An Accelerated Hybrid Proximal Extragradient Method for Convex Optimization and its Implications to Second-Order Methods

Renato D.C. Monteiro * B. F. Svaiter†

May 10, 2011 (Revised: May 24, 2012)

Abstract

This paper presents an accelerated variant of the hybrid proximal extragradient (HPE) method for convex optimization, referred to as the accelerated HPE (A-HPE) framework. Iteration-complexity results are established for the A-HPE framework, as well as a special version of it, where a large stepsize condition is imposed. Two specific implementations of the A-HPE framework are described in the context of a structured convex optimization problem whose objective function consists of the sum of a smooth convex function and an extended real-valued non-smooth convex function. In the first implementation, a generalization of a variant of Nesterov's method is obtained for the case where the smooth component of the objective function has Lipschitz continuous gradient. In the second implementation, an accelerated Newton proximal extragradient (A-NPE) method is obtained for the case where the smooth component of the objective function has Lipschitz continuous Hessian. It is shown that the A-NPE method has a $\mathcal{O}(1/k^{7/2})$ convergence rate, which improves upon the $\mathcal{O}(1/k^3)$ convergence rate bound for another accelerated Newton-type method presented by Nesterov. Finally, while Nesterov's method is based on exact solutions of subproblems with cubic regularization terms, the A-NPE method is based on inexact solutions of subproblems with quadratic regularization terms, and hence is potentially more tractable from a computational point of view.

Key words: complexity, extragradient, variational inequality, maximal monotone operator, proximal point, ergodic convergence, hybrid, convex programming, accelerated gradient, accelerated Newton

1 Introduction

A broad class of optimization, saddle point, equilibrium and variational inequality (VI) problems can be posed as the *monotone inclusion problem* (MI), namely: finding x such that $0 \in T(x)$, where T is a maximal monotone point-to-set operator defined in a real Hilbert space. The proximal point

*School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (email: monteiro@isye.gatech.edu). The work of this author was partially supported by NSF Grants CCF-0808863 and CMMI-0900094 and ONR Grant ONR N00014-11-1-0062.

†IMPA, Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, Brazil (benar@impa.br). The work of this author was partially supported by CNPq grants 303583/2008-8, 302962/2011-5, 480101/2008-6, 474944/2010-7, FAPERJ grants E-26/102.821/2008, E-26/102.940/2011 and by PRONEX-Optimization.

method, proposed by Martinet [6], and further studied by Rockafellar [16, 15], is a classical iterative scheme for solving the MI problem which generates a sequence $\{x_k\}$ according to

$$x_k = (\lambda_k T + I)^{-1}(x_{k-1}),$$

It has been used as a generic framework for the design and analysis of several implementable algorithms. The classical inexact version of the proximal point method allows for the presence of a sequence of summable errors in the above iteration, i.e.:

$$\|x_k - (\lambda_k T + I)^{-1}(x_{k-1})\| \leq e_k, \qquad \sum_{k=1}^{\infty} e_k < \infty,$$

where $\|\cdot\|$ is the canonical norm of the Hilbert space. Convergence results under the above error condition have been establish in [16] and have been used in the convergence analysis of other methods that can be recast in the above framework [15].

New inexact versions of the proximal point method with relative error tolerance were proposed by Solodov and Svaiter [18, 19, 20, 21]. Iteration complexity results for one of these inexact versions of the proximal point method introduced in [18, 19], namely the hybrid proximal extragradient (HPE) method, were established in [10]. Application of this framework in the iteration-complexity analysis of several zero-order (or, in the context of optimization, first-order) methods for solving monotone variational inequalities, and monotone inclusion and saddle-point problems, are discussed in [10] and in the subsequent papers [8, 9].

The HPE framework was also used to study the iteration-complexity of a first-order (or, in the context of optimization, second-order) method for solving monotone nonlinear equations (see [10]) and, more generally, for monotone smooth variational inequalities and inclusion problems consisting of the sum of a smooth monotone map and a maximal monotone point-to-set operator (see [11]).

Iteration-complexity bounds for accelerated inexact versions of the proximal point method for convex optimization have been obtained in [4, 17] under suitable *absolute* error asymptotic conditions. The purpose of this paper is to present an accelerated variant of the HPE method for convex optimization (based on a *relative* error condition), which we refer to as the accelerated HPE (A-HPE) framework. This framework builds on the ideas introduced in [10, 18, 19, 12]. Iteration-complexity results are established for the A-HPE method, as well as a special version of it, where a large stepsize condition is imposed. We then give two specific implementations of the A-HPE framework in the context of a structured convex optimization problem whose objective function consists of the sum of a smooth convex function and an extended real-valued non-smooth convex function. In the first implementation, we obtain a generalization of a variant of Nesterov's method for the case where the smooth component of the objective function has Lipschitz continuous gradient. In the second implementation, we obtain a accelerated Newton proximal extragradient (A-NPE) method for the case where the smooth component of the objective function has Lipschitz continuous Hessian. We show that the A-NPE method has a $\mathcal{O}(1/k^{7/2})$ convergence rate, which improves upon the $\mathcal{O}(1/k^3)$ convergence rate bound for another accelerated Newton-type method presented in Nesterov [13]. As opposed to the method in the latter paper, which is based on exact solutions of subproblems with cubic regularization terms, the A-NPE method is based on inexact solutions of subproblems with quadratic regularization terms, and hence is potentially more tractable from a computational point of view. In addition, the method in [13] is described only in the context of unconstrained convex optimization, while A-HPE framework applies to constrained, as well as more general, convex optimization problems.

This paper is organized as follows. Section 2 introduces some basic definitions and facts about convex functions and ε-enlargement of maximal monotone operators. Section 3 describes in the context of convex optimization an accelerated variant of the HPE method introduced in [18, 19] and studies its computational complexity. Section 4 analyzes the convergence rate of a special version of the A-HPE framework, namely the large-step A-HPE framework, which imposes a large-step condition on the sequence of stepsizes. Section 5 describes a first-order implementation of the A-HPE framework which leads to a generalization of a variant of Nesterov's method. Section 6 describes a second-order implementation of the large-step A-HPE framework, namely the A-NPE method. Section 7 describes a line-search procedure which is used to compute the stepsize at each iteration of the A-NPE method.

2 Basic definition and notation

In this section, we review some basic definitions and facts about convex functions and ε-enlargement of monotone multi-valued maps.

Throughout this paper, $\mathbb{E}$ will denote a finite dimensional inner product real vector space with inner product and induced norm denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. For a nonempty closed convex set $\Omega \subseteq \mathbb{E}$, we denote the orthogonal projection operator onto Ω by P_Ω. We denote the set of real numbers by $\mathbb{R}$ and the set of extended reals, namely $\mathbb{R}\cup\{\pm\infty\}$ by $\bar{\mathbb{R}} = \mathbb{R}\cup\{\pm\infty\}$. We let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote for the set of nonnegative and positive real numbers, respectively. The domain of definition of a point-to-point function F is denoted by $\operatorname{Dom} F$.

A point-to-set operator $T : \mathbb{E} \rightrightarrows \mathbb{E}$ is a relation $T \subseteq \mathbb{E}\times\mathbb{E}$. Alternatively, one can consider T as a multi-valued function of $\mathbb{E}$ into the family $\wp(\mathbb{E}) = 2^{(\mathbb{E})}$ of subsets of $\mathbb{E}$, namely:

$$T(z) = \{v \in \mathbb{E} \mid (z,v) \in T\}, \qquad \forall z \in \mathbb{E}.$$

Regardless of the approach, it is usual to identify T with its graph defined as

$$Gr(T) = \{(z,v) \in \mathbb{E}\times\mathbb{E} \mid v \in T(z)\}.$$

The domain of T, denoted by $\operatorname{Dom} T$, is defined as

$$\operatorname{Dom} T := \{z \in \mathbb{E} : T(z) \neq \emptyset\}.$$

The operator $T : \mathbb{E} \rightrightarrows \mathbb{E}$ is *monotone* if

$$\langle v - \tilde{v}, z - \tilde{z}\rangle \geq 0, \qquad \forall (z,v),(\tilde{z},\tilde{v}) \in Gr(T),$$

and T is *maximal monotone* if it is monotone and maximal in the family of monotone operators with respect to the partial order of inclusion, i.e., $S : \mathbb{E} \rightrightarrows \mathbb{E}$ monotone and $Gr(S) \supset Gr(T)$ implies that $S = T$.

In [2], Burachik, Iusem and Svaiter introduced the ε-enlargement of maximal monotone operators. In [10] this concept was extended to a generic point-to-set operator in $\mathbb{E}$ as follows. Given $T : \mathbb{E} \rightrightarrows \mathbb{E}$ and a scalar ε, define $T^\varepsilon : \mathbb{E} \rightrightarrows \mathbb{E}$ as

$$T^\varepsilon(z) = \{v \in \mathbb{E} \mid \langle z - \tilde{z}, v - \tilde{v}\rangle \geq -\varepsilon, \quad \forall \tilde{z} \in \mathbb{E},\ \forall \tilde{v} \in T(\tilde{z})\}, \quad \forall z \in \mathbb{E}. \tag{1}$$

We now state a few useful properties of the operator T^ε that will be needed in our presentation.

Proposition 2.1. *Let $T, T' : \mathbb{E} \rightrightarrows \mathbb{E}$. Then,*

a) if $\varepsilon_1 \leq \varepsilon_2$, then $T^{\varepsilon_1}(z) \subseteq T^{\varepsilon_2}(z)$ for every $z \in \mathbb{E}$;

b) $T^{\varepsilon}(z) + (T')^{\varepsilon'}(z) \subseteq (T + T')^{\varepsilon+\varepsilon'}(z)$ for every $z \in \mathbb{E}$ and $\varepsilon, \varepsilon' \in \mathbb{R}$;

c) T is monotone if, and only if, $T \subseteq T^0$.

Proof. Statements a) and b) follow immediately from definition (1) and statement c) is proved in [7, Proposition 21]. □

Proposition 2.2 ([3, Corrolary 3.8])**.** *Let $T : \mathbb{E} \rightrightarrows \mathbb{E}$ be a maximal monotone operator. Then,* $\mathrm{Dom}\,(T^{\varepsilon}) \subseteq \overline{\mathrm{Dom}\,(T)}$ *for any $\varepsilon \geq 0$.*

For a scalar $\varepsilon \geq 0$, the ε-subdifferential of a proper closed convex function $f : \mathbb{E} \to \bar{\mathbb{R}}$ is the operator $\partial_{\varepsilon} f : \mathbb{E} \rightrightarrows \mathbb{E}$ defined as

$$\partial_{\varepsilon} f(x) = \{v \mid f(y) \geq f(x) + \langle y - x, v\rangle - \varepsilon,\ \forall y \in \mathbb{E}\}, \quad \forall x \in \mathbb{E}. \tag{2}$$

When $\varepsilon = 0$, the operator $\partial_{\varepsilon} f$ is simply denoted by ∂f and is referred to as the subdifferential of f. The operator ∂f is trivially monotone if f is proper. If f is a proper lower semi-continuous convex function, then ∂f is maximal monotone [14]. The next proposition states some useful properties of the ε-subdifferential.

Proposition 2.3. *Let $f : \mathbb{E} \to \bar{\mathbb{R}}$ be a proper convex function. Then,*

a) $\partial_{\varepsilon} f(x) \subseteq (\partial f)^{\varepsilon}(x)$ for any $\varepsilon \geq 0$ and $x \in \mathbb{E}$, where $(\partial f)^{\varepsilon}$ stands for the ε-enlargement of ∂f;

b) if $v \in \partial f(x)$ and $f(y) < \infty$, then $v \in \partial_{\varepsilon} f(y)$, where $\varepsilon := f(y) - [f(x) + \langle y - x, v\rangle]$.

Proof. Statement a) is proved in [2, Proposition 3] and b) is a classical result which can be found, for example, in Proposition 4.2.2 of [5]. □

Let $X \subseteq \mathbb{E}$ be a non-empty closed convex set. The *indicator function* of X is the function $\delta_X : \mathbb{E} \to \bar{\mathbb{R}}$ defined as

$$\delta_X(x) = \begin{cases} 0, & x \in X, \\ \infty, & \text{otherwise}, \end{cases}$$

and the *normal cone operator* of X is the point-to-set map $N_X : \mathbb{E} \rightrightarrows \mathbb{E}$ given by

$$N_X(x) = \begin{cases} \emptyset, & x \notin X, \\ \{v \in \mathbb{E},\ \mid \langle y - x, v\rangle \leq 0,\ \forall y \in X\}, & x \in X. \end{cases} \tag{3}$$

Clearly, the normal cone operator N_X of X can be expressed in terms of δ_X as $N_X = \partial \delta_X$.

3 An accelerated hybrid proximal extragradient framework

In this section, we describe in the context of convex optimization an accelerated variant of the hybrid proximal extragradient method introduced in [18, 19] and study its computational complexity. This variant uses ideas similar to the ones used in Nesterov's optimal method but generalizes the later method in a significant way.

Our problem of interest is the convex optimization problem

$$f_* := \inf\{f(x) : x \in \mathbb{E}\}, \tag{4}$$

where $f : \mathbb{E} \to \mathbb{R} \cup \{\infty\}$ is a proper closed convex function. We assume throughout the paper that $f_* \in \mathbb{R}$ and the set of optimal solutions X_* of (4) is nonempty.

The accelerated hybrid proximal extragradient framework studied in this section is as follows.

A-HPE Framework

0) Let $x_0, y_0 \in \mathbb{E}$ and $0 \le \sigma \le 1$ be given, and set $A_0 = 0$ and $k = 0$.

1) Compute $\lambda_{k+1} > 0$ and a triple $(\tilde{y}_{k+1}, v_{k+1}, \varepsilon_{k+1}) \in \mathbb{E} \times \mathbb{E} \times \mathbb{R}_+$ such that

$$v_{k+1} \in \partial_{\varepsilon_{k+1}} f(\tilde{y}_{k+1}), \qquad \|\lambda_{k+1} v_{k+1} + \tilde{y}_{k+1} - \tilde{x}_k\|^2 + 2\lambda_{k+1}\varepsilon_{k+1} \le \sigma^2 \|\tilde{y}_{k+1} - \tilde{x}_k\|^2, \tag{5}$$

where

$$\tilde{x}_k = \frac{A_k}{A_k + a_{k+1}}\, y_k + \frac{a_{k+1}}{A_k + a_{k+1}}\, x_k, \tag{6}$$

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1} A_k}}{2}. \tag{7}$$

2) Find y_{k+1} such that $f(y_{k+1}) \le f(\tilde{y}_{k+1})$ and let

$$A_{k+1} = A_k + a_{k+1}, \tag{8}$$

$$x_{k+1} = x_k - a_{k+1} v_{k+1}. \tag{9}$$

3) Set $k \leftarrow k + 1$, and go to step 1.

end

We now make several remarks about the A-HPE Framework. First, the framework obtained by replacing (8) by the equation $A_{k+1} = 0$, which as consequence imply that $\tilde{x}_j = x_j$, $a_j = \lambda_j$ and $x_{j+1} = x_j - \lambda_{j+1} v_{j+1}$ for all $j \in \mathbb{N}$, is exactly the HPE method proposed by Solodov and Svaiter [18]. Second, the A-HPE Framework does not specify how to compute λ_{k+1} and $(\tilde{y}_{k+1}, v_{k+1}, \varepsilon_{k+1})$ as in step 1. Specific computation of these quantities will depend on the particular implementation of the framework and properties of function f. In Sections 5 and 6, we will describe procedures for finding these quantities in the context of two specific implementations of the A-HPE framework, namely: a first-order method which is a variant of Nesterov's algorithm, and a second-order accelerated method. Third, for an arbitrary λ_{k+1} and $\tilde{x}_k$ as in (6)-(7), the exact proximal point iterate $\tilde{y}$ and the vector

v defined as

$$\tilde{y} := \arg\min_{x\in\mathbb{E}} \left(\lambda_{k+1} f(x) + \frac{1}{2}\|x - \tilde{x}_k\|^2 \right), \qquad v := \frac{1}{\lambda_{k+1}}(\tilde{x}_k - \tilde{y}),$$

are characterized by

$$v \in \partial f(\tilde{y}), \qquad \lambda_{k+1} v + \tilde{y} - \tilde{x}_k = 0, \tag{10}$$

and hence $\varepsilon_{k+1} := 0$, $\tilde{y}_{k+1} := \tilde{y}$ and $v_{k+1} := v$ satisfy the error tolerance (5) with $\sigma = 0$. Therefore, the error criterion (5) is a relaxation of the characterization (10) of the proximal point iterate in that the inclusion in (10) is relaxed to $v \in \partial_\varepsilon f(\tilde{y})$ and the equation in (10) is allowed to have a residual $r = \lambda_{k+1} v + \tilde{y} - \tilde{x}_k$ such that the residual pair (r, ε) is small relative to $\|\tilde{y} - \tilde{x}_k\|$ in that $\|r\|^2 + 2\lambda_{k+1}\varepsilon \le \sigma^2 \|\tilde{y} - \tilde{x}_k\|^2$. Fourth, the error tolerance (5) is the optimization version of the HPE relative error tolerance introduced in [18] in that an inclusion in terms of the ε-enlargement of a maximal monotone operator is replaced by an inclusion in terms of the ε-subdifferential of a proper closed convex function. Fifth, there are two readily available rules for choosing y_{k+1} in step 2) of the A-HPE framework, namely:

- either set $y_{k+1} = \tilde{y}_{k+1}$;
- or, $y_{k+1} = \operatorname{argmin}\{f(y) : y \in \{y_k, \tilde{y}_{k+1}\}\}$.

The advantage of the latter rule is that it forces the sequence $\{f(y_k)\}$ to be non-increasing.

For the sake of future reference, we state the following trivial result.

Lemma 3.1. *For every integer $k \ge 0$, we have $\lambda_{k+1} A_{k+1} = a_{k+1}^2 > 0$.*

Proof. Clearly, $a_{k+1} > 0$ satisfies (7) if, and only if, $a = a_{k+1} > 0$ satisfies

$$a^2 - \lambda_{k+1} a - \lambda_{k+1} A_k = 0,$$

or equivalently, $a_{k+1}^2 = \lambda_{k+1}(A_k + a_{k+1}) = \lambda_{k+1} A_{k+1}$, where the last equality is due to (8). □

In order to analyze the properties of the sequences $\{x_k\}$ and $\{y_k\}$, define the affine maps $\gamma_k : \mathbb{E} \to \mathbb{R}$ as

$$\gamma_k(x) = f(\tilde{y}_k) + \langle x - \tilde{y}_k, v_k \rangle - \varepsilon_k, \qquad \forall x \in \mathbb{E},\ k \ge 1. \tag{11}$$

and the aggregate affine maps $\Gamma_k : \mathbb{E} \to \mathbb{R}$ recursively as

$$\Gamma_0 \equiv 0, \qquad \Gamma_{k+1} = \frac{A_k}{A_{k+1}} \Gamma_k + \frac{a_{k+1}}{A_{k+1}} \gamma_{k+1}, \quad \forall k \ge 0. \tag{12}$$

Lemma 3.2. *For every integer $k \ge 0$, there hold:*

a) γ_{k+1} is affine and $\gamma_{k+1} \le f$;

b) Γ_k is affine and $A_k \Gamma_k \le A_k f$;

c) $x_k = \arg\min_{x\in\mathbb{E}} \left(A_k \Gamma_k(x) + \frac{1}{2}\|x - x_0\|^2 \right)$.

Proof. Statement a) follows from (11), the inclusion in (5), and definition (2). Statement b) follows from a), (12) and a simple induction argument. To prove c), first note that (12) imply by induction that

$$A_k \nabla \Gamma_k(x) = \sum_{j=1}^{k} a_j v_j, \qquad \forall x \in \mathbb{E}.$$

Moreover, (9) imply that $x_k = x_0 - \sum_{j=1}^{k} a_j v_j$. The last two conclusions then imply that $A_k \nabla \Gamma_k(x_k) + x_k - x_0 = 0$, and hence that x_k satisfies the optimality condition for the minimization problem in c). Hence, c) follows. □

The following elementary result will be used in the analysis of the A-HPE framework.

Lemma 3.3. *Let vectors $\tilde{x}, \tilde{y}, \tilde{v} \in \mathbb{E}$ and scalars $\lambda > 0$, $\varepsilon, \sigma \geq 0$ be given. Then, the inequality*

$$\|\lambda \tilde{v} + \tilde{y} - \tilde{x}\|^2 + 2\lambda\varepsilon \leq \sigma^2 \|\tilde{y} - \tilde{x}\|^2$$

is equivalent to the inequality

$$\min_{x \in \mathbb{E}} \left\{ \langle \tilde{v}, x - \tilde{y} \rangle - \varepsilon + \frac{1}{2\lambda} \|x - \tilde{x}\|^2 \right\} \geq \frac{1 - \sigma^2}{2\lambda} \|\tilde{y} - \tilde{x}\|^2.$$

Lemma 3.4. *For integer $k \geq 0$, define*

$$\beta_k = \inf_{x \in \mathbb{E}} \left(A_k \Gamma_k(x) + \frac{1}{2} \|x - x_0\|^2 \right) - A_k f(y_k). \tag{13}$$

Then, $\beta_0 = 0$ and

$$\beta_{k+1} \geq \beta_k + \frac{(1 - \sigma^2) A_{k+1}}{2\lambda_{k+1}} \|\tilde{y}_{k+1} - \tilde{x}_k\|^2, \qquad \forall k \geq 0. \tag{14}$$

Proof. Since $A_0 = 0$, we trivially have $\beta_0 = 0$. We will now prove (14). Let an arbitrary $x \in \mathbb{E}$ be given. Define

$$\tilde{x} = \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} x \tag{15}$$

and note that, by (6), (8) and the affinity of γ_{k+1}, we have

$$\tilde{x} - \tilde{x}_k = \frac{a_{k+1}}{A_{k+1}} (x - x_k), \tag{16}$$

$$\gamma_{k+1}(\tilde{x}) = \frac{A_k}{A_{k+1}} \gamma_{k+1}(y_k) + \frac{a_{k+1}}{A_{k+1}} \gamma_{k+1}(x). \tag{17}$$

Using the definition of Γ_{k+1} in (12), and items b) and c) of Lemma 3.2, we conclude that for any $x \in \mathbb{E}$,

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x_0\|^2 &= a_{k+1}\gamma_{k+1}(x) + A_k\Gamma_k(x) + \frac{1}{2}\|x - x_0\|^2 \\ &= a_{k+1}\gamma_{k+1}(x) + A_k\Gamma_k(x_k) + \frac{1}{2}\|x_k - x_0\|^2 + \frac{1}{2}\|x - x_k\|^2 \\ &= a_{k+1}\gamma_{k+1}(x) + A_k f(y_k) + \beta_k + \frac{1}{2}\|x - x_k\|^2, \end{aligned}$$

where the third equality follows from the definition of β_k in (13). Combining the above relation with Lemma 3.2(a), the definition of $\tilde{x}$ in (15), and relations (16) and (17), we conclude that

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x_0\|^2 &\geq a_{k+1}\gamma_{k+1}(x) + A_k\gamma_{k+1}(y_k) + \beta_k + \frac{1}{2}\|x - x_k\|^2 \\ &= \beta_k + A_{k+1}\gamma_{k+1}(\tilde{x}) + \frac{A_{k+1}^2}{2a_{k+1}^2}\|\tilde{x} - \tilde{x}_k\|^2 \\ &= \beta_k + A_{k+1}\gamma_{k+1}(\tilde{x}) + \frac{A_{k+1}}{2\lambda_{k+1}}\|\tilde{x} - \tilde{x}_k\|^2, \end{aligned}$$

where the last equality is due to Lemma 3.1. Using definition (11) with $k = k+1$, the inequality in (5), and Lemma 3.3 with $\tilde{v} = v_{k+1}$, $\tilde{y} = \tilde{y}_{k+1}$, $\tilde{x} = \tilde{x}_k$ and $\varepsilon = \varepsilon_{k+1}$, we conclude that

$$\begin{aligned} \gamma_{k+1}(\tilde{x}) + \frac{1}{2\lambda_{k+1}}\|\tilde{x} - \tilde{x}_k\|^2 &= f(\tilde{y}_{k+1}) + \left(\langle \tilde{x} - \tilde{y}_{k+1}, v_{k+1}\rangle - \varepsilon_{k+1} + \frac{1}{2\lambda_{k+1}}\|\tilde{x} - \tilde{x}_k\|^2\right) \\ &\geq f(\tilde{y}_{k+1}) + \frac{1-\sigma^2}{2\lambda_{k+1}}\|\tilde{y}_{k+1} - \tilde{x}_k\|^2. \end{aligned}$$

Using the non-negativity of A_{k+1} and the above two relations, we conclude that

$$\inf_{x\in\mathbb{E}}\left(A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x_0\|^2\right) \geq \beta_k + A_{k+1}f(\tilde{y}_{k+1}) + \frac{(1-\sigma^2)A_{k+1}}{2\lambda_{k+1}}\|\tilde{y}_{k+1} - \tilde{x}_k\|^2,$$

which, combined with definition (13) with $k = k+1$, and the inequality in step 2) of the A-HPE framework, proves that (14) holds. □

We now make a remark about Lemma 3.4. The only conditions we have used about x_k, y_k, A_k and Γ_k in order to show that (14) holds are that: $A_k \geq 0$, $y_k \in \operatorname{dom} f$, Γ_k is an affine function such that $A_k\Gamma_k \leq A_k f$,

$$x_k = \operatorname{argmin}_{x\in\mathbb{E}}\left(A_k\Gamma_k(x) + \frac{1}{2}\|x - x_0\|^2\right),$$

and that $(x_{k+1}, y_{k+1}, A_{k+1})$ is obtained according to an iteration of the A-HPE framework initialized with the triple (x_k, y_k, A_k).

The next proposition follows directly from Lemma 3.4.

Proposition 3.5. *For every integer $k \geq 1$ and $x \in \mathbb{E}$,*

$$A_k f(y_k) + \frac{1-\sigma^2}{2}\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|\tilde{y}_j - \tilde{x}_{j-1}\|^2 + \frac{1}{2}\|x - x_k\|^2 \leq A_k\Gamma_k(x) + \frac{1}{2}\|x - x_0\|^2.$$

Proof. Adding (14) from $k = 0$ to $k = k-1$ and using the fact that $\beta_0 = 0$, we conclude that

$$\frac{1-\sigma^2}{2}\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|\tilde{y}_j - \tilde{x}_{j-1}\|^2 \leq \beta_k,$$

which together with (13) then imply that

$$A_k f(y_k) + \frac{1-\sigma^2}{2}\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|\tilde{y}_j - \tilde{x}_{j-1}\|^2 \leq \inf_{x'\in\mathbb{E}}\left(A_k\Gamma_k(x) + \frac{1}{2}\|x' - x_0\|^2\right). \tag{18}$$

Using b) and c) of Lemma 3.2, we conclude that for any $x \in \mathbb{E}$,

$$A_k\Gamma_k(x) + \frac{1}{2}\|x - x_0\|^2 = \inf_{x' \in \mathbb{E}} \left(A_k\Gamma_k(x') + \frac{1}{2}\|x' - x_0\|^2 \right) + \frac{1}{2}\|x - x_k\|^2.$$

To end the proof, add $\|x - x_k\|^2/2$ to both sides of (18) and use the above equality. □

The following main result, which establishes the rate of convergence of $f(y_k) - f_*$ and the boundedness of $\{x_k\}$, follows as an immediate consequence of the previous result.

Theorem 3.6. *Let x_* be the projection of x_0 onto X_* and d_0 be the distance of x_0 to X_*. Then, for every integer $k \geq 1$,*

$$\frac{1}{2}\|x_* - x_k\|^2 + A_k\left[f(y_k) - f_*\right] + \frac{1-\sigma^2}{2}\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|\tilde{y}_j - \tilde{x}_{j-1}\|^2 \leq \frac{1}{2}\, d_0^2.$$

As a consequence, for every integer $k \geq 1$,

$$f(y_k) - f_* \leq \frac{d_0^2}{2A_k}, \qquad \|x_k - x_*\| \leq d_0, \tag{19}$$

and, if $\sigma < 1$, then

$$\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|\tilde{y}_j - \tilde{x}_{j-1}\|^2 \leq \frac{d_0^2}{1-\sigma^2}. \tag{20}$$

Proof. This result follows immediately from Proposition 3.5 with $x = x_*$ and Lemma 3.2(b). □

The following result shows how fast A_k grows in terms of the sequence of stepsizes $\{\lambda_k\}$.

Lemma 3.7. *For every integer $k \geq 0$,*

$$\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2}\sqrt{\lambda_{k+1}}. \tag{21}$$

As a consequence, the following statements hold:

a) for every integer $k \geq 1$,

$$A_k \geq \frac{1}{4}\left(\sum_{j=1}^{k}\sqrt{\lambda_j}\right)^2; \tag{22}$$

b) if $\sigma < 1$, then $\sum_{j=1}^{\infty}\|\tilde{y}_j - \tilde{x}_{j-1}\|^2 \leq 4d_0^2/(1-\sigma^2)$.

Proof. Noting that the definition of a_{k+1} in (7) implies that

$$a_{k+1} \geq \frac{\lambda_{k+1}}{2} + \sqrt{\lambda_{k+1}A_k}$$

and using (8), we conclude that

$$A_{k+1} \geq A_k + \left(\frac{\lambda_{k+1}}{2} + \sqrt{\lambda_{k+1}A_k}\right) \geq \left(\sqrt{A_k} + \frac{1}{2}\sqrt{\lambda_{k+1}}\right)^2, \quad \forall k \geq 0,$$

and hence that (21) holds for every $k \geq 0$. Adding (21) from $k = 0$ to $k = k - 1$ and using the fact that $A_0 = 0$, we conclude that

$$\sqrt{A_k} \geq \frac{1}{2} \sum_{j=1}^{k} \sqrt{\lambda_j}, \quad \forall k \geq 1,$$

and hence that a) holds. Statement b) follows from a) and (20). □

The following result follows as an immediate consequence of Theorem 3.6 and Lemma 3.7.

Theorem 3.8. *For every integer $k \geq 1$, we have*

$$f(y_k) - f_* \leq 2d_0^2 \left(\sum_{j=1}^{k} \sqrt{\lambda_j} \right)^{-2}.$$

Theorem 3.8 gives an upper bound on $f(y_k) - f_*$ in terms of the sequence $\{\lambda_k\}$. Depending on the specific instance of the A-HPE, it is possible to further refine this upper bound to obtain an upper bound depending on the iteration count k only. Specific illustrations of that will be given in Sections 4, 5 and 6.

Recall that $v_k \in \partial_{\varepsilon_k} f(\tilde{y}_k)$ in view of the formulation of the A-HPE framework. Since the set of solutions of the inclusion $0 \in \partial f(x)$ is exactly X_*, it follows that the size of $\|v_k\|$ and ε_k provides an optimality measure for $\tilde{y}_k$. The following result provides a certain estimate on these quantities in terms of the sequences $\{\lambda_k\}$ and $\{A_k\}$, and hence $\{\lambda_k\}$ only, in view of Lemma 3.7.

Proposition 3.9. *Assume that $\sigma < 1$. For every integer $k \geq 1$, we have $v_k \in \partial_{\varepsilon_k} f(\tilde{y}_k)$, and there exists $1 \leq i \leq k$ such that*

$$\sqrt{\lambda_i}\|v_i\| \leq \sqrt{\frac{1+\sigma}{1-\sigma}} \frac{d_0}{\sqrt{\sum_{j=1}^{k} A_j}}, \qquad \varepsilon_i \leq \frac{\sigma^2}{2(1-\sigma^2)} \frac{d_0^2}{\sum_{j=1}^{k} A_j}.$$

Proof. For every integer $k \geq 1$, define

$$\tau_k := \max\left\{ \frac{2\varepsilon_k}{\sigma^2}, \frac{\lambda_k \|v_k\|^2}{(1+\sigma)^2} \right\},$$

with the convention $0/0 = 0$. The inequality in (5) with $k = k - 1$, the non-negativity of ε_k and triangle inequality imply that

$$\begin{aligned} &2\lambda_k \varepsilon_k \leq \sigma^2 \|\tilde{y}_k - \tilde{x}_{k-1}\|^2, \\ &\|\lambda_k v_k\| \leq \|\tilde{y}_k - \tilde{x}_{k-1}\| + \|\lambda_k v_k + \tilde{y}_k - \tilde{x}_{k-1}\| \leq (1+\sigma)\|\tilde{y}_k - \tilde{x}_{k-1}\|. \end{aligned}$$

Therefore,

$$\lambda_k \tau_k \leq \|\tilde{y}_k - \tilde{x}_{k-1}\|^2, \quad \forall k \geq 1.$$

Hence, it follows from (20) that

$$\frac{d_0^2}{1-\sigma^2} \geq \sum_{j=1}^{k} A_j \tau_j \geq \left(\min_{j=1,\ldots,k} \tau_j \right) \left(\sum_{j=1}^{k} A_j \right).$$

Noting the definition of τ_k, we easily see that the last inequality implies the conclusion of the proposition. □

Theorem 3.6 shows that the sequence $\{x_k\}$ is bounded. The following result establishes boundedness of $\{y_k\}$, and hence of $\{\tilde{x}_k\}$ in view of (6), under the condition that y_{k+1} is chosen to be $\tilde{y}_{k+1}$ at every iteration of the A-HPE framework.

Theorem 3.10. *Let x_* be the projection of x_0 onto X_* and d_0 be the distance of x_0 to X_*. Consider the A-HPE framework with $\sigma < 1$ and y_{k+1} chosen as $y_{k+1} = \tilde{y}_{k+1}$ for every $k \geq 0$. Then, for every $k \geq 1$,*

$$\|y_k - x_*\| \leq \left(\frac{2}{\sqrt{1-\sigma^2}} + 1\right) d_0.$$

Proof. We first claim that for every integer $k \geq 1$, there holds

$$\|y_k - x_*\| \leq \frac{1}{A_k}\left[\sum_{j=1}^{k} A_j \|y_j - \tilde{x}_{j-1}\|\right] + d_0. \tag{23}$$

We will show this claim by induction on k. Note first that the triangle inequality, relations (6) and (8), the convexity of the norm, and Theorem 3.6, imply that

$$\begin{aligned}
\|y_{k+1} - x_*\| &\leq \|y_{k+1} - \tilde{x}_k\| + \|\tilde{x}_k - x_*\| \\
&\leq \|y_{k+1} - \tilde{x}_k\| + \frac{A_k}{A_{k+1}}\|y_k - x_*\| + \frac{a_{k+1}}{A_{k+1}}\|x_k - x_*\| \\
&\leq \|y_{k+1} - \tilde{x}_k\| + \frac{A_k}{A_{k+1}}\|y_k - x_*\| + \frac{a_{k+1}}{A_{k+1}} d_0, \qquad \forall k \geq 0.
\end{aligned}$$

The previous inequality with $k = 0$ and the fact that $A_0 = 0$ clearly imply (23) with $k = 1$. Assume now that (23) holds for k and let us show it also holds for $k+1$. Indeed, the previous inequality, relation (8) and the induction hypothesis imply that

$$\begin{aligned}
\|y_{k+1} - x_*\| &\leq \|y_{k+1} - \tilde{x}_k\| + \frac{A_k}{A_{k+1}}\left\{\frac{1}{A_k}\left[\sum_{j=1}^{k} A_j \|y_j - \tilde{x}_{j-1}\|\right] + d_0\right\} + \frac{a_{k+1}}{A_{k+1}} d_0 \\
&= \frac{1}{A_{k+1}}\left[\sum_{j=1}^{k+1} A_j \|y_j - \tilde{x}_{j-1}\|\right] + d_0,
\end{aligned}$$

and hence that (23) holds for $k+1$. Hence, the claim follows.

Letting $s_k := \|y_k - \tilde{x}_{k-1}\|$, it follows from Theorem 3.6 and the assumption that $y_k = \tilde{y}_k$ for every $k \geq 1$ that

$$\sum_{j=1}^{k} \frac{A_j}{\lambda_j} s_j^2 \leq \frac{d_0^2}{1-\sigma^2}.$$

Hence, in view of Lemma A.2 with $C = d_0^2/(1-\sigma^2)$, $\alpha_j = A_j$ and $\beta_j = A_j/\lambda_j$ for every $j = 1, \ldots, k$, we conclude that

$$\sum_{j=1}^{k} A_j \|y_j - \tilde{x}_{j-1}\| \leq \frac{d_0}{\sqrt{1-\sigma^2}}\sqrt{\sum_{j=1}^{k} A_j \lambda_j} \leq \frac{d_0}{\sqrt{1-\sigma^2}}\sqrt{A_k}\sqrt{\sum_{j=1}^{k} \lambda_j} \leq \frac{d_0}{\sqrt{1-\sigma^2}}\sqrt{A_k}\sum_{j=1}^{k}\sqrt{\lambda_j},$$

where the second inequality follows from the fact $\{A_k\}$ is increasing and the last one from the fact that the 2-norm is majorized by the 1-norm. The latter inequality together with (23) then imply that

$$\|y_k - x_*\| \leq \frac{d_0}{\sqrt{1-\sigma^2}} \frac{1}{\sqrt{A_k}} \sum_{j=1}^{k} \sqrt{\lambda_j} + d_0.$$

The conclusion of the proposition now follows from the latter inequality and Lemma 3.7. □

4 Large-step A-HPE Framework

In this section, we analyze the convergence rate of a special version of the A-HPE Framework which ensures that the sequence of stepsizes $\{\lambda_k\}$ is not too small. This special version, referred to as the large-step A-HPE framework, will be useful in the analysis of second-order inexact proximal methods for solving (4) discussed in Section 6.

We start by stating the whole large-step A-HPE Framework.

Large-step A-HPE Framework

0) Let $x_0, y_0 \in \mathbb{E}$, $0 \leq \sigma < 1$ and $\theta > 0$ be given, and set $A_0 = 0$ and $k = 0$.

1) If $0 \in \partial f(x_k)$, then **stop**.

2) Otherwise, compute $\lambda_{k+1} > 0$ and a triple $(\tilde{y}_{k+1}, v_{k+1}, \varepsilon_{k+1}) \in \mathbb{E} \times \mathbb{E} \times \mathbb{R}_+$ such that

$$v_{k+1} \in \partial_{\varepsilon_{k+1}} f(\tilde{y}_{k+1}), \qquad \|\lambda_{k+1} v_{k+1} + \tilde{y}_{k+1} - \tilde{x}_k\|^2 + 2\lambda_{k+1}\varepsilon_{k+1} \leq \sigma^2 \|\tilde{y}_{k+1} - \tilde{x}_k\|^2, \tag{24}$$

$$\lambda_{k+1}\|\tilde{y}_{k+1} - \tilde{x}_k\| \geq \theta, \tag{25}$$

where

$$\tilde{x}_k = \frac{A_k}{A_k + a_{k+1}}\, y_k + \frac{a_{k+1}}{A_k + a_{k+1}}\, x_k, \tag{26}$$

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1} A_k}}{2}. \tag{27}$$

3) Choose y_{k+1} such that $f(y_{k+1}) \leq f(\tilde{y}_{k+1})$ and let

$$\begin{aligned} A_{k+1} &= A_k + a_{k+1}, \\ x_{k+1} &= x_k - a_{k+1} v_{k+1}. \end{aligned} \tag{28}$$

4) Set $k \leftarrow k + 1$, and go to step 1.

end

We now make a few remarks about the Large-step A-HPE Framework. First, the large-step A-HPE Framework is similar to the A-HPE Framework, except that it adds a stopping criterion and requires the 'large-step' condition (25) on the quantities λ_{k+1}, $\tilde{x}_k$ and $\tilde{y}_{k+1}$ computed at step

1 of the A-HPE Framework. Clearly, ignoring the stopping criterion in step 1, any implementation of the Large-step A-HPE Framework is also an instance of the A-HPE Framework. Second, as in the A-HPE Framework, the above framework does not specify how the quantities λ_{k+1} and $(\tilde{y}_{k+1}, v_{k+1}, \varepsilon_{k+1})$ are computed in step 2. An specific implementation of the above framework will be described in Sections 6 and 7 in which these quantities are computed by solving subproblems based on second-order approximations of f. Third, due to statement b) of Lemma 3.7 and the first remark above, any instance of the large-step A-HPE Framework satisfies $\lim_{k\to\infty} \|\tilde{y}_{k+1} - \tilde{x}_k\| = 0$, and hence $\lim_{k\to\infty} \lambda_k = \infty$, due to the 'large-step' condition (25). As a result, the Large-step A-HPE Framework does not contain variants of Nesterov's method where a prox-type subproblem based on a first-order approximation of f and with bounded stepsize λ_k is solved at the k-th iteration.

In what follows, we study the complexity of the large-step A-HPE framework. For simplicity of exposition, the convergence results presented below implicitly assume that the framework does not stop in a finite number of iterations. However, they can easily be restated without assuming such condition by saying that either the conclusion stated thereof holds or x_k is a solution of (4).

The main result we intend to prove in this section is as follows.

Theorem 4.1. *Let d_0 denote the distance of x_0 to X_*. For every integer $k \geq 1$, the following statements hold:*

a) there holds

$$f(y_k) - f_* \leq \frac{3^{7/2}}{2\sqrt{2}} \frac{d_0^3}{\theta\sqrt{1-\sigma^2}} \frac{1}{k^{7/2}}; \tag{29}$$

b) $v_k \in \partial_{\varepsilon_k} f(\tilde{y}_k)$ and there exists $i \leq k$ such that

$$\|v_i\| = \mathcal{O}\left(\frac{d_0^2}{\theta k^3}\right), \qquad \varepsilon_i = \mathcal{O}\left(\frac{d_0^3}{\theta\, k^{9/2}}\right). \tag{30}$$

Before giving the proof of the above result, we will first establish a number of technical results. Note that we are assuming $0 \leq \sigma < 1$.

Lemma 4.2. *If, for some constants $b > 0$ and $\xi \geq 0$, there holds*

$$A_k \geq b k^{\xi}, \qquad \forall k \geq 1, \tag{31}$$

then

$$A_k \geq \frac{w b^{1/3}}{(\xi/7+1)^{7/3}} k^{(\xi+7)/3}, \qquad \forall k \geq 1,$$

where

$$w = \frac{1}{4}\left(\frac{\theta^2(1-\sigma^2)}{d_0^2}\right)^{1/3}. \tag{32}$$

Proof. First we claim that for every integer $k \geq 1$,

$$\sum_{j=1}^{k} \frac{A_j}{\lambda_j^3} \leq \frac{d_0^2}{\theta^2(1-\sigma^2)}. \tag{33}$$

To prove this claim, use the large step condition (25) and inequality (20) on Theorem 3.6 to conclude that

$$\sum_{j=1}^{k} \frac{A_j}{\lambda_j^3}\theta^2 \leq \sum_{j=1}^{k} \frac{A_j}{\lambda_j^3}\left(\lambda_j\|\tilde{y}_j - \tilde{x}_{j-1}\|\right)^2 = \sum_{j=1}^{k} \frac{A_j}{\lambda_j}\|\tilde{y}_j - \tilde{x}_{j-1}\|^2 \leq \frac{d_0^2}{1-\sigma^2}.$$

and divide the above inequalities by θ^2.

Using (33) and Lemma A.1 with

$$C = \frac{d_0^2}{\theta^2(1-\sigma^2)}, \qquad \alpha_j = A_j, \qquad t_j = \sqrt{\lambda_j}, \qquad \forall j = 1, \ldots, k,$$

we conclude that

$$\sum_{j=1}^{k} \sqrt{\lambda_j} \geq \frac{1}{C^{1/6}} \left(\sum_{j=1}^{k} A_j^{1/7}\right)^{7/6},$$

which, combined with Lemma 3.7 shows that, for every $k \geq 1$,

$$A_k \geq w \left(\sum_{j=1}^{k} A_j^{1/7}\right)^{7/3}, \tag{34}$$

Assume that (31) holds. Then, using (34), we have

$$A_k \geq wb^{1/3} \left(\sum_{j=1}^{k} j^{\xi/7}\right)^{7/3}, \qquad \forall k \geq 1.$$

Since $0 \leq t \mapsto t^{\xi/7}$ is non-decreasing, we have

$$\sum_{j=1}^{k} j^{\xi/7} \geq \int_0^k t^{\xi/7}\, dt = \frac{1}{\xi/7+1} k^{\xi/7+1}.$$

The conclusion of the lemma now follows by combining the above two inequalities.

□

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1: We first claim that for every integer $i \geq 0$, we have

$$A_k \geq b_i k^{\xi_i}, \qquad \forall k \geq 1, \tag{35}$$

where

$$b_i := \tilde{w}\left(\frac{A_1}{\tilde{w}}\right)^{\frac{1}{3^i}}, \qquad \xi_i := \frac{7}{2}\left(1 - 3^{-i}\right), \qquad \tilde{w} := \left(\frac{w}{(3/2)^{7/3}}\right)^{3/2}. \tag{36}$$

We will prove this claim by induction on i. The claim is obviously true for $i = 0$ since in this case $b_0 = A_1$ and $\xi_0 = 0$, and hence $b_0 k^{\xi_0} = A_1 \leq A_k$ for every $k \geq 1$. Assume now that the result holds for $i \geq 0$. Using this assumption and Lemma 4.2 with $b = b_i$ and $\xi = \xi_i$, we conclude that

$$A_k \geq \frac{wb_i^{1/3}}{(\xi_i/7+1)^{7/3}} k^{(\xi_i+7)/3}, \qquad \forall k \geq 1,$$

Since (36) implies that

$$\frac{\xi_i+7}{3}=\frac{1}{3}\left[\frac{7}{2}(1-3^{-i})+7\right]=\frac{7}{2}(1-3^{-(i+1)})=\xi_{i+1}$$

and

$$\frac{wb_i^{1/3}}{(\xi_i/7+1)^{7/3}}\geq\frac{wb_i^{1/3}}{(3/2)^{7/3}}=\tilde{w}^{2/3}b_i^{1/3}=\tilde{w}^{2/3}\left(\tilde{w}\left(\frac{A_1}{\tilde{w}}\right)^{\frac{1}{3^i}}\right)^{1/3}=\tilde{w}\left(\frac{A_1}{\tilde{w}}\right)^{\frac{1}{3^{i+1}}}=b_{i+1},$$

we conclude that (35) holds for $i+1$. Hence, we conclude that (35) holds for every $i\geq 0$.

Now letting i goes to ∞ in (35) and noting that $\lim_{i\to\infty}b_i=\tilde{w}$ and $\lim_{i\to\infty}\xi_i=7/2$, we conclude that

$$A_k\geq\tilde{w}k^{7/2}=\left(\frac{2}{3}\right)^{\frac{7}{2}}w^{3/2}k^{7/2}=\left(\frac{2}{3}\right)^{\frac{7}{2}}\left(\frac{\theta(1-\sigma^2)^{1/2}}{8d_0}\right)k^{7/2}.$$

Statement a) now follows from the last inequality and the first inequality in (19). Moreover, the above inequality together with Proposition 3.9 imply the existence of $i\leq k$ such that the second estimate in (30) holds and

$$\sqrt{\lambda_i}\|v_i\|=\mathcal{O}\left(\frac{d_0^{3/2}}{\theta^{1/2}k^{9/4}}\right).$$

Now, it is easily seen that the inequality in (24) and the large step condition (25) imply that

$$\lambda_i^2\|v_i\|\geq\theta(1-\sigma).$$

Combining the last two inequalities, we now easily see that the first estimate in (30) also holds.

5 Application I: First-order methods

In this section, we use the theory presented in the previous sections to analyze a first-order implementation of the A-HPE framework for solving structured convex optimization problems.

The problem of interest in this section is

$$f_*:=\min\{f(x):=g(x)+h(x):x\in\mathbb{E}\},\tag{37}$$

where the following conditions are assumed to hold:

A.1) $g,h:\mathbb{E}\to\bar{\mathbb{R}}$ are proper closed convex functions;

A.2) g is differentiable on a closed convex set $\Omega\supseteq\mathrm{dom}\,(h)$;

A.3) ∇g is L_0-Lipschitz continuous on Ω;

A.4) the solution set X_* of (37) is non-empty.

Under the above assumptions, it can be easily shown that problem (37) is equivalent to the inclusion

$$0 \in (\nabla g + \partial h)(x). \tag{38}$$

We now state a specific implementation of the A-HPE framework for solving (37) under the above assumptions. In the following sections, we let P_Ω denote the projection operator onto Ω.

Algorithm I:

0) Let $x_0, y_0 \in \mathbb{E}$ and $0 < \sigma \leq 1$ and set be given, and set $A_0 = 0$, $\lambda = \sigma^2/L_0$ and $k = 0$.

1) Define

$$a_{k+1} = \frac{\lambda + \sqrt{\lambda^2 + 4\lambda A_k}}{2} \tag{39}$$

$$\tilde{x}_k = \frac{A_k}{A_k + a_{k+1}}\, y_k + \frac{a_{k+1}}{A_k + a_{k+1}}\, x_k, \tag{40}$$

and compute

$$x'_k = P_\Omega(\tilde{x}_k), \qquad y_{k+1} = (I + \lambda \partial h)^{-1}(\tilde{x}_k - \lambda \nabla g(x'_k)). \tag{41}$$

2) Define

$$\begin{aligned} A_{k+1} &= A_k + a_{k+1}, \\ x_{k+1} &= x_k - \frac{a_{k+1}}{\lambda}(\tilde{x}_k - y_{k+1}). \end{aligned} \tag{42}$$

3) Set $k \leftarrow k + 1$, and go to step 1.

end

Note that the computation of y_{k+1} in (41) is equivalent to solving

$$y_{k+1} = \arg\min_{x \in \mathbb{E}} \left(\lambda \left[\langle \nabla g(x'_k), x \rangle + h(x) \right] + \frac{1}{2}\|x - \tilde{x}_k\|^2 \right). \tag{43}$$

The following result shows that Algorithm I is a special case of the A-HPE framework with $\lambda_k = \lambda$ for all $k \geq 1$.

Proposition 5.1. *Define for every $k \geq 0$,*

$$\lambda_{k+1} = \lambda, \qquad \tilde{y}_{k+1} = y_{k+1}, \qquad v_{k+1} = \frac{1}{\lambda}(\tilde{x}_k - y_{k+1}), \tag{44}$$

$$\varepsilon_{k+1} = g(y_{k+1}) - [g(x'_k) + \langle y_{k+1} - x'_k, \nabla g(x'_k) \rangle]. \tag{45}$$

Then

$$v_{k+1} \in \partial_{\varepsilon_{k+1}}(g + h)(\tilde{y}_{k+1}), \qquad \|\lambda_{k+1} v_{k+1} + \tilde{y}_{k+1} - \tilde{x}_k\|^2 + 2\lambda_{k+1}\varepsilon_{k+1} \leq \sigma^2 \|\tilde{y}_{k+1} - \tilde{x}_k\|^2.$$

As a consequence, Algorithm I is a particular case of the A-HPE framework.

Proof. Using (41), (44), (45) and Proposition 2.3(b) we easily see that

$$v_{k+1} - \nabla g(x'_k) \in \partial h(y_{k+1}), \qquad \nabla g(x'_k) \in \partial_{\varepsilon_{k+1}} g(y_{k+1}), \tag{46}$$

Since $(\partial_\varepsilon g + \partial h)(y) \subseteq \partial_\varepsilon (g+h)(y)$, we conclude that

$$v_{k+1} \in \partial_{\varepsilon_{k+1}} (g+h)(y_{k+1}).$$

Noting that $y_{k+1} \in \operatorname{dom} \partial h \subseteq \Omega$ and $x'_k = P_\Omega(\tilde{x}_k) \in \Omega$, it follows from (45), Assumption A.3, and a well-known property of $P_\Omega(\cdot)$, that

$$\varepsilon_{k+1} \leq \frac{L_0}{2}\|y_{k+1} - x'_k\|^2 \leq \frac{L_0}{2}\|y_{k+1} - \tilde{x}_k\|^2,$$

This inequality together with (44) and the fact that $\lambda = \sigma^2/L_0$ then imply that

$$\|\lambda v_{k+1} + y_{k+1} - \tilde{x}_k\|^2 + 2\lambda\varepsilon_{k+1} = 2\lambda\varepsilon_{k+1} \leq \lambda L_0 \|y_{k+1} - \tilde{x}_k\|^2 = \sigma^2 \|y_{k+1} - \tilde{x}_k\|^2.$$

Moreover, note that (44) implies that the update formula for x_{k+1} in step 2 of Algorithm I is the same as that of the A-HPE framework. Hence, the result follows. □

The following result gives the complexity estimation of Algorithm I as a consequence of Proposition 5.1 and the complexity results for the A-HPE framework on Section 3.

Proposition 5.2. *Consider the sequence $\{y_k\}$ generated by Algorithm I, the sequences $\{v_k\}$ and $\{\varepsilon_k\}$ defined as in Proposition 5.1 and the sequence $\{w_k\}$ defined as*

$$w_k = v_k + \nabla g(y_k) - \nabla g(x'_{k-1}) \qquad k \geq 1.$$

Then, the the following statements hold:

a) for every $k \geq 1$,

$$y_k \in \operatorname{dom}(h) \subseteq \Omega, \quad f(y_k) - f_* \leq \frac{2L_0 d_0^2}{k^2\sigma^2};$$

b) if $\sigma < 1$, then

$$\|y_k - x_*\| \leq \left(\frac{2}{\sqrt{1-\sigma^2}} + 1\right) d_0;$$

c) if $\sigma < 1$, then for every $k \geq 1$, $v_k \in \partial_{\varepsilon_k} g(y_k) + \partial h(y_k) \subseteq \partial_{\varepsilon_k}(g+h)(y_k)$, and there exists $i \leq k$ such that

$$\|v_i\| = \mathcal{O}\left(\frac{L_0 d_0}{k^{3/2}}\right), \qquad \varepsilon_i = \mathcal{O}\left(\frac{L_0 d_0^2}{k^3}\right);$$

d) if $\sigma < 1$, then for every $k \geq 1$, $w_k \in (\nabla g + \partial h)(y_k)$, and there exists $i \leq k$ such that

$$\|w_i\| = \mathcal{O}\left(\frac{L_0 d_0}{k^{3/2}}\right).$$

Proof. Statements a) and b) follow from Proposition 5.1, Theorem 3.8 with $\lambda_k = \lambda := \sigma^2/L_0$ and Theorem 3.10. Statement c) follows from Proposition 5.1, Lemma 3.7(a), Proposition 3.9 and the fact that $\lambda_k = \lambda := \sigma^2/L_0$. The inclusion in d) follows from the first inclusion in (46) with $k = k-1$ and the definition of w_k. Noting that $y_k \in \operatorname{dom} \partial h \subseteq \Omega$ and $x'_{k-1} = P_\Omega(\tilde{x}_{k-1}) \in \Omega$, it follows from the definition of w_k, Assumption A.3, the non-expansiveness of $P_\Omega(\cdot)$ and the last equality in (44), that

$$\|w_k\| \leq \|v_k\| + \|\nabla g(y_k) - \nabla g(x'_{k-1})\| \leq \|v_k\| + L_0\|y_k - \tilde{x}_{k-1}\| = (1 + \lambda L_0)\,\|v_k\|.$$

The complexity estimation in d) now follows from the previous inequality, statement c) and the definition of λ. □

It can be shown that Algorithm I for the case in which $\Omega = \mathbb{E}$, and hence ∇g is defined and is L_0-Lipschitz continuous on the whole $\mathbb{E}$, reduces to a variant of Nesterov's method, namely the FISTA method in [1] (see also Algorithm 2 of [22]). In this case, $x'_k = \tilde{x}_k$ and only the resolvent in (41), or equivalently the minimization subproblem (43), has to be computed at iteration k, while in the general case where $\Omega \neq \mathbb{E}$, the projection of $\tilde{x}_k$ onto Ω must also be computed in order to determine x'_k.

In summary, we have seen above that the A-HPE contains a variant of Nesterov's optimal method for (37) when $\Omega = \mathbb{E}$, and also an extension of this variant when $\Omega \neq \mathbb{E}$. The example of this section also shows that the A-HPE is a natural framework for generalizing Nesterov's acceleration schemes. We will also see in the next section that it is a suitable framework for designing and analyzing second-order proximal methods for (37).

6 Application II: Second-order methods

We will now apply the theory outlined in Sections 3 and 4 to analyze an accelerated Newton proximal extragradient (A-NPE) method, which is an accelerated version of the method presented in [10] for solving a monotone nonlinear equation.

In this section, our problem of interest is the same as that of Section 5, namely problem (37). However, in this section, we impose a different set of assumptions on (37), i.e.:

C.1) g and h are proper closed convex functions;

C.2) g is twice-differentiable on a closed convex set Ω such that $\Omega \supseteq \operatorname{Dom}(\partial h)$;

C.3) $g''(\cdot)$ is L_1-Lipschitz continuous on Ω.

Recall that, for the monotone inclusion problem (38), the exact proximal iteration y from x with stepsize $\lambda > 0$ is defined as

$$y = \operatorname{argmin}_{u \in \mathbb{E}} g(u) + h(u) + \frac{1}{2\lambda}\|u - x\|^2. \tag{47}$$

The accelerated NPE method of this section is based on inexact solutions of the minimization problem

$$\min_{u \in \mathbb{E}} g_x(u) + h(u) + \frac{1}{2\lambda}\|u - x\|^2, \tag{48}$$

where g_x is the second-order approximation of g at x with respect to Ω defined as

$$g_x(y) = g(\bar{x}) + \langle \nabla g(\bar{x}), y - \bar{x} \rangle + \frac{1}{2}\langle y - \bar{x}, g''(\bar{x})(y - \bar{x}) \rangle, \qquad \bar{x} := P_\Omega(x). \tag{49}$$

Note that the unique solution y_x of (48) together with the vector $v := (x - y_x)/\lambda$ are characterized by the optimality condition

$$v \in (\nabla g_x + \partial h)(y_x), \qquad \lambda v + y_x - x = 0. \tag{50}$$

We will consider the following notion of approximate solution for (50), and hence of (48).

Definition 6.1. *Given $(\lambda, x) \in \mathbb{R}_{++} \times \mathbb{E}$ and $\hat{\sigma} \geq 0$, the triple $(y, u, \varepsilon) \in \mathbb{E} \times \mathbb{E} \times \mathbb{R}_+$ is called a $\hat{\sigma}$-approximate Newton solution of (47) at (λ, x) if*

$$u \in (\nabla g_x + \partial_\varepsilon h)(y), \qquad \|\lambda u + y - x\|^2 + 2\lambda\varepsilon \leq \hat{\sigma}^2 \|y - x\|^2.$$

We now make a few remarks about the above definition. First, if (y_x, v) is the solution pair of (50), then $(y_x, v, 0)$ is a $\hat{\sigma}$-approximate Newton solution of (47) at (λ, x) for any $\hat{\sigma} \geq 0$. Second, if h is the indicator function δ_X of a closed convex set $X \subseteq E$, then exact computation of the pair (y_x, v) boils down to minimizing a strongly convex quadratic function over X.

The next result shows how an approximate Newton solution can be used to generate a triple (y, v, ε) satisfying (5) with $f = g + h$.

Lemma 6.2. *Suppose that the triple $(y, u, \varepsilon) \in \mathbb{E} \times \mathbb{E} \times \mathbb{R}_+$ is a $\hat{\sigma}$-approximate Newton solution of (47) at (λ, x) and define*

$$v = \nabla g(y) + u - \nabla g_x(y), \qquad \sigma = \hat{\sigma} + \frac{L_1}{2}\lambda \|y - x\|.$$

Then,

$$v \in (\nabla g + \partial_\varepsilon h)(y) \subseteq \partial_\varepsilon (g + h)(y), \qquad \|\lambda v + y - x\|^2 + 2\lambda\varepsilon \leq \sigma^2 \|y - x\|^2.$$

Proof. Direct use of the inclusion in Definition 6.1 together with the definition of v shows that

$$v = u + \nabla g(y) - \nabla g_x(y) \in (\nabla g_x + \partial_\varepsilon h)(y) + \nabla g(y) - \nabla g_x(y) = (\nabla g + \partial_\varepsilon h)(y),$$

which proves the first inclusion. The second inclusion follows trivially from the first one and basic properties of the ε-subdifferential. Now, Definition 6.1, Assumption C.2 and a basic property of the ε-subdifferential imply that

$$y \in \mathrm{Dom}\,(\partial_\varepsilon h) \subseteq \mathrm{cl}\,(\partial h) \subseteq \Omega.$$

Letting $\bar{x} = P_\Omega(x)$ and using the above conclusion, Assumption C.3 and (49), we conclude that

$$\|v - u\| = \|\nabla g(y) - \nabla g_x(y)\| = \left\|\nabla g(y) - (\nabla g(\bar{x}) + g''(\bar{x})(y - \bar{x})\right\| \leq \frac{L_1}{2}\|y - \bar{x}\|^2 \leq \frac{L_1}{2}\|y - x\|^2,$$

where the last inequality follows from the fact that P_Ω is a non-expansive map. The above inequality, the triangle inequality for norms, the definition of σ and the inequality in Definition 6.1 then imply that

$$\begin{aligned}
\|\lambda v + y - x\|^2 + 2\lambda\varepsilon &\leq (\lambda\|v - u\| + \|\lambda u + y - x\|)^2 + 2\lambda\varepsilon \\
&\leq \left(\lambda\|v - u\| + \sqrt{\|\lambda u + y - x\|^2 + 2\lambda\varepsilon}\right)^2 \\
&\leq \left(\frac{L_1}{2}\lambda\|y - x\|^2 + \hat{\sigma}\|y - x\|\right)^2 = \sigma^2\|y - x\|^2.
\end{aligned}$$

□

We now state the accelerated NPE method based on the above notion of approximate solutions.

Accelerated Newton Proximal Extragradient (A-NPE) Method:

0) Let $x_0, y_0 \in \mathbb{E}$, $\hat{\sigma} \geq 0$ and $0 < \sigma_\ell < \sigma_u < 1$ such that

$$\sigma := \hat{\sigma} + \sigma_u < 1, \qquad \sigma_\ell(1+\hat{\sigma}) < \sigma_u(1-\hat{\sigma}) \tag{51}$$

be given, and set $A_0 = 0$ and $k = 1$.

1) If $0 \in \partial f(x_k)$, then **stop**.

2) Otherwise, compute a positive scalar λ_{k+1} and a $\hat{\sigma}$-approximate Newton solution $(y_{k+1}, u_{k+1}, \varepsilon_{k+1}) \in \mathbb{E} \times \mathbb{E} \times \mathbb{R}_+$ of (47) at $(\lambda_{k+1}, \tilde{x}_k)$ satisfying

$$\frac{2\sigma_\ell}{L_1} \leq \lambda_{k+1}\|\tilde{y}_{k+1} - \tilde{x}_k\| \leq \frac{2\sigma_u}{L_1}, \tag{52}$$

where

$$\tilde{x}_k = \frac{A_k}{A_k + a_{k+1}}\, y_k + \frac{a_{k+1}}{A_k + a_{k+1}} x_k, \tag{53}$$

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1}A_k}}{2}. \tag{54}$$

3) Choose y_{k+1} such that $f(y_{k+1}) \leq f(\tilde{y}_{k+1})$ and let

$$v_{k+1} = \nabla g(\tilde{y}_{k+1}) + u_{k+1} - \nabla g_{\tilde{x}_k}(\tilde{y}_{k+1}), \tag{55}$$

$$A_{k+1} = A_k + a_{k+1}, \tag{56}$$

$$x_{k+1} = x_k - a_{k+1}v_{k+1}.$$

4) Set $k \leftarrow k+1$, and go to step 1.

end

Define, for each k,

$$\sigma_k := \hat{\sigma} + \frac{L_1}{2}\lambda_k\|\tilde{y}_k - \tilde{x}_{k-1}\|. \tag{57}$$

We will now establish that the inexact accelerated NPE method can be viewed as a special case of the large-step A-HPE framework.

Proposition 6.3. *Let σ be defined as in* (51)*. Then, for each $k \geq 0$, $\sigma_{k+1} \leq \sigma$ and*

$$v_{k+1} \in \left(\nabla g + \partial_{\varepsilon_{k+1}} h\right)(\tilde{y}_{k+1}) \subseteq \partial_{\varepsilon_{k+1}}(g+h)(\tilde{y}_{k+1}), \tag{58}$$

$$\|\lambda_{k+1}v_{k+1} + \tilde{y}_{k+1} - \tilde{x}_k\|^2 + 2\lambda_{k+1}\varepsilon_{k+1} \leq \sigma_{k+1}^2\|\tilde{y}_{k+1} - \tilde{x}_k\|^2.$$

As a consequence of (52) *and* (55)*, it follows that: the accelerated NPE method is a special case of the large-step A-HPE framework stated in Section 3 with $f = g + h$ and $\theta = 2\sigma_\ell/L_1$.*

Proof. The inequality on σ_{k+1} follows from (52) and the definition of σ in (51). The inclusion and the other inequality follow from the fact that $(y_{k+1}, u_{k+1}, \varepsilon_{k+1})$ is a $\hat{\sigma}$-approximate Newton solution of (47) at $(\lambda_{k+1}, \tilde{x}_k)$, relation (55) and Lemma 6.2 with $(x, \lambda) = (\tilde{x}_k, \lambda_{k+1})$ and $(y, u, \varepsilon) = (y_{k+1}, u_{k+1}, \varepsilon_{k+1})$. The last claim of the proposition follows from its first part and the first inequality in (52). □

As a consequence of the above result, it follows that all the convergence rate and complexity results derived for the A-HPE and the large-step A-HPE framework hold for A-NPE method.

Theorem 6.4. *Let d_0 denote the distance of x_0 to X_* and consider the sequences $\{x_k\}$, $\{y_k\}$, $\{\tilde{y}_k\}$, $\{v_k\}$ and $\{\varepsilon_k\}$ generated by the A-NPE method. Then, for every $k \geq 1$, the following statements hold:*

a) $\|x_k - x_*\| \leq d_0$ *and*

$$f(y_k) - f_* \leq \frac{3^{7/2}}{4\sqrt{2}} \frac{L_1 d_0^3}{\sigma_\ell \sqrt{1-\sigma^2}} \frac{1}{k^{7/2}}, \tag{59}$$

where σ is given by (51);

b) $v_k \in \nabla g(\tilde{y}_k) + \partial_{\varepsilon_k} h(\tilde{y}_k)$, *and there exists $i \leq k$ such that*

$$\|v_i\| = \mathcal{O}\left(\frac{L_1 d_0^2}{k^3}\right), \qquad \varepsilon_i = \mathcal{O}\left(\frac{L_1 d_0^3}{k^{9/2}}\right). \tag{60}$$

Proof. This result follows immediately from Theorem 4.1, Proposition 6.3 and the fact that $\theta = 2\sigma_\ell / L_1$. □

Assuming that we have at our disposal a black-box which is able to compute a $\hat{\sigma}$-approximate Newton solution at any given pair $(x, \lambda) \in \mathbb{E} \times \Re_{++}$, the next section describes a line-search procedure, and corresponding complexity bounds, for finding a stepsize $\lambda_{k+1} > 0$, and hence the corresponding $\tilde{x}_k$ given by (53)-(54), which, together with a $\hat{\sigma}$-approximate Newton solution $(y_{k+1}, u_{k+1}, \varepsilon_{k+1})$ at $(\lambda_{k+1}, \tilde{x}_k)$ output by the black-box, satisfy condition (52).

We note that a simpler line-search procedure has been described in [11] which accomplishes the same goal under the simpler assumption that the base point $\tilde{x}_k$ does not depend on the choice of λ_{k+1}.

7 Line Search

The main goal of this section is to present a line search procedure for implementing step 2 of the A-NPE method. This section contains four subsections as follows. Subsection 7.1 reviews some technical results about the resolvent of a maximal monotone operator. Subsection 7.2 introduces a certain structured monotone inclusion problem of which (38) is a special case, and studies some properties of the resolvents of maximal monotone operators obtained by linearizing the operator of this inclusion problem. Subsection 7.3 presents a line search procedure in the more general setting of the structured monotone inclusion problem. Finally, Subsection 7.4 specializes the line search procedure of the previous subsection to the context of (38) in order to obtain an implementation of step 2 of the A-NPE method.

7.1 Preliminary results

Let a maximal monotone operator $B : \mathbb{E} \rightrightarrows \mathbb{E}$ and $x \in \mathbb{E}$ be given and define for each $\lambda > 0$,

$$y_B(\lambda; x) := (I + \lambda B)^{-1}(x), \qquad \varphi_B(\lambda; x) := \lambda \|y_B(\lambda; x) - x\|. \tag{61}$$

In subsection, we describe some basic properties of φ_B that will be needed in our presentation.

The point $y_B(\lambda; x)$ is the exact proximal point iteration from x with stepsize $\lambda > 0$ with respect to the inclusion $0 \in B(x)$. Note that $y_B(\lambda; x)$ is the unique solution y of the inclusion

$$0 \in (\lambda B + I)(y) - x = \lambda B(y) + y - x, \tag{62}$$

or equivalently, the y-component of the unique solution (y, v) of the inclusion/equation

$$v \in B(y), \qquad \lambda v + y - x = 0. \tag{63}$$

The following result, whose proof can be found in Lemma 4.3 of [10], describes some basic properties of the function $\lambda \mapsto \varphi_B(\lambda; x)$.

Proposition 7.1. *For every $x \in \mathbb{E}$, the following statements hold:*

a) $\lambda > 0 \to \varphi_B(\lambda; x)$ is a continuous function;

b) for every $0 < \tilde{\lambda} \leq \lambda$,

$$\frac{\lambda}{\tilde{\lambda}} \varphi_B(\tilde{\lambda}; x) \leq \varphi_B(\lambda; x) \leq \left(\frac{\lambda}{\tilde{\lambda}}\right)^2 \varphi_B(\tilde{\lambda}; x). \tag{64}$$

The following definition introduces the notion of an approximate solution of (63).

Definition 7.2. *Given $\hat{\sigma} \geq 0$, the triple (y, v, ε) is said to be a $\hat{\sigma}$-approximate solution of (63) at (λ, x) if*

$$v \in B^{\varepsilon}(y), \quad \|\lambda v + y - x\|^2 + 2\lambda\varepsilon \leq \hat{\sigma}^2 \|y - x\|^2. \tag{65}$$

Note that $(y, v, \varepsilon) = (\tilde{y}_{k+1}, v_{k+1}, \varepsilon_{k+1})$, where $\tilde{y}_{k+1}$, v_{k+1} and ε_{k+1} are as in step 2 of the Large-step A-HPE Framework, is a σ-approximate solution of (63) with $B = \partial f$ at $(\lambda_{k+1}, \tilde{x}_k)$, due to Proposition 2.3(a). Note also that $(y, v, \varepsilon) = (y_{k+1}, u_{k+1}, \varepsilon_{k+1})$, where y_{k+1}, u_{k+1} and ε_{k+1} are as in step 2 of the A-NPE method, is a σ-approximate solution of (63) with $B = (\nabla g_{\tilde{x}_k} + \partial h)$ at $(\lambda_{k+1}, \tilde{x}_k)$, due to the fact that

$$(\nabla g_{\tilde{x}_k} + \partial_\varepsilon h)(y) \subseteq \partial_\varepsilon (g_{\tilde{x}_k} + h)(y) \subseteq [\partial (g_{\tilde{x}_k} + h)]^\varepsilon (y), \qquad \forall y \in \mathbb{E}.$$

Note also that the conditions (25) and (52) that appear in these methods are conditions on the quantity $\lambda\|y - x\|$. The following result, whose proof can be found in Lemma 4.1 of [10], shows that the quantity $\lambda\|y - x\|$, where (y, v, ε) is a $\hat{\sigma}$-approximate solution of (63) at (λ, x), can be well-approximated by $\varphi_B(x; \lambda)$.

Proposition 7.3. *Let $x \in \mathbb{E}$, $\lambda > 0$ and $\hat{\sigma} \geq 0$ be given. If (y, v, ε) is a $\hat{\sigma}$-approximate solution of (63) at (λ, x), then*

$$(1 - \hat{\sigma})\lambda\|y - x\| \leq \varphi_B(\lambda; x) \leq (1 + \hat{\sigma})\lambda\|y - x\|. \tag{66}$$

7.2 Technical Results

In this subsection, we describe the monotone inclusion problem in the context of which the line search procedure of Subsection 7.3 will be presented. It contains the inclusion (38) as a special case, and hence any line search procedure described in the context of this inclusion problem will also work in the setting of (38). We will also establish a number of preliminary results for the associated function φ_B in this setting.

In this subsection, we consider the monotone inclusion problem

$$0 \in T(x) := (G + H)(x), \tag{67}$$

where $G : \operatorname{Dom} G \subseteq \mathbb{E} \to \mathbb{E}$ and $H : \mathbb{E} \rightrightarrows \mathbb{E}$ satisfy

C.1) H is a maximal monotone operator;

C.2) G is monotone and differentiable on a closed convex set Ω such that $\operatorname{Dom} H \subseteq \Omega \subseteq \operatorname{Dom} G$;

C.3) G' is L-Lipschitz continuous on Ω.

Observe that the monotone inclusion problem (38) is a special case of (67) in which $G = \nabla g$ and $H = \partial h$. Also, under the above assumptions, it can be shown using Proposition A.1 of [8] that $T = G + H$ is a maximal monotone operator.

Recall that, for the monotone inclusion problem (67), the exact proximal iteration from x with stepsize $\lambda > 0$ is the unique solution y of the inclusion

$$0 \in \lambda(G + H)(y) + y - x, \tag{68}$$

or equivalently, the y-component of the unique solution (y, v) of the inclusion/equation

$$v \in (G + H)(y), \qquad \lambda v + y - x = 0. \tag{69}$$

For $x \in \mathbb{E}$, define the 'first-order approximation' of $T_x : \mathbb{E} \rightrightarrows \mathbb{E}$ of T at x as

$$T_x(y) = G_x(y) + H(y), \quad \forall y \in \mathbb{E},$$

where $G_x : \mathbb{E} \to \mathbb{E}$ is the *first-order approximation of G at x with respect to Ω* given by

$$G_x(y) = G(P_\Omega(x)) + G'(P_\Omega(x))(y - x).$$

Lemma 7.4. *For every $x \in \mathbb{E}$ and $y \in \Omega$,*

$$\|G(y) - G_x(y)\| \leq \frac{L}{2}\|y - x\|^2.$$

Proof. Use the fact that $G(y) - G_x(y)$ is a linearization error, Assumption C.3 and the fact that P_Ω is non-expansive. □

Finally, note that when $G = \nabla g$ and $H = \partial h$, where g and h are as in Section 6, then $T_x = \nabla g_x + \partial h$. In view of Proposition 7.3 and the fact that the approximate Newton solutions generated by A-NPE method are approximate solutions of operators of the form $T_x = \nabla g_x + \partial h$ where the base point x depends on the choice of the stepsize λ, it is important to understand how the quantity $\varphi_{T_x}(\lambda; x)$ behaves in terms of λ and x in order to develop a scheme for computing $\lambda = \lambda_{k+1}$ satisfying condition (52). The dependence of $\varphi_{T_x}(\lambda; x)$ in terms of λ follows from Proposition 7.1, while its dependence in terms of x follows from the next result.

Lemma 7.5. *Let $x, \tilde{x} \in \mathbb{E}$ and $\lambda > 0$ be given and define $B := T_x$ and $\tilde{B} := T_{\tilde{x}}$. Then,*

$$|\varphi_B(\lambda; x) - \varphi_{\tilde{B}}(\lambda; \tilde{x})| \leq \lambda\|\tilde{x} - x\| + L\lambda^2\|\tilde{x} - x\|^2 + 2L\lambda\|\tilde{x} - x\|\eta. \tag{70}$$

where

$$\eta := \min\left\{\varphi_B(\lambda; x)\,,\, \varphi_{\tilde{B}}(\lambda; \tilde{x})\right\}.$$

As a consequence,

$$\varphi_B(\lambda; x) \leq \lambda\|\tilde{x} - x\| + L\lambda^2\|\tilde{x} - x\|^2 + (2L\lambda\|\tilde{x} - x\| + 1)\,\varphi_{\tilde{B}}(\lambda; \tilde{x}).$$

Proof. To simplify notation, let $y = y_B(\lambda; x)$ and $\tilde{y} = y_{\tilde{B}}(\lambda; \tilde{x})$. Then, there exist unique $v \in B(y)$ and $\tilde{v} \in \tilde{B}(\tilde{y})$ such that

$$\lambda v + y - x = 0, \qquad \lambda\tilde{v} + \tilde{y} - \tilde{x} = 0. \tag{71}$$

Clearly,

$$\varphi_B(\lambda; x) = \lambda^2\|v\|, \qquad \varphi_{\tilde{B}}(\lambda; \tilde{x}) = \lambda^2\|\tilde{v}\|. \tag{72}$$

Let $u := v + G_{\tilde{x}}(y) - G_x(y)$ and note that the fact that $v \in B(y)$ and the first identity (71) imply that

$$u \in B(y) + G_{\tilde{x}}(y) - G_x(y) = T_x(y) + G_{\tilde{x}}(y) - G_x(y) = G_{\tilde{x}}(y) + H(y) = G_{\tilde{x}}(y) = \tilde{B}(y)$$

and

$$\lambda u + y - \tilde{x} = \lambda v + y - x + (\tilde{x} - x) + \lambda(u - v) = (\tilde{x} - x) + \lambda(u - v).$$

Subtracting the second equation in (71) from the last identity, we conclude that

$$\lambda(u - \tilde{v}) + (y - \tilde{y}) = (\tilde{x} - x) + \lambda(u - v).$$

Since $u \in \tilde{B}(y)$ and $\tilde{v} \in \tilde{B}(\tilde{y})$, it follows from the monotonicity of $\tilde{B}$ that

$$\langle u - \tilde{v}, y - \tilde{y}\rangle \geq 0,$$

which together with the previous relation and the triangle inequality for norms imply that

$$\lambda\|u - \tilde{v}\| \leq \|\tilde{x} - x\| + \lambda\|u - v\|,$$

and hence that

$$\lambda\|v - \tilde{v}\| \leq \|\tilde{x} - x\| + 2\lambda\|u - v\|,$$

The latter conclusion together with (72) then imply that

$$|\varphi_B(\lambda; x) - \varphi_{\tilde{B}}(\lambda; \tilde{x})| = \lambda^2\,|\,\|v\| - \|\tilde{v}\|\,| \leq \lambda^2\|v - \tilde{v}\| \leq \lambda\left[\|\tilde{x} - x\| + 2\lambda\|u - v\|\right]. \tag{73}$$

Now, letting $x_p = P_\Omega(x)$ and $\tilde{x}_p = P_\Omega(\tilde{x})$, and using the definition of u, we have

$$\begin{aligned} u - v = G_{\tilde{x}}(y) - G_x(y) &= G(\tilde{x}_p) + G'(\tilde{x}_p)(y - \tilde{x}_p) - [G(x_p) + G'(x_p)(y - x_p)] \\ &= [G(\tilde{x}_p) + G'(\tilde{x}_p)(x_p - \tilde{x}_p) - G(x_p)] + [G'(\tilde{x}_p) - G'(x_p)](y - x_p) \end{aligned}$$

and hence

$$\begin{aligned}\lambda\|u-v\| &\le \lambda\|G(\tilde{x}_p)+G'(\tilde{x}_p)(x_p-\tilde{x}_p)-G(x_p)\|+\lambda\|G'(\tilde{x}_p)-G'(x_p)\|\|y-x_p\| \\ &\le \frac{\lambda L}{2}\|\tilde{x}_p-x_p\|^2+\lambda L\|\tilde{x}_p-x_p\|\|y-\tilde{x}_p\| \le \frac{\lambda L}{2}\|\tilde{x}-x\|^2+L\lambda\|\tilde{x}-x\|\|y-x\| \\ &= \frac{\lambda L}{2}\|\tilde{x}-x\|^2+L\|\tilde{x}-x\|\varphi_B(\lambda;x).\end{aligned}$$

Combining the latter inequality with (73), we then conclude that

$$|\varphi_B(\lambda;x)-\varphi_{\tilde{B}}(\lambda;\tilde{x})| \le \lambda\left[\|\tilde{x}-x\|+L\lambda\|\tilde{x}-x\|^2+2L\varphi_B(\lambda;x)\|\tilde{x}-x\|\right].$$

This inequality and the symmetric one obtained by interchanging x and $\tilde{x}$ in the latter relation then imply (70). □

The following definition extends the notion of a $\hat{\sigma}$-approximate Newton solution to the context of (69).

Definition 7.6. *Given $(\lambda,x)\in\mathbb{R}_{++}\times\mathbb{E}$ and $\hat{\sigma}\ge 0$, the triple $(y,u,\varepsilon)\in\mathbb{E}\times\mathbb{E}\times\mathbb{R}_+$ is called a $\hat{\sigma}$-approximate Newton solution of (69) at (λ,x) if*

$$u\in(G_x+H^{\varepsilon})(y), \qquad \|\lambda u+y-x\|^2+2\lambda\varepsilon\le\hat{\sigma}^2\|y-x\|^2.$$

Note that when $G=\nabla g$ and $H=\partial h$, where g and h are as in Section 6, then a $\hat{\sigma}$-approximate Newton solution according to Definition 6.1 is $\hat{\sigma}$-approximate Newton solution according to Definition 7.6, due to Proposition 2.3(a).

The following result shows that $\hat{\sigma}$-approximate Newton solutions of (69) yield approximate solutions of (69).

Proposition 7.7. *Let $(\lambda,x)\in\mathbb{R}_{++}\times\mathbb{E}$ and a $\hat{\sigma}$-approximate Newton solution (y,u,ε) of (69) at (λ,x) be given, and define $v:=G(y)+u-G_x(y)$. Then,*

$$v\in(G+H^{\varepsilon})(y)\subseteq T^{\varepsilon}(y), \qquad \|\lambda v+y-x\|^2+2\lambda\varepsilon\le\left(\hat{\sigma}+\frac{L\lambda}{2}\|y-x\|\right)^2\|y-x\|^2 \tag{74}$$

and

$$\|v\|\le\frac{1}{\lambda}\left(1+\hat{\sigma}+\frac{L\lambda}{2}\|y-x\|\right)\|y-x\|, \qquad \varepsilon\le\frac{\hat{\sigma}^2}{2\lambda}\|y-x\|^2. \tag{75}$$

Proof. The first inclusion and the inequality in (74) have been established in Lemma 3.2 of [11]. The second inclusion in (74) can be easily proved using Assumptions C.1 and C.2, the definition of T, and b) and c) of Proposition 2.1. Moreover, the inequalities in (75) follow either from (74) or Definition 7.6. □

As a consequence of Proposition 7.7, we can now establish the following result which will be used to obtain the upper endpoint of the initial bracketing interval used in the line search procedure of Subsection 7.3 for computing the stepsize λ_{k+1} satisfying (51).

Lemma 7.8. *Let tolerances $\bar{\rho} > 0$ and $\bar{\varepsilon} > 0$ and scalars $\hat{\sigma} \geq 0$ and $\alpha > 0$ be given. Then, for any scalar*

$$\lambda \geq \max\left\{ \sqrt{\frac{\alpha}{\bar{\rho}}\left(1 + \hat{\sigma} + \frac{L\alpha}{2}\right)}, \left(\frac{\hat{\sigma}^2\alpha^2}{2\bar{\varepsilon}}\right)^{\frac{1}{3}} \right\}, \tag{76}$$

vector $x \in \mathbb{E}$ and $\hat{\sigma}$-approximate Newton solution (y, u, ε) of (69) at (λ, x), one of the following statements hold:

a) either, $\lambda\|y - x\| > \alpha$;

b) or, the vector $v := G(y) - G_x(y) + u$ satisfies

$$v \in (G + H^{\varepsilon})(y), \qquad \|v\| \leq \bar{\rho}, \qquad \varepsilon \leq \bar{\varepsilon}. \tag{77}$$

Proof. To prove the lemma, let λ satisfying (76) be given and assume that a) does not hold, i.e.,

$$\lambda\|y - x\| \leq \alpha. \tag{78}$$

Then, it follows from Proposition 7.7 and relations (76) and (78) that the inclusion in (77) holds and

$$\|v\| \leq \frac{1}{\lambda}\left(1 + \hat{\sigma} + \frac{L\lambda}{2}\|y - x\|\right)\|y - x\| \leq \left(1 + \hat{\sigma} + \frac{L\alpha}{2}\right)\frac{\alpha}{\lambda^2} \leq \bar{\rho},$$

and

$$\varepsilon \leq \frac{\hat{\sigma}^2\|y - x\|^2}{2\lambda} \leq \frac{\hat{\sigma}^2\alpha^2}{2\lambda^3} \leq \bar{\varepsilon}.$$

□

We now make a few remarks about Lemma 7.8. First, note that condition (77) is a natural relaxation of an exact solution of (67), where two levels of relaxations are introduced, namely: the scalar $\varepsilon \geq 0$ in the enlargement of H and the residual v in place of 0 as in (67). Hence, if a triple (y, u, ε) satisfying condition (77) is found for some user-supplied tolerance pair $(\bar{\rho}, \bar{\varepsilon})$, then y can be considered a sufficiently accurate approximate solution of (67) and the pair (v, ε) provides a certificate of such accuracy. Second, when the triple (y, u, ε) fails to satisfy b), then (76) describes how large λ should be chosen so as to guarantee that the quantity $\lambda\|y - x\|$ be larger than a given scalar $\alpha > 0$.

The following result describes the idea for obtaining the lower endpoint of the initial bracket interval for the line search procedure of Subsection 7.3.

Lemma 7.9. *Let $x_-^0 \in \mathbb{E}$, $(\lambda_+^0, x_+^0) \in \mathbb{R}_{++} \times \mathbb{E}$ and a $\hat{\sigma}$-approximate Newton solution $(y_+^0, u_+^0, \varepsilon_+^0)$ of (69) at (λ_+^0, x_+^0) be given. Then, for any scalar α such that*

$$0 < \alpha \leq \lambda_+^0\|y_+^0 - x_+^0\|, \tag{79}$$

scalar $\lambda_-^0 > 0$ such that

$$\lambda_-^0 \leq \frac{\alpha(1 - \hat{\sigma})\lambda_+^0}{(1 + \hat{\sigma})(1 + 2L\theta_+^0)\lambda_+^0\|y_+^0 - x_+^0\| + \theta_+^0 + L(\theta_+^0)^2}, \qquad \theta_+^0 := \lambda_+^0\|x_+^0 - x_-^0\|, \tag{80}$$

and $\hat{\sigma}$-approximate Newton solution $(y_-^0, u_-^0, \varepsilon_-^0)$ of (69) at (x_-^0, λ_-^0), we have

$$(1 + \hat{\sigma})\lambda_-^0 \leq (1 - \hat{\sigma})\lambda_+^0, \tag{81}$$

$$\lambda_-^0\|y_-^0 - x_-^0\| \leq \alpha. \tag{82}$$

Proof. First, note that (81) follows immediately from (79) and (80). Let $B_- := T_{x_-^0}$ and $B_+ = T_{x_+^0}$. Since a $\hat\sigma$-approximate Newton solution of (69) at (λ_-^0, x_-^0) is obviously a $\hat\sigma$-approximate solution of (63) with $B = B_-$, it follows from Lemma 7.3 with $B = B_-$ that

$$\lambda_-^0 \|y_-^0 - x_-^0\| \le \frac{\varphi_{B_-}(\lambda_-^0; x_-^0)}{1-\hat\sigma} \le \frac{\lambda_-^0 \varphi_{B_-}(\lambda_+^0; x_-^0)}{(1-\hat\sigma)\lambda_+^0},$$

where the last inequality is due to (81) and Proposition 7.1(b) with $B = B_-$, $\tilde\lambda = \lambda_-^0$ and $\lambda = \lambda_+^0$. Also, Lemma 7.5 with $\lambda = \lambda_+^0$, $x = x_-^0$ and $\tilde x = x_+^0$, and the definition of θ_+^0 in (80) imply that

$$\begin{aligned}\varphi_{B_-}(\lambda_+^0; x_-^0) &\le \left(1 + 2L\theta_+^0\right) \varphi_{B_+}(\lambda_+^0; x_+^0) + \theta_+^0 + L(\theta_+^0)^2 \\ &\le \left(1 + 2L\theta_+^0\right)(1+\hat\sigma)\lambda_+^0 \|y_+^0 - x_+^0\| + \theta_+^0 + L(\theta_+^0)^2 \le \frac{\alpha(1-\hat\sigma)\lambda_+^0}{\lambda_-^0},\end{aligned}$$

where the last inequality follows from the definition of λ_-^0. Combining the above two inequalities, we then conclude that (82) holds. □

7.3 Line search problem and procedure

In this subsection, we describe a line search procedure whose goal is to implement step 2 of the A-NPE method. For the sake of generality and simplicity of notation, it will be described in a setting slightly more general than that of (37), or equivalently (38). Namely, we will consider the setting of the inclusion problem (67) with G and H satisfying conditions C.1 to C.3.

Throughout this subsection, we assume that we have at our disposal the following Newton Black-Box.

Newton Black-Box: For any given $\sigma \ge 0$ and $(\lambda, x) \in \mathbb{R}_{++} \times \mathbb{E}$, it computes a σ-approximate Newton solution (y, u, ε) of (69) at (λ, x).

With the aid of the above black-box, the goal of the line search procedure described in this subsection is to solve the following line search problem.

Line Search Problem: Given tolerances $\hat\sigma \ge 0$, $\bar\rho > 0$ and $\bar\varepsilon > 0$, bounds $\alpha_+ > \alpha_- > 0$ and a continuous curve $x : [0, \infty) \to \mathbb{E}$ satisfying the property that, for some constant $M_0 \ge 0$ and $M_1 \ge 0$,

$$\|x(s) - x(t)\| \le \min\left\{\frac{M_0}{t}(s-t)\,,\, M_1\|x(s) - x(0)\|\right\}, \qquad \forall s \ge t > 0, \tag{83}$$

the problem is to find a stepsize $\lambda > 0$ and a $\hat\sigma$-approximate Newton solution $(y_\lambda, u_\lambda, \varepsilon_\lambda)$ of (69) at $(\lambda, x(\lambda))$ such that

a) either $\lambda\|y_\lambda - x(\lambda)\| \in [\alpha_-, \alpha_+]$, or;

b) the triple $(y, v, \varepsilon) = (y_\lambda, v_\lambda, \varepsilon_\lambda)$ satisfies (77), where

$$v_\lambda := G(y_\lambda) + u_\lambda - G_{x(\lambda)}(y_\lambda).$$

We now state the procedure for solving the above line-search problem.

Bracketing/Bisection Procedure:

Input: Curve $x : [0, \infty) \to \mathbb{E}$ satisfying (83), tolerances $\hat{\sigma} \geq 0$, $\bar{\rho} > 0$ and $\bar{\varepsilon} > 0$, and bounds $\alpha_+ > \alpha_- > 0$ satisfying

$$\alpha_-(1 + \hat{\sigma}) < \alpha_+(1 - \hat{\sigma}); \tag{84}$$

Output: stepsize $\lambda > 0$ and a $\hat{\sigma}$-approximate Newton solution $(y_\lambda, u_\lambda, \varepsilon_\lambda$ of (69) at $(\lambda, x(\lambda))$ such that either a) or b) above holds.

1) (**Bracketing stage)** compute

$$\lambda_+^0 := \max\left\{\sqrt{\frac{\alpha_+}{\bar{\rho}}\left(1 + \hat{\sigma} + \frac{L\alpha_+}{2}\right)}, \left(\frac{\hat{\sigma}^2\alpha_+^2}{2\bar{\varepsilon}}\right)^{\frac{1}{3}}\right\}, \tag{85}$$

and set $x_+^0 = x(\lambda_+^0)$; use the Newton Black-Box to compute a $\hat{\sigma}$-approximate Newton solution $(y_+^0, u_+^0, \varepsilon_+^0)$ of (69) at (λ_+^0, x_+^0), and set

$$v_+^0 = G(y_+^0) - G_{x_+^0}(y_+^0) + u_+^0;$$

1.a) if (v_+^0, ε_+^0) satisfies $\|v_+^0\| \leq \bar{\rho}$ and $\varepsilon_+^0 \leq \bar{\varepsilon}$, then output $\lambda = \lambda_+^0$ and $(y_\lambda, u_\lambda, \varepsilon_\lambda = (y_+^0, u_+^0, \varepsilon_+^0)$, and **stop**;

1.b) otherwise, compute $\gamma_0 := M_1\lambda_+^0\|x_+^0 - x(0)\|$ and

$$\lambda_-^0 := \frac{(1 - \hat{\sigma})\alpha_-\lambda_+^0}{(1 + \hat{\sigma})(1 + 2L\gamma_0)\lambda_+^0\|y_+^0 - x_+^0\| + \gamma_0 + L(\gamma_0)^2}, \qquad x_-^0 := x(\lambda_-^0), \tag{86}$$

and use the Newton Black-Box to compute a $\hat{\sigma}$-approximate Newton solution $(y_-^0, u_-^0, \varepsilon_-^0)$ of (69) at (λ_-^0, x_-^0).

2) (**Bisection stage)** set $\lambda_- = \lambda_-^0$ and $\lambda_+ = \lambda_+^0$;

2.a) set $\lambda = (\lambda_- + \lambda_+)/2$, use the Newton Black-Box to compute a $\hat{\sigma}$-approximate Newton solution $(y_\lambda, u_\lambda, \varepsilon_\lambda)$ of (69) at $(\lambda, x(\lambda))$, and set

$$v_\lambda := G(y_\lambda) + u_\lambda - G_{x(\lambda)}(y_\lambda);$$

if $(v_\lambda, \varepsilon_\lambda)$ satisfies $\|v_\lambda\| \leq \bar{\rho}$ and $\varepsilon_\lambda \leq \bar{\varepsilon}$, then output λ and $(y_\lambda, u_\lambda, \varepsilon_\lambda)$, and **stop**;

2.b) if $\lambda\|y_\lambda - x(\lambda)\| \in [\alpha_-, \alpha_+]$, then output λ and (y, u, ε), and **stop**;

2.c) if $\lambda\|y_\lambda - x(\lambda)\| > \alpha_+$, then set $\lambda_+ \leftarrow \lambda$; else set $\lambda_- \leftarrow \lambda$;

2.d) go to step 2.a.

end

Proposition 7.10. *If the Bracketing/Bisection procedure does not stop during the Bracketing stage, then at the end of the Bracketing stage, the following conditions hold:*

$$\lambda_-^0\|y_-^0 - x_-^0\| \leq \alpha_-, \qquad \lambda_+^0\|y_+^0 - x_+^0\| \geq \alpha_+.$$

Proof. Assume that Bracketing/Bisection procedure does not stop during the Bracketing stage. In view of Lemma 7.8 with $\alpha = \alpha_+$, $\lambda = \lambda_+^0$, $x = x_+^0$, $(y, u, \varepsilon) = (y_+^0, u_+^0, \varepsilon_+^0)$ and $v = v_+^0$, we conclude that $\lambda_+^0 \|y_+^0 - x_+^0\| > \alpha_+$. Observe that (84) implies that $\alpha_- < \alpha_+$, and hence that $\lambda_+^0 \|y_+^0 - x_+^0\| > \alpha_-$. Also, the assumption on the curve $x : [0, \infty) \to \mathbb{E}$ and the definition of γ_0 in step 1.b) implies that

$$\lambda_+^0 \|x_+^0 - x_-^0\| = \lambda_+^0 \|x(\lambda_+^0) - x(\lambda_-^0)\| \leq \lambda_+^0 \|x(\lambda_+^0) - x(0)\| = M_1 \lambda_+^0 \|x_+^0 - x(0)\| = \gamma_0,$$

and hence that λ_-^0 given by (86) satisfies the inequality in (80). Hence, using Lemma 7.9 with $\alpha = \alpha_-$, we conclude that $\lambda_-^0 \|y_-^0 - x_-^0\| \leq \alpha_-$. □

The proof of the following result can be found in Proposition 4.8 of [10].

Proposition 7.11. *Assume that $x_* \in T^{-1}(0) = (G + H)^{-1}(0)$ and let $\bar{x}, x \in \mathbb{E}$ be given. Then,*

$$\|x - (I + \lambda T_{\bar{x}})^{-1}(x)\| \leq \|x - x_*\| + \lambda L \|\bar{x} - x_*\|^2.$$

As a consequence, for every $x \in \mathbb{E}$, $x_* \in T^{-1}(0)$ and $\lambda > 0$, there holds

$$\varphi_{T_x}(\lambda; x) \leq \lambda \|x - x_*\| + \lambda^2 L \|x - x_*\|^2.$$

We are now ready to give the complexity of the Bracketing/Bisection procedure.

Lemma 7.12. *The Bracketing/Bisection procedure makes at most*

$$3 + \log \left[C_0^2 M_0 \lambda_+^0 \left(\frac{1 + L M_0 \lambda_+^0 + 2 \left(L + \frac{1}{M_0 \lambda_+^0} \right) (1 + \hat{\sigma}) \alpha_-}{(1 - \hat{\sigma}) \alpha_+ - (1 + \hat{\sigma}) \alpha_-} \right) \right]$$

Black-Box calls, where λ_+^0 is as in (85)*,*

$$C_0 := \frac{(1 + \hat{\sigma})(1 + 2L\gamma_0) \left[\lambda_+^0 d(x_+^0) + (\lambda_+^0)^2 L d(x_+^0)^2 \right] + \gamma_0 + L(\gamma_0)^2}{(1 - \hat{\sigma})^2 \alpha_-},$$

γ_0 is defined in step 1.b) of the procedure, and $d(x_+^0)$ denotes the distance of x_+^0 to $T^{-1}(0)$.

Proof. First observe that the Bracketing/Bisection procedure performs at most two Black-Box call during the Bracketing stage. Assume that the procedure enters the Bisection stage and let us estimate the number of Black-Box calls within this stage. Since $\lambda = (\lambda_- + \lambda_+)/2$, it follows that, after j bisection iterations, the scalars λ_- and λ_+ computed at step 2.c) satisfy

$$\lambda_+ - \lambda_- = \frac{1}{2^j} (\lambda_+^0 - \lambda_-^0) \leq \frac{1}{2^j} \lambda_+^0, \tag{87}$$

and hence

$$j \leq \log \left(\frac{\lambda_+^0}{\lambda_+ - \lambda_-} \right). \tag{88}$$

Assume now that the method does not stop at the j-th bisection iteration. Then, the values of λ_- and λ_+ at step 2.c) of this iteration satisfy

$$\lambda_+ \|y_{\lambda_+} - x(\lambda_+)\| > \alpha_+, \qquad \lambda_- \|y_{\lambda_-} - x(\lambda_-)\| < \alpha_-.$$

and let $x_+ := x(\lambda_+)$, $x_- := x(\lambda_-)$, $B_+ := T_{x_+}$ and $B_- := T_{x_-}$. Hence, applying Lemma 7.3 twice, one time with $B = B_+$, $x = x_+$ and $(y, u, \varepsilon) = (y_{\lambda_+}, u_{\lambda_+}, \varepsilon_{\lambda_+})$, and the other with $B = B_-$, $x = x_-$ and $(y, u, \varepsilon) = (y_{\lambda_-}, u_{\lambda_-}, \varepsilon_{\lambda_-})$, we conclude that

$$\varphi_+ := \varphi_{B_+}(\lambda_+; x_+) > (1-\hat{\sigma})\alpha_+, \qquad \varphi_- := \varphi_{B_-}(\lambda_-; x_-) < (1+\hat{\sigma})\alpha_-, \tag{89}$$

On the other hand, it follows from Proposition 7.1(b) with $B = B_+$, $x = x_+$, $\tilde{\lambda} = \lambda_-$ and $\lambda = \lambda_+$, and Lemma 7.5 with $\lambda = \lambda_-$, $x = x_+$ and $\tilde{x} = x_-$ that

$$\begin{aligned}\varphi_+ = \varphi_{B_+}(\lambda_+; x_+) &\le \left(\frac{\lambda_+}{\lambda_-}\right)^2 \varphi_{B_+}(\lambda_-; x_+) \\ &\le \left(\frac{\lambda_+}{\lambda_-}\right)^2 \left[\theta + L\theta^2 + (1+2L\theta)\varphi_-)\right],\end{aligned} \tag{90}$$

where

$$\theta := \lambda_- \|x_+ - x_-\| \le M_0(\lambda_+ - \lambda_-), \tag{91}$$

in view of the property assumed for the curve $x(\cdot)$. Hence, we conclude that

$$\begin{aligned}\varphi_+ - \varphi_- &\le \left(\frac{\lambda_+}{\lambda_-}\right)^2 \theta[1 + L\theta + 2L\varphi_-] + \left[\left(\frac{\lambda_+}{\lambda_-}\right)^2 - 1\right]\varphi_- \\ &\le \left(\frac{\lambda_+^0}{\lambda_-^0}\right)^2 M_0\left(1 + LM_0\lambda_+^0 + 2L\varphi_-\right)(\lambda_+ - \lambda_-) + \frac{(\lambda_+ + \lambda_-)}{(\lambda_-)^2}\varphi_-(\lambda_+ - \lambda_-) \\ &\le \left(\frac{\lambda_+^0}{\lambda_-^0}\right)^2 M_0\left(1 + LM_0\lambda_+^0 + 2L\varphi_-\right)(\lambda_+ - \lambda_-) + 2\frac{\lambda_+^0}{(\lambda_-^0)^2}\varphi_-(\lambda_+ - \lambda_-) \\ &= (\lambda_+ - \lambda_-)\left(\frac{\lambda_+^0}{\lambda_-^0}\right)^2 M_0\left[1 + LM_0\lambda_+^0 + 2\left(L + \frac{1}{M_0\lambda_+^0}\right)\varphi_-\right].\end{aligned}$$

Combining the latter inequality with (89), we then conclude that

$$\frac{1}{\lambda_+ - \lambda_-} \le \left(\frac{\lambda_+^0}{\lambda_-^0}\right)^2 M_0\left(\frac{1 + LM_0\lambda_+^0 + 2\left(L + \frac{1}{M_0\lambda_+^0}\right)(1+\hat{\sigma})\alpha_-}{(1-\hat{\sigma})\alpha_+ - (1+\hat{\sigma})\alpha_-}\right). \tag{92}$$

We will now estimate the ratio λ_+^0/λ_-^0. First note that Lemma 7.3 with $\lambda = \lambda_+^0$, $x = x_+^0$, $B = T_{x_+^0}$ and $(y, u, \varepsilon) = (y_+^0, u_+^0, \varepsilon_+^0)$ and Proposition 7.11 with $\lambda = \lambda_+^0$ and $x = x_+^0$ imply that

$$\lambda_+^0 \|y_+^0 - x_+^0\| \le \frac{\varphi_{T_{x_+^0}}(\lambda_+^0; x_+^0)}{1-\hat{\sigma}} \le \frac{\lambda_+^0 d(x_+^0) + (\lambda_+^0)^2 L d(x_+^0)^2}{1-\hat{\sigma}}.$$

The latter inequality together with relations (85) and (80) then imply that

$$\begin{aligned}\frac{\lambda_+^0}{\lambda_-^0} &= \frac{(1+\hat{\sigma})(1+2L\gamma_0)\lambda_+^0\|y_+^0 - x_+^0\| + \gamma_0 + L(\gamma_0)^2}{(1-\hat{\sigma})\alpha_-} \\ &\le \frac{(1+\hat{\sigma})(1+2L\gamma_0)\left[\lambda_+^0 d(x_+^0) + (\lambda_+^0)^2 L d(x_+^0)^2\right] + \gamma_0 + L(\gamma_0)^2}{(1-\hat{\sigma})^2\alpha_-}\end{aligned}$$

The result now follows from the above inequality and relations (88) and (92). □

7.4 Complexity of implementing step 2 of the A-NPE method

In this subsection, we study a special case of the line search procedure introduced in the previous subsection whose goal is to implement step 2 of the A-NPE method. We will also derive its computational complexity, and as a by-product the overall complexity of the A-NPE method, in terms of number of calls to a given optimization Newton Black-Box.

Throughout this subsection without further mentioning, we consider only the version of the A-NPE method in which $\sigma < 1$ and $y_{k+1} = \tilde{y}_{k+1}$ for every $k \geq 0$, and hence the sequence $\{y_k\}$ is bounded due to Theorem 3.10.

We assume throughout this subsection that an optimization Newton Black-Box for (37) is available which, given $\hat{\sigma}$ and $(\lambda, x) \in \mathbb{R}_{++} \times \mathbb{E}$, finds a $\hat{\sigma}$-approximate Newton optimal solution for (37) at (λ, x), i.e., a triple $(y, u, \varepsilon) \in \mathbb{E} \times \mathbb{E} \times \mathbb{R}_+$ such that

$$u \in (\nabla g_x + \partial_\varepsilon h)(y), \qquad \|\lambda u + y - x\|^2 + 2\lambda\varepsilon \leq \hat{\sigma}^2 \|y - x\|^2. \tag{93}$$

Clearly, by Proposition 2.3(a), it follows that a $\hat{\sigma}$-approximate Newton optimal solution for (47) is a $\hat{\sigma}$-approximate Newton solution for (50) in the sense of Definition 7.6, and hence an optimization Newton Black-Box for (47) is a Newton Black-Box for (50) in the sense of Subsection 7.3. Note also that the triple $(\tilde{y}_{k+1}, u_{k+1}, \varepsilon_{k+1})$ as in Step 2 of the A-NPE method is a $\hat{\sigma}$-approximate Newton solution for (47) at $(\lambda_{k+1}, \tilde{x}_k)$.

Consider the curve

$$\tilde{x}_k(\lambda) = \frac{A_k}{A_k + a_{k+1}(\lambda)} y_k + \frac{a_{k+1}(\lambda)}{A_k + a_{k+1}(\lambda)} x_k. \tag{94}$$

where

$$a_{k+1}(\lambda) = \frac{\lambda + \sqrt{\lambda^2 + 4A_k\lambda}}{2}, \tag{95}$$

In order to compute the stepsize λ_{k+1}, and the triple $(\tilde{y}_{k+1}, u_{k+1}, \varepsilon_{k+1})$ as in Step 2 of the A-NPE method, the strategy used is based on invoking the line search procedure described in the previous subsection to look for a stepsize $\lambda_{k+1} > 0$ and a $\hat{\sigma}$-approximate Newton optimal solution $(\tilde{y}_{k+1}, u_{k+1}, \varepsilon_{k+1})$ at $(\lambda_{k+1}, \tilde{x}_k(\lambda_{k+1}))$ such that

$$\frac{2\sigma_\ell}{L} \leq \lambda_{k+1} \|\tilde{y}_{k+1} - \tilde{x}_k(\lambda_{k+1})\| \leq \frac{2\sigma_u}{L}. \tag{96}$$

and then setting $(a_{k+1}, \tilde{x}_k) = (a_{k+1}(\lambda_{k+1}), \tilde{x}_k(\lambda_{k+1}))$. More precisely, with the aid of the optimization Newton Black-Box at our disposal, we can use the Bracketing/Bisection Procedure of the previous subsection with tolerances $\hat{\sigma} \geq 0$, $\bar{\rho} > 0$ and $\bar{\varepsilon} > 0$, curve $x(\cdot) = \tilde{x}_k(\cdot)$, and bounds $\alpha_- = 2\sigma_\ell/L$, $\alpha_+ = 2\sigma_u/L$ to compute a stepsize $\lambda_{k+1} > 0$ and a $\hat{\sigma}$-approximate Newton optimal solution $(\tilde{y}_{k+1}, u_{k+1}, \varepsilon_{k+1})$ at $(\lambda_{k+1}, \tilde{x}_k(\lambda_{k+1}))$ such that

a) either, the residual pair $(v_{k+1}, \varepsilon_{k+1})$ (see relation (58)) with v_{k+1} given by (55) satisfy $\|v_{k+1}\| \leq \bar{\rho}$ and $\varepsilon_{k+1} \leq \bar{\varepsilon}$;

b) or relation (96) holds, and as a consequence, λ_{k+1} and $(\tilde{y}_{k+1}, u_{k+1}, \varepsilon_{k+1})$ fulfil the conditions of step 2 of the A-NPE method;

The following result shows that the curve (94) satisfies property (83) as required by the Bracketing/Bisection procedure.

Lemma 7.13. *The curve $\tilde{x}_k(\cdot)$, where $\tilde{x}_k(\cdot)$ is given by* (94)*, satisfies* (83) *with*

$$M_0 := 2\left(\frac{1}{\sqrt{1-\sigma^2}}+1\right)d_0, \qquad M_1 = 1. \tag{97}$$

Proof. To show the lemma, let $s > t > 0$ be given. Define

$$\tau(\lambda) = \frac{a_{k+1}(\lambda)}{A_k + a_{k+1}(\lambda)}, \qquad \forall \lambda > 0,$$

where $a_{k+1}(\lambda)$ is given by (95), and note that $\tau(\cdot)$ is an increasing function. In view of (94), we have

$$\tilde{x}_k(\lambda) = y_k + \tau(\lambda)(x_k - y_k), \qquad \forall \lambda > 0.$$

and hence

$$\|\tilde{x}_k(s) - \tilde{x}_k(t)\| = (\tau(s) - \tau(t))\|x_k - y_k\|. \tag{98}$$

Since $\tau(\cdot)$ is increasing and $\tau(0) = 0$, we conclude

$$\|\tilde{x}_k(s) - \tilde{x}_k(t)\| \le \|\tilde{x}_k(s) - \tilde{x}_k(0)\|. \tag{99}$$

Clearly (95) implies that $a = a_{k+1}(\lambda)$ satisfies the second-order equation $a^2 - \lambda_{k+1}a - \lambda_{k+1}A_k = 0$, and hence

$$\tau(\lambda) = \frac{a_{k+1}(\lambda)}{A_k + a_{k+1}(\lambda)} = \frac{\lambda}{a_{k+1}(\lambda)} = \frac{2}{1+\sqrt{1+4A_k\lambda^{-1}}},$$

where the last equality is due to (95). Differentiating the last expression for $\tau(\lambda)$, we conclude that

$$\dot{\tau}(\lambda) = \frac{4A_k\lambda^{-2}}{\left(1+\sqrt{1+4A_k\lambda^{-1}}\right)^2\sqrt{1+4A_k\lambda^{-1}}} \le \frac{1}{\lambda}, \qquad \forall \lambda > 0.$$

Hence, by the mean value theorem, we have

$$\tau(s) - \tau(t) = \dot{\tau}(\xi)(s-t) \le \frac{1}{\xi}(s-t) \le \frac{1}{t}(s-t),$$

for some $\xi \in (t, s)$. Combining the last inequality with (98), we then obtain

$$\|\tilde{x}_k(s) - \tilde{x}_k(t)\| = \frac{\|x_k - y_k\|}{t}(s-t).$$

Now, Theorem 3.6, Theorem 3.10, and the triangle inequality for norms, imply that

$$\|x_k - y_k\| \le 2\left(\frac{1}{\sqrt{1-\sigma^2}}+1\right)d_0.$$

The latter two inequalities together with (99) then imply that the curve $\tilde{x}(\cdot) = \tilde{x}_k(\cdot)$ satisfies (83) with M_0 and M_1 given by (97). □

Viewing σ_ℓ, σ_u, $\hat{\sigma}$ as universal constants, the following result establishes the complexity of the line search procedure when used to implement step 2 of the A-NPE method.

Theorem 7.14. *Let $\bar{\rho} > 0$ and $\bar{\varepsilon} > 0$ be given and suppose that the Bracketing/Bisection procedure is used to implement step 2 of the A-NPE method as explained in the paragraphs preceding Lemma 7.13. Then, the procedure performs at most*

$$\mathcal{O}\left(\max\left\{\log \bar{\varepsilon}^{-1}, \log \bar{\rho}^{-1}, \log d_0, \log L, \log L^{-1}\right\}\right). \tag{100}$$

optimization Newton Black-Box calls to compute a stepsize $\lambda_{k+1} > 0$ and a $\hat{\sigma}$-approximate Newton optimal solution $(\tilde{y}_{k+1}, u_{k+1}, \varepsilon_{k+1})$ at $(\lambda_{k+1}, \tilde{x}_k(\lambda_{k+1}))$ such that one of the statements a) or b) preceding Lemma 7.13 holds.

Proof. Viewing σ_ℓ, σ_u, $\hat{\sigma}$ as universal constants, it follows from the definition of γ_0 in step 1.b of Bracketing/Bisection procedure and Lemmas 7.12 and 7.13 that the latter procedure when used to implement step 2 of the A-NPE method will find a stepsize $\lambda_{k+1} > 0$ and a $\hat{\sigma}$-approximate Newton optimal solution $(\tilde{y}_{k+1}, u_{k+1}, \varepsilon_{k+1})$ at $(\lambda_{k+1}, \tilde{x}_k(\lambda_{k+1}))$ such that either statements a) or b) holds in at most

$$\mathcal{O}\left(\max\left\{\log \lambda_+^0\,,\, \log d_0\,,\, \log L\,,\, \log d(\tilde{x}_k(\lambda_+^0))\,,\, \log\left\|\tilde{x}_k(\lambda_+^0) - y_k\right\|\right\}\right).$$

optimization Newton Black-Box calls, where $d(\tilde{x}_k(\lambda_+^0))$ denotes the distance of $\tilde{x}_k(\lambda_+^0)$ to X_*. Now, using (85), the fact that $\alpha_+ = 2\sigma_u/L$, Theorems 3.6 and 3.10 and the triangle inequality for norms, we easily see that the above bound is majorized by (100). □

Assuming that an upper bound on d_0 is known, the following result describes the computational complexity (in terms of number of optimization Newton Black-Box calls) for the A-NPE method to find an iterate y_k such that $f(y_k) - f_* \leq \delta$, for some given tolerance $\delta > 0$.

Theorem 7.15. *Assume that an upper bound D_0 on the distance d_0 from x_0 to X_* is known and let tolerance $\delta > 0$ be given. Consider the A-NPE method with step 2 implemented as explained in this subsection with tolerances $\bar{\rho}$ and $\bar{\varepsilon}$ chosen as $\bar{\rho} = \delta/(2D_0)$ and $\bar{\varepsilon} = \delta/2$. Then, an iterate y_k satisfying $f(y_k) - f_* \leq \delta$ will be found in no more than*

$$\mathcal{O}\left(\left(\frac{L_1 d_0^3}{\delta}\right)^{2/7}\right)$$

iterations of this method, making no more than

$$\mathcal{O}\left(\left(\frac{L_1 d_0^3}{\delta}\right)^{2/7} \max\left\{\log \delta^{-1},\, \log L,\, \log L^{-1},\, \log D_0\right\}\right)$$

calls to the optimization Newton Black-Box.

Proof. This result follows immediately from Theorems 6.4(a) and 7.14. □

The following result gives the complexity of computing of an approximate solution of (38).

Theorem 7.16. *Consider the A-NPE method with step 2 implemented as explained in this subsection with given tolerances $\bar{\rho} > 0$ and $\bar{\varepsilon} > 0$. Then, the A-NPE method finds a triple $(\tilde{y}_k, v_k, \varepsilon_k)$ satisfying*

$$v_k \in \nabla g(\tilde{y}_k) + \partial_{\varepsilon_k} h(\tilde{y}_k), \qquad \|v_k\| \leq \bar{\rho}, \qquad \varepsilon_k \leq \bar{\varepsilon}$$

in at most

$$\mathcal{O}\left(d_0^{2/3}\max\left\{\left(\frac{L_1}{\bar{\rho}}\right)^{1/3},\left(\frac{L_1}{\bar{\varepsilon}}\right)^{2/9}\right\}\right)$$

iterations of this method, making no more than

$$\mathcal{O}\left(d_0^{2/3}\max\left\{\log\bar{\rho}^{-1},\log\bar{\varepsilon}^{-1},\log L,\log L^{-1},\log d_0\right\}\max\left\{\left(\frac{L_1}{\bar{\rho}}\right)^{1/3},\left(\frac{L_1}{\bar{\varepsilon}}\right)^{2/9}\right\}\right)$$

calls to the optimization Newton Black-Box.

Proof. This result follows immediately from Theorems 6.4(b) and 7.14. □

8 Concluding remarks

In this section, we discuss the relationship between the A-HPE framework and the accelerated inexact proximal point method studied in [4, 17]. First, although they are presented with different notation, it can be shown that the exact method in [4] and the exact case of the A-HPE framework, namely the one in which $\sigma = 0$ in (5), coincide. The other remarks below concern the inexact case. Second, instead of using the relative error condition (5), they assume that one of the (or both) residuals $r_{k+1} := \lambda_{k+1}v_{k+1} + \tilde{y}_{k+1} - \tilde{x}_k$ and ε_{k+1} are $\mathcal{O}(1/k^p)$, for some scalar $p > 0$. Hence, their method is based on an absolute error asymptotic condition rather than a relative error condition such as the one, namely (5), used by the A-HPE Framework. Third, aside from the use of different error criteria, all steps of the A-HPE and the one in [4, 17] are the same with the exception of the update formula (9). More specifically, instead of (9), they use the formula

$$x_{k+1} = x_k - \frac{a_{k+1}}{\lambda_{k+1}}(\tilde{y}_{k+1} - \tilde{x}_k),$$

and also assume that $\tilde{y}_{k+1} = y_{k+1}$. Clearly, when $r_{k+1} \neq 0$, the two aforementioned formulae, and hence the respective methods, differ. We believe that the use of the relative error condition and the update formula (9) play an important role in making the A-HPE Framework a powerful tool in the design and/or analysis of accelerated methods for convex optimization, as illustrated by the discussion in Sections 5 and 6.

A Auxiliary technical results

First we state a technical result which proof is simple, and hence is omitted.

Lemma A.1. *If $C > 0$, $t_1, \ldots, t_k > 0$ and $\alpha_1, \ldots, \alpha_k > 0$ are such that*

$$\sum_{j=1}^{k}\frac{\alpha_j}{t_j^6} \leq C,$$

then

$$\sum_{j=1}^{k} t_j \geq \frac{1}{C^{1/6}}\left(\sum_{j=1}^{k}\alpha_j^{1/7}\right)^{7/6}.$$

Lemma A.2. *For any $C \geq 0$, $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ and $\beta_1, \ldots, \beta_k > 0$, there holds*

$$\max \left\{ \sum_{j=0}^{k} \alpha_j s_j : \sum_{j=0}^{k} \beta_j s_j^2 \leq C \right\} = \sqrt{C \sum_{j=0}^{k} \frac{\alpha_j^2}{\beta_j}}.$$

Proof. The lemma holds trivially when $C = 0$. Suppose then that $C > 0$. Consider the maximization problem and associated Lagrangean

$$\max \left\{ \sum_{j=0}^{k} \alpha_j s_j : \sum_{j=0}^{k} \beta_j s_j^2 \leq C \right\}, \qquad L(s, \mu) = \sum_{j=0}^{n} \alpha_j s_j + \frac{\mu}{2} \left(C - \sum_{j=0}^{k} \beta_j s_j^2 \right).$$

Since the feasible set of this problem is a compact convex set and the objective function is linear, there exists a solution to this problem, say s^*, in the boundary of the feasible set, that is

$$\sum_{j=0}^{k} \beta_j (s_j^*)^2 = C. \tag{101}$$

Moreover, since 0 is an interior point of the feasible set, there exists $\mu^* \geq 0$ such that

$$\nabla_s L(s^*, \mu^*) = 0,$$

whence

$$\mu^* \beta_j s_j^* = \alpha_j, \qquad j = 0, \ldots, n. \tag{102}$$

If $\mu^* = 0$, then $\alpha_1 = \cdots = \alpha_n = 0$, and the lemma holds trivially. So, assume that $\mu^* > 0$. Multiplying each of the $n + 1$ equalities in (102) by the corresponding s_j^* and adding the resulting equalities, we conclude that

$$\mu^* \sum_{j=0}^{n} \beta_j (s_j^*)^2 = \sum_{j=0}^{n} \alpha_j s_j^*,$$

which, combined with (101) yields

$$\mu^* C = \sum_{j=0}^{n} \alpha_j s_j^*.$$

Multiplying each of the $n+1$ equalities in (102) by the corresponding α_j/β_j and adding the resulting equalities, we conclude that

$$\mu^* \sum_{j=0}^{n} \alpha_j s_j^* = \sum_{j=0}^{n} \alpha_j^2/\beta_j,$$

To end the proof, multiply the above equality by C and use the previous equality. □

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[2] Regina S. Burachik, Alfredo N. Iusem, and B. F. Svaiter. Enlargement of monotone operators with applications to variational inequalities. *Set-Valued Anal.*, 5(2):159–180, 1997.

[3] Regina Sandra Burachik and B. F. Svaiter. ϵ-enlargements of maximal monotone operators in Banach spaces. *Set-Valued Anal.*, 7(2):117–132, 1999.

[4] Osman Güler. New proximal point algorithms for convex minimization. *SIAM J. Optim.*, 2(4):649–664, 1992.

[5] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms. II*, volume 306 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. Advanced theory and bundle methods.

[6] B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle*, 4(Ser. R-3):154–158, 1970.

[7] J.-E. Martínez-Legaz and B. F. Svaiter. Monotone operators representable by l.s.c. convex functions. *Set-Valued Anal.*, 13(1):21–46, 2005.

[8] R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle point and convex optimization problems. *SIAM Journal on Optimization*, 21:1688–1720, 2010.

[9] R. D. C. Monteiro and B. F. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating minimization augmented lagrangian method. Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA, August 2010.

[10] R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20:2755–2787, 2010.

[11] R. D. C. Monteiro and B. F. Svaiter. Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems. Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA, April 2011.

[12] Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[13] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Math. Program.*, 112(1, Ser. B):159–181, 2008.

[14] R. T. Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific J. Math.*, 33:209–216, 1970.

[15] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.*, 1(2):97–116, 1976.

[16] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14(5):877–898, 1976.

[17] S. Salzo and S. Villa. Inexact and accelerated proximal point algorithms. *J. Convex Anal.*, 19(4), 2012.

[18] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Anal.*, 7(4):323–345, 1999.

[19] M. V. Solodov and B. F. Svaiter. A hybrid projection-proximal point algorithm. *J. Convex Anal.*, 6(1):59–70, 1999.

[20] M. V. Solodov and B. F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Math. Oper. Res.*, 25(2):214–230, 2000.

[21] M. V. Solodov and B. F. Svaiter. A unified framework for some inexact proximal point algorithms. *Numer. Funct. Anal. Optim.*, 22(7-8):1013–1035, 2001.

[22] P Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, Department of Mathematics, University of Washington, Washington, 98195, USA, 2008.