_x$.

In order to minimize q_x, the only task that should be performed is to keep track of its value (or equivalently of its quadratic expression) between consecutive elements of $\mathcal{D}_x$. Here, we can use the fact that elements of $\mathcal{D}_x$ corresponds to values of α for which one coordinate of $p(x, \alpha)$ goes to zero or becomes active (non-zero). A careful implementation of the minimization of q_x amounts to sort the values in $\mathcal{D}_x$, placing them in increasing order, keeping track of the corresponding quadratic expression and the minimal value. We provide a few details for completeness.

- The vector $d_x(s) := \left(\frac{\partial [p(x,s)]_i}{\partial s}\right)_{i=1}^n \in \mathbb{R}^n$ is constant between consecutive elements of $\mathcal{D}_x$. Furthermore the elements of $\mathcal{D}_x$ (counted with multiple ties) corresponds to value of α for which a single coordinate of $d_x(s)$ is modified.
- Suppose that $\alpha_1 < \alpha_2$ are two consecutive elements of $\mathcal{D}_x$. Then for all $\alpha \in [\alpha_1, \alpha_2]$, letting $d_x(\alpha) = d$ on this segment, we have $p(x, \alpha) = p(x, \alpha_1) + (\alpha - \alpha_1)d$, hence,

$$\begin{aligned} &\frac{1}{2}\|Ap(x,\alpha) - b\|_2^2 + \lambda\|p(x,\alpha)\|_1 \\ =\ &\frac{1}{2}\|Ap(x,\alpha_1) - b\|_2^2 + \lambda\|p(x,\alpha_1)\|_1 \\ &+ \frac{(\alpha - \alpha_1)}{n}\langle Ad, Ax - b\rangle + \frac{(\alpha - \alpha_1)^2}{2n}\|Ad\|_2^2 + \lambda(\alpha - \alpha_1)\left\langle \bar{d}, d\right\rangle, \end{aligned}$$

  where $\bar{d} \in \mathbb{R}^p$ is a vector which depends on the sign pattern of $p(x, \alpha_1)$ and d.
- For $\alpha = \alpha_2$, the sign pattern of $p(x, \alpha_2)$ and the corresponding value of d and $\bar{d}$ (for the next interval) are modified only at a single coordinate, the same for the three of them. In other words, updating the quadratic expression of q_x at α_2 only requires the knowledge of this coordinate, the value of the corresponding column in A and can be done by computing inner products in $\mathbb{R}^p$. This requires $O(p)$ operations.
- Given these properties, we can perform minimization of q_x by an active set strategy, keeping track only of the sign pattern of $p(x, \alpha)$, the value of $\left\langle \bar{d}, d\right\rangle$, the value of Ad, $Ap(x, \alpha) - b$ and $\|p(x, \alpha)\|_1$ which cost is of the order of $O(p)$. This should not be repeated more than $2n$ times.

Using this active set procedure provides the quadratic expression of q_x for all intervals represented by consecutive values in $\mathcal{D}_x$. From these expressions, it is not difficult to compute the global minimum of q_x. The overall cost of this operation is of the order of $O(np)$ plus the cost of sorting $2n$ elements in $\mathbb{R}$. This is comparable to the cost of computing the gradient of f. Hence in this specific setting, performing exact line search does not add much overhead in term of computational cost compared to existing step-size strategies.

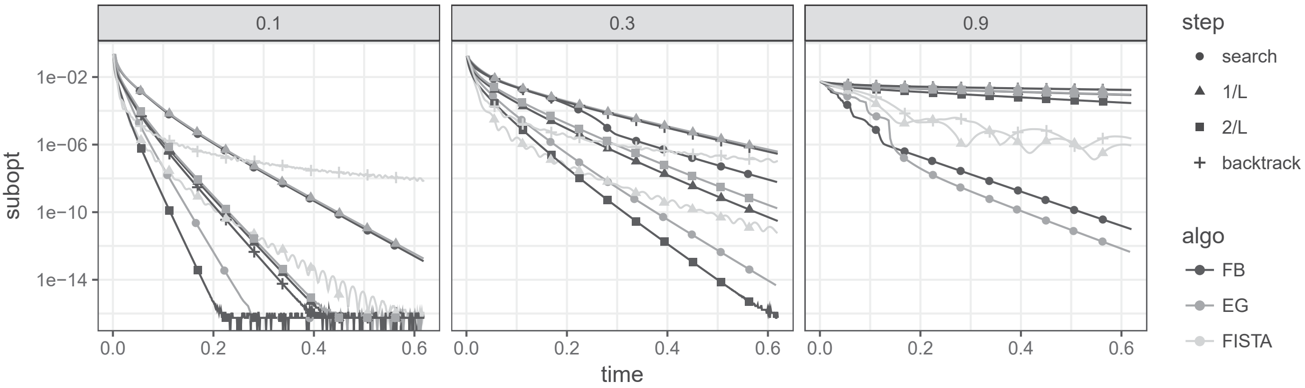

Figure 1: Suboptimality $(F(x_k) - F^*)$ as a function of time for simulated ℓ_1 regularized least squares data. FB stands for Forward-Backward and EG for Extra-Gradient. The color is related to the algorithm used and the dots are related to the step size used. Different windows show different values of the parameter δ (see the main text for a precise description). On the left we have a well conditioned problem and on the right the conditioning is much worse. For $\delta = 0.1$, 0.3, 0.9, the condition number of the matrix A are approximately 6, 15 and 300 respectively.

4.3 Simulation and Results

We generate a matrix A and vector b using the following process.

- Set $n = 600$ and $p = 300$.
- Set $A = DX$ where X has standard Gaussian independent entries and D is a diagonal matrix which i-th diagonal entry is $\frac{1}{i^\delta}$ where δ is a positive parameter controlling the good conditioning of the matrix A (the smaller δ, the better).
- Choose b with independant Gaussian entries.
- We set $\lambda = 1/n \simeq 0.001$.

We compare the forward-backward splitting algorithm, FISTA [3] and the proposed extragradient method with different step size rules (L is the Lipschitz constant of f computed from the singular values of A).

- A step of size $1/L$.
- A step of size $2/L$.
- A step given by backtracking line search (see e.g. [3]). The original guess for L is chosen to be 1 and the multiplicative parameter is 1.2.
- A step given by exact line search as described in the previous section.

For the extragradient method, we always choose $s = 1/L$ and determine α by the chosen step-size rule. For FISTA algorithm, we do not implement the $2/L$ and exact line search step size rules as they produce diverging sequences. The exact line search active set procedure is implemented in compiled `C` code in order keep a reasonable level of efficiency compared to linear algebra operations which have efficient implementations. The algorithms are initialized at the origin. We keep track of decrease of the objective value, the iteration counter k and the total time spent since initialization. The iteration counter is related to analytical complexity while the total time spent is related to the arithmetical complexity (see the introduction in [19] for more details). Comparing algorithms in term of analytical complexity does not reflect the fact that iterations are more costly for some of them compared to others so we only focus on arithmetical complexity which in our case is roughly proportional to computational time.

Computational times for a generic LASSO problem are presented in Figure 1 for $\delta = 0.1$, 0.3, 0.9. The main comments are as follows:

- For well conditioned problems, the forward-backward algorithm with step size $2/L$ performs the best. This is not the case for the less well conditioned problem where exact line search method shows some advantage.
- The extragradient method with exact line search performs reasonably well, independently of the conditioning of the problem.
- FISTA algorithm is outperformed by other methods in terms of asymptotic convergence. Furthermore, FISTA's performance is very sensitive to step-size tuning.

This experiment illustrates that exact line search can improve performances for ill-conditioned problems and that the proposed extragradient method is able to take advantage of it, independently of the problem's conditioning. This observation is based on a "generic" instance of the LASSO problem. Further experiments on real data are required to confirm generality of the observation. This is a matter of future research.

5 Conclusions

In this paper, we presented an extension of extragradient method, EEG, and used it to tackle the problem of minimizing the sum of two functions. Under step size conditions, we showed that EEG is a first order descent method. By using the KL inequality, we obtained the convergence of the sequence produced by EEG method and estimated the complexity of EEG method via the small-prox method. In the convex setting, we obtained a classical sublinear convergence rate for the objective function value. Finally, we described an exact line search strategy for the ℓ_1 regularized least square problem and conducted numerical comparisons with existing algorithms on a generic instance of the LASSO problem.

Acknowledgement of Support and Disclaimer This work is sponsored by a grant from the Air Force Office of Scientific Research, Air Force Material Command (grant number FA9550-15-1-0500). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force Research Laboratory. The collaboration with Emile Richard mostly took place during his postdoctoral stay in Mines ParisTech, Paris, France in 2013.

The authors would like to thank Professor Jérôme Bolte for his suggestions, the associate editor and anonymous referee for helpful remarks which helped improve the quality of this manuscript.

References

[1] Patrick Louis Combettes and Jean Christophe Pesquet. Proximal splitting methods in signal processing. In Heinz H Bauschke, Regina Burachik, Patrick Louis Combettes, Veit Elser, D Russell Luke,

and Henry Wolkowicz, editors, *Fixed-point algorithms for inverse problems in science and engineering*, volume 49, pages 185–212. Springer, 2011.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[4] Patrick Louis Combettes and Valérie Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[5] GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[6] Yair Censor, Aviv Gibali, and Simeon Reich. The subgradient extragradient method for solving variational inequalities in Hilbert space. *Journal of Optimization Theory and Applications*, 148(2):318–335, 2011.

[7] Renato Monteiro and Benar Svaiter. Complexity of variants of Tseng's modified forward–backward splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

[8] Zhi Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[9] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.

[10] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[11] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[12] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[13] Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[14] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[15] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[16] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[17] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton University Press, 1972.

[18] Heinz H Bauschke and Patrick Louis Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer Science & Business Media, 2011.

[19] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[20] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[21] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.