Variants of the A-HPE and large-step A-HPE algorithms for strongly convex problems with applications to accelerated high-order tensor methods

M. Marques Alves *

February 3, 2021

Abstract

For solving strongly convex optimization problems, we propose and study the global convergence of variants of the A-HPE and large-step A-HPE algorithms of Monteiro and Svaiter [18]. We prove *linear* and the *superlinear* $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ global rates for the proposed variants of the A-HPE and large-step A-HPE methods, respectively. The parameter $p \geq 2$ appears in the (high-order) large-step condition of the new large-step A-HPE algorithm. We apply our results to high-order tensor methods, obtaning a new inexact (relative-error) tensor method for (smooth) strongly convex optimization with iteration-complexity $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$. In particular, for $p = 2$, we obtain an inexact Newton-proximal algorithm with fast global $\mathcal{O}\left(k^{-k/3}\right)$ convergence rate.

2000 Mathematics Subject Classification: 90C60, 90C25, 47H05, 65K10.

Key words: Convex optimization, strongly convex, accelerated methods, proximal-point algorithm, large-step, high-order tensor methods, superlinear convergence, proximal-Newton method.

1 Introduction

The *proximal-point method* [14, 25] is one of the most popular algorithms for solving nonsmooth convex optimization problems. For the general problem of minimizing a convex function $h(\cdot)$, its *exact* version can be described by the iteration

$$x^{k+1} = \operatorname{Arg}\min_x \left\{ h(x) + \frac{1}{2\lambda}\|x - x^k\|^2 \right\}, \qquad k \geq 0, \tag{1}$$

where $\lambda = \lambda_{k+1} > 0$ and x^k is the current iterate. Motivated by the fact that in many cases the computation of x^{k+1} is numerically expensive, several authors have proposed *inexact* versions of (1). Among them, inexact proximal-point methods based on *relative-error* criterion for the subproblems are currently quite popular. For the more abstract setting of solving inclusions for maximal monotone operators, this approach was initially developed by Solodov and Svaiter (see, e.g., [26, 27, 28, 29]),

*Departamento de Matemática, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 88040-900 (maicon.alves@ufsc.br). The work of this author was partially supported by CNPq grants no. 304692/2017-4.

subsequently studied, from the viewpoint of computational complexity, by Monteiro and Svaiter (see, e.g., [15, 16, 17, 18]) and has gained a lot of attention by different authors and research groups (see, e.g., [4, 5, 8, 11, 13]) with many applications in optimization algorithms and related topics such as variational inequalities, saddle-point problems, etc.

The starting point of this contribution is [18], where the relative-error inexact hybrid proximal extragradient (HPE) method [16, 26] was accelerated for convex optimization, by using Nesterov's acceleration [19]. The resulting accelerated HPE-type algorithms, called A-HPE and large-step A-HPE, were applied to first- and second-order optimization, with iteration-complexities $\mathcal{O}\left(1/k^2\right)$ and $\mathcal{O}\left(1/k^{7/2}\right)$, respectively. The A-HPE and/or the large-step A-HPE algorithms were recently studied also in [3, 5, 6, 9, 11, 13], with applications in high-order optimization, machine learning and tensor methods.

In this paper, we consider the (unconstrained) convex optimization problem

$$\min_x \{h(x) := f(x) + g(x)\}, \tag{2}$$

where f is convex and g is *strongly convex*. For solving (2), we propose and study the convergence rates of variants of the A-HPE and large-step A-HPE algorithms. The new algorithms are designed especially for strongly convex problems, and the resulting global convergence rates are *linear* and $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ for the variants of the A-HPE and large-step A-HPE, respectively. (the parameter $p \geq 2$ appears in the high-order large-step condition (see also [11, 13].) We also apply our study to tensor algorithms for high-order convex optimization, a topic which has been the object of investigation of several authors (see, e.g., [6, 7, 10, 11, 13, 21, 22] and references therein). The proposed inexact (relative-error) p-th order tensor algorithm has global *superlinear* $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ convergence rate. We also mention that, for $p = 2$ we obtain, as a by-product of our approach to high-order optimization, a fast $\mathcal{O}\left(k^{-k/3}\right)$ proximal-Newton method for strongly convex optimization.

The main contributions of this paper can be summarized as follows:

(i) A variant of the A-HPE algorithm for strongly convex objectives (Algorithm 1) and its iteration-complexity analysis as in Theorems 2.6 and 2.9.

(ii) A large-step A-HPE-type algorithm for strongly convex problems (Algorithm 2) with a high-order large-step condition and its iteration-complexity (see Theorem 3.3).

(iii) A new inexact high-order tensor algorithm (Algorithm 3) for strongly convex problems and its global convergence analysis (see Theorem 4.4). Here and in item (ii) above we highlight the fast global convergence rate $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$.

(iv) An inexact relative-error forward-backward algorithm for strongly convex optimization (see Algorithm 4 and Theorem 5.4).

Additionally to the contributions described in (i)–(iv) above, we refer the reader to the remarks/comments following Algorithms 1, 2, 3 and 4.

Some previous contributions. Based on the A-HPE framework of Monteiro and Svaiter [18], pth-order tensor methods with iteration-complexity $\mathcal{O}\left(1/k^{\frac{3p+1}{2}}\right)$ were studied in [3, 6, 9, 11, 13].

When combined with restart techniques, improved rates for the uniformly- and/or strongly-convex case were also obtained in [3, 9] (see also [12]). The A-HPE for strongly-convex problems was also recently studied in [5] within the framework of "performance estimation problems (PEPs)" (see remark (iv) following Algorithm 1). We also mention that local superlinear convergence rates for tensor methods were obtained in [7].

General notation. We denote by $\mathcal{H}$ a finite-dimensional real vector space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. We also use the standard notation and definitions of convex analysis [24] for subdifferentials, set-valued maps, etc. Recall that $g : \mathcal{H} \to (-\infty, \infty]$ is μ-strongly convex if $\mu > 0$ and, for all $x, y \in \mathcal{H}$,

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) - \frac{1}{2}\mu\lambda(1-\lambda)\|x-y\|^2, \qquad \forall \lambda \in [0,1]. \tag{3}$$

2 A variant of the A-HPE algorithm for strongly convex problems

In this section, we consider the convex optimization problem (2), i.e.,

$$\min_{x \in \mathcal{H}} \{h(x) := f(x) + g(x)\},$$

where $f, g : \mathcal{H} \to (-\infty, \infty]$ are proper, closed and convex functions, $\operatorname{dom} h \neq \emptyset$, and g is *μ-strongly convex*, for some $\mu > 0$. We will denote by x^* the unique solution of (2).

Next we present the main algorithm of this section for solving (2), whose the complexity analysis will be presented in Theorems 2.6 and 2.9.

Algorithm 1. A variant of the A-HPE algorithm for solving the (strongly convex) problem (2)

0) Choose $x^0, y^0 \in \mathcal{H}$, $\sigma \in [0,1]$, let $A_0 = 0$ and set $k = 0$.

1) Compute $\lambda_{k+1} > 0$ and $(y^{k+1}, v^{k+1}, \varepsilon_{k+1}) \in \mathcal{H} \times \mathcal{H} \times \mathbb{R}_{++}$ such that

$$\begin{aligned} &v^{k+1} \in \partial_{\varepsilon_{k+1}} f(y^{k+1}) + \partial g(y^{k+1}), \\ &\frac{\|\lambda_{k+1} v^{k+1} + y^{k+1} - \widetilde{x}^k\|^2}{1 + \lambda_{k+1}\mu} + 2\lambda_{k+1}\varepsilon_{k+1} \le \sigma^2 \|y^{k+1} - \widetilde{x}^k\|^2, \end{aligned} \tag{4}$$

where

$$\widetilde{x}^k = \left(\frac{a_{k+1} - \mu A_k \lambda_{k+1}}{A_k + a_{k+1}}\right) x^k + \left(\frac{A_k + \mu A_k \lambda_{k+1}}{A_k + a_{k+1}}\right) y^k, \tag{5}$$

$$a_{k+1} = \frac{(1 + 2\mu A_k)\lambda_{k+1} + \sqrt{(1 + 2\mu A_k)^2 \lambda_{k+1}^2 + 4(1 + \mu A_k) A_k \lambda_{k+1}}}{2}. \tag{6}$$

2) Let

$$A_{k+1} = A_k + a_{k+1}, \tag{7}$$

$$x^{k+1} = \left(\frac{1 + \mu A_k}{1 + \mu A_{k+1}}\right) x^k + \left(\frac{\mu a_{k+1}}{1 + \mu A_{k+1}}\right) y^{k+1} - \left(\frac{a_{k+1}}{1 + \mu A_{k+1}}\right) v^{k+1}. \tag{8}$$

3) Set $k = k + 1$ and go to step 1.

Next we make the following remarks concerning Algorithm 1:

(i) By letting $\mu = 0$ in Algorithm 1, we obtain a special instance of the A-HPE algorithm of Monteiro and Svaiter (see [18, Section 3]), whose global convergence rate is $\mathcal{O}\left(1/k^2\right)$ (see [18, Theorem 3.8]). On the other hand, thanks to the strong-convexity assumption on g, in Theorems 2.6 and 2.9 we obtain *linear convergence* for Algorithm 1. We will also study a *high-order* large-step version of Algorithm 1 (see Algorithm 2 in Section 3), for which *superlinear* $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ global convergence rates are proved, where $p \ge 2$. Applications of the latter result to high-order tensor methods for convex optimization will also be discussed in Section 3.

(ii) Since steps (5)–(8) are negligible (from a computational viewpoint), it follows that the computational burden of Algorithm 1 is represented by the computation of $\lambda_{k+1} > 0$ and $(y^{k+1}, v^{k+1}, \varepsilon_{k+1})$ as in (4). In this regard, note that if $\mathrm{prox}_{\lambda h} := (\lambda \partial h + I)^{-1}$ of h is computable, for $\lambda > 0$, then $\lambda_{k+1} := \lambda$ and $(y^{k+1}, v^{k+1}, \varepsilon_{k+1}) := \left(\mathrm{prox}_{\lambda h}(\widetilde{x}^k), \frac{\widetilde{x}^k - y^{k+1}}{\lambda_{k+1}}, 0\right)$ clearly satisfy the conditions in (4) with $\sigma = 0$. On the other hand, in the more general setting of $\sigma > 0$, Algorithm 1 can be used both as a framework for the design and analysis of practical algorithms [18] and as a

bilevel method, in which the inequality in (4) is used as a stopping criterion for some *inner* algorithm applied to the regularized inclusion $0 \in \lambda h(x) + x - \widetilde{x}^k$. In this case, note that the *error-criterion* in (4) is *relative* and controlled by the parameter $\sigma \in (0, 1]$.

(iii) We emphasize that the inequality in (4) is specially tailored to strongly convex problems, in the sense that it is more general than the usual inequality appearing in relative-error HPE-type methods (see, e.g., [1, 8, 16, 17, 26]), which in the context of this paper would read as

$$\|\lambda_{k+1} v^{k+1} + y^{k+1} - \widetilde{x}^k\|^2 + 2\lambda_{k+1}\varepsilon_{k+1} \leq \sigma^2 \|y^{k+1} - \widetilde{x}^k\|^2.$$

(iv) We also mention that Algorithm 1 is closely related to a variant of the A-HPE for strongly convex objectives presented and studied in [5, Section 4.2]. However, in constrast to the analysis in [5], which is supported on "performance estimation problems (PEPs)", in this contribution we take an approach similar to the one which was taken in [18, 23]. In doing so, we obtain global convergence rates for Algorithm 1 in terms of *function values*, *sequences* and *(sub-)gradients* (see Theorems 2.6 and 2.9). In contrast to [5], in this paper we also consider a *large-step* version of Algorithm 1, namely Algorithm 2, for which the (global) superlinear $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ convergence rate is proved (see Theorems 3.3 and 4.4).

(v) We note that condition (6) yields

$$\frac{(1 + \mu A_k) A_{k+1} \lambda_{k+1}}{a_{k+1}^2} + \frac{\mu A_k \lambda_{k+1}}{a_{k+1}} = 1. \tag{9}$$

Indeed, substituion of A_{k+1} by $A_k + a_{k+1}$ (see (7)) and some simple algebra give that (9) is equivalent to

$$a_{k+1}^2 - (1 + 2\mu A_k)\lambda_{k+1} a_{k+1} - (1 + \mu A_k) A_k \lambda_{k+1} = 0. \tag{10}$$

Note now that a_{k+1} as in (6) is exactly the largest root of the quadratic equation in (10).

(vi) Using (7) and the fact that $A_0 = 0$ (see step 0) we obtain $A_1 = A_0 + a_1 = a_1$. On the other hand, direct substituion of $A_0 = 0$ in (6) with $k = 0$ yields $a_1 = \lambda_1$. As a consequence, we conclude that

$$A_1 = a_1 = \lambda_1. \tag{11}$$

Before analyzing the convergence rates of Algorithm 1 we will need the following:
Define, for $k \geq 1$,

$$\gamma_k(x) = h(y^k) + \langle v^k, x - y^k \rangle - \varepsilon_k + \frac{\mu}{2}\|x - y^k\|^2 \qquad (x \in \mathcal{H}) \tag{12}$$

and

$$\Gamma_0 = 0 \text{ and, for } k \geq 1, \ \Gamma_k = \sum_{j=1}^{k} \frac{a_j}{A_k} \gamma_j. \tag{13}$$

Note that

$$\nabla\gamma_k(x) = v^k + \mu(x - y^k) \text{ and } \nabla^2\gamma_k(x) = \mu I \tag{14}$$

and observe that A_k $(k = 0, 1, \dots)$ as in Algorithm 1 satisfies

$$A_0 = 0 \text{ and, for } k \geq 1, \ A_k = \sum_{j=1}^{k} a_j. \tag{15}$$

From (13)–(15) we obtain, for $k \geq 1$,

$$\nabla^2\Gamma_k(x) = \mu I, \qquad x \in \mathcal{H}. \tag{16}$$

Note also that the following holds trivialy from (13) and (15): for all $k \geq 0$,

$$A_{k+1}\Gamma_{k+1} = A_k\Gamma_k + a_{k+1}\gamma_{k+1}. \tag{17}$$

Define also, for all $k \geq 0$,

$$\beta_k = \inf_{x\in\mathcal{H}} \left\{ A_k\Gamma_k(x) + \frac{1}{2}\|x - x^0\|^2 \right\}. \tag{18}$$

Note that $\beta_0 = 0$.

The following three technical lemmas will be useful to prove the first result on the iteration-complexity of Algorithm 1, namely Proposition 2.4 below.

Lemma 2.1. *Let $\gamma_k(\cdot)$ and $\Gamma_k(\cdot)$ be as in (12) and (13), respectively. The following holds:*

(a) *For all $k \geq 1$, we have $\gamma_k(x) \leq h(x), \quad \forall x \in \mathcal{H}$.*

(b) *For all $k \geq 0$, we have $x^k = \arg\min_{x\in\mathcal{H}}\{A_k\Gamma_k(x) + \frac{1}{2}\|x - x^0\|^2\}$.*

Proof. (a) In view of the inclusion in (4) we have, for all $k \geq 1$, $v^k = r^k + s^k$, where $r^k \in \partial_{\varepsilon_k} f(y^k)$ and $s^k \in \partial g(y^k)$. Using the assumption that g is μ-strongly convex and the definition of the ε-subdifferential of f we obtain, for all $x \in \mathcal{H}$,

$$\begin{aligned} f(x) &\geq f(y^k) + \langle r^k, x - y^k\rangle - \varepsilon_k, \\ g(x) &\geq g(y^k) + \langle s^k, x - y^k\rangle + \frac{\mu}{2}\|x - y^k\|^2, \end{aligned}$$

which in turn combined with the definition of $h(\cdot)$ in (2), the fact that $v^k = r^k + s^k$ and (12) yields the desired result.

(b) Let us proceed by induction on $k \geq 0$. The result is trivially true for $k = 0$ (since $A_0\Gamma_0 = 0$). Assume now that it is true for some $k \geq 0$, i.e., assume that $x^k = \arg\min_x\{A_k\Gamma_k(x) + \frac{1}{2}\|x - x^0\|^2\}$. Using the latter identity, (16)–(18) and Taylor's theorem we find

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 &= A_k\Gamma_k(x) + \frac{1}{2}\|x - x^0\|^2 + a_{k+1}\gamma_{k+1}(x) \\ &= \beta_k + \left(\frac{1 + \mu A_k}{2}\right)\|x - x^k\|^2 + a_{k+1}\gamma_{k+1}(x). \end{aligned} \tag{19}$$

From the definition of $\gamma_{k+1}(\cdot)$ (see (12)) and some simple calculus one can check that x^{k+1} as in (8) is exactly the (unique) minimizer of $x \mapsto \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2 + a_{k+1}\gamma_{k+1}(x)$. Hence, from this fact and (19) we obtain that $x^{k+1} = \arg\min_{x\in\mathcal{H}}\{A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2\}$, completing the induction argument. □

Lemma 2.2. *Consider the sequences evolved by* Algorithm 1. *The following holds for all $x \in \mathcal{H}$:*

(a) *For all $k \geq 0$,*

$$A_k\Gamma_k(x) + \frac{1}{2}\|x - x^0\|^2 = \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2.$$

(b) *For all $k \geq 0$,*

$$A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 = \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2 + a_{k+1}\gamma_{k+1}(x).$$

(c) *For all $k \geq 0$,*

$$A_k h(y^k) + A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 \geq \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2 + a_{k+1}\gamma_{k+1}(x) + A_k\gamma_{k+1}(y^k).$$

Proof. (a) First note that the result is trivial for $k = 0$, since $\beta_0 = A_0 = 0$ and $\Gamma_0 = 0$. Now note that in view of (16) we obtain, for $k \geq 1$,

$$\nabla^2\left(A_k\Gamma_k(\cdot) + \frac{1}{2}\|\cdot - x^0\|^2\right)(x) = 1 + \mu A_k.$$

Using the latter identity, Lemma 2.1(b), (18) and Taylor's theorem we find

$$\begin{aligned} A_k\Gamma_k(x) + \frac{1}{2}\|x - x^0\|^2 &= \underbrace{A_k\Gamma_k(x^k) + \frac{1}{2}\|x^k - x^0\|^2}_{\beta_k} + \frac{1}{2}\langle(1+\mu A_k)(x - x^k), x - x^k\rangle \\ &= \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2. \end{aligned}$$

(b) From (17) and item (a), we obtain, for all $k \geq 0$,

$$\begin{aligned} A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 &= A_k\Gamma_k(x) + \frac{1}{2}\|x - x^0\|^2 + a_{k+1}\gamma_{k+1}(x) \\ &= \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2 + a_{k+1}\gamma_{k+1}(x). \end{aligned}$$

(c) From (b) and Lemma 2.1(a) with $k = k+1$ and $x = y^k$,

$$\begin{aligned} A_k h(y^k) + A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 &= \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2 + a_{k+1}\gamma_{k+1}(x) + A_k h(y^k) \\ &\geq \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2 + a_{k+1}\gamma_{k+1}(x) + A_k\gamma_{k+1}(y^k). \end{aligned}$$

□

Lemma 2.3. *Consider the sequences evolved by* Algorithm 1. *The following holds:*

(a) *For all $k \geq 0$ and $x \in \mathcal{H}$,*

$$a_{k+1}\gamma_{k+1}(x) + A_k\gamma_{k+1}(y^k) = A_{k+1}\gamma_{k+1}(\widetilde{x}) + \left(\frac{\mu\, a_{k+1}A_k}{2A_{k+1}}\right)\|x - y^k\|^2,$$

where

$$\widetilde{x} := \frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}y^k. \tag{20}$$

(b) *For all $k \geq 0$ and $x \in \mathcal{H}$,*

$$A_k h(y^k) + A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 \geq \beta_k + A_{k+1}\Big[\gamma_{k+1}(\widetilde{x}) + \Delta_k\Big],$$

where, for all $k \geq 0$, $\widetilde{x}$ is as in (20) *and*

$$\Delta_k := \left(\frac{(1+\mu A_k)A_{k+1}}{2a_{k+1}^2}\right)\|\widetilde{x} - z^k\|^2 + \left(\frac{\mu A_k}{2a_{k+1}}\right)\|\widetilde{x} - y^k\|^2, \tag{21}$$

$$z^k := \frac{a_{k+1}}{A_{k+1}}x^k + \frac{A_k}{A_{k+1}}y^k. \tag{22}$$

(c) *For all $k \geq 0$,*

$$\Delta_k = \frac{1}{2\lambda_{k+1}}\left[\|\widetilde{x} - \widetilde{x}^k\|^2 + \left(\frac{\mu(1+\mu A_k)\lambda_{k+1}^2 A_k}{a_{k+1}A_{k+1}}\right)\|x^k - y^k\|^2\right], \tag{23}$$

where $\widetilde{x}$ is as in (20).

(d) *For all $k \geq 0$ and $x \in \mathcal{H}$,*

$$A_k h(y^k) + A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 \geq \beta_k + A_{k+1}h(y^{k+1}) + \left(\frac{1-\sigma^2}{2}\right)\left(\frac{A_{k+1}}{\lambda_{k+1}}\|y^{k+1} - \widetilde{x}^k\|^2\right) + \left(\frac{\mu(1+\mu A_k)\lambda_{k+1}A_k}{2a_{k+1}}\right)\|x^k - y^k\|^2.$$

Proof. (a) First recall that (see (12))

$$\gamma_{k+1}(x) = \underbrace{h(y^{k+1}) + \langle v^{k+1}, x - y^{k+1}\rangle - \varepsilon_{k+1}}_{\ell_{k+1}(x)} + \frac{\mu}{2}\|x - y^{k+1}\|^2, \qquad \forall x \in \mathcal{H}. \tag{24}$$

Let $p = \frac{a_{k+1}}{A_{k+1}}$, $q = \frac{A_k}{A_{k+1}}$ and note that $p, q \geq 0$, $p + q = 1$ and $\widetilde{x} = px + qy^k$. Since $\ell_{k+1}(\cdot)$ is affine, we find

$$\begin{aligned}\ell_{k+1}(\widetilde{x}) = \ell_{k+1}(px + qy^k) &= p\ell_{k+1}(x) + q\ell_{k+1}(y^k)\\ &= \frac{1}{A_{k+1}}\left[a_{k+1}\ell_{k+1}(x) + A_k\ell_{k+1}(y^k)\right].\end{aligned} \tag{25}$$

On the other hand, using the well-know identity $\|pz+qw\|^2 = p\|z\|^2 + q\|w\|^2 - pq\|z-w\|^2$, for all $z, w \in \mathcal{H}$, we also find

$$\begin{aligned}\|\widetilde{x} - y^{k+1}\|^2 &= \|p(x-y^{k+1}) + q(y^k - y^{k+1})\|^2 \\ &= p\|x-y^{k+1}\|^2 + q\|y^k - y^{k+1}\|^2 - pq\|x-y^k\|^2 \\ &= \frac{1}{A_{k+1}}\left[a_{k+1}\|x-y^{k+1}\|^2 + A_k\|y^k - y^{k+1}\|^2 - \left(\frac{a_{k+1}A_k}{A_{k+1}}\right)\|x-y^k\|^2\right]. \end{aligned} \tag{26}$$

Combinding (24)–(26), we then obtain

$$\begin{aligned}\gamma_{k+1}(\widetilde{x}) &= \ell_{k+1}(\widetilde{x}) + \frac{\mu}{2}\left\|\widetilde{x} - y^{k+1}\right\|^2 \\ &= \frac{1}{A_{k+1}}\left[a_{k+1}\left(\ell_{k+1}(x) + \frac{\mu}{2}\|x-y^{k+1}\|^2\right) + A_k\left(\ell_{k+1}(y^k) + \frac{\mu}{2}\|y^k - y^{k+1}\|^2\right) - \left(\frac{\mu a_{k+1}A_k}{2A_{k+1}}\right)\|x-y^k\|^2\right] \\ &= \frac{1}{A_{k+1}}\left[a_{k+1}\gamma_{k+1}(x) + A_k\gamma_{k+1}(y^k) - \left(\frac{\mu a_{k+1}A_k}{2A_{k+1}}\right)\|x-y^k\|^2\right],\end{aligned}$$

which is clearly equivalent to the desired identity.

(b) First note that in view of (20) and (22) we have $\widetilde{x} - z^k = \frac{a_{k+1}}{A_{k+1}}(x - x^k)$ and, analogously, we also have $\widetilde{x} - y^k = \frac{a_{k+1}}{A_{k+1}}(x - y^k)$. Hence,

$$\|x-x^k\|^2 = \frac{A_{k+1}^2}{a_{k+1}^2}\|\widetilde{x} - z^k\|^2 \quad \text{and} \quad \|x-y^k\|^2 = \frac{A_{k+1}^2}{a_{k+1}^2}\|\widetilde{x} - y^k\|^2. \tag{27}$$

Using Lemma 2.2(c) and item (a) we find

$$\begin{aligned}A_k h(y^k) + A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x-x^0\|^2 &\geq \beta_k + \left(\frac{1+\mu A_k}{2}\right)\|x-x^k\|^2 + a_{k+1}\gamma_{k+1}(x) + A_k\gamma_{k+1}(y^k) \\ &= \beta_k + A_{k+1}\gamma_{k+1}(\widetilde{x}) \\ &\quad + \left(\frac{1+\mu A_k}{2}\right)\|x-x^k\|^2 + \left(\frac{\mu\, a_{k+1}A_k}{2A_{k+1}}\right)\|x-y^k\|^2,\end{aligned}$$

which in turn combined with (27) and (21) finishes the proof of item (b).

(c) First let $p = \frac{(1+\mu A_k)A_{k+1}\lambda_{k+1}}{a_{k+1}^2}$, $q = \dfrac{\mu A_k \lambda_{k+1}}{a_{k+1}}$ and note that $p, q \geq 0$ and, in view of (9), $p + q = 1$. From (21) and the above definitions of p and q, we obtain

$$\begin{aligned}\Delta_k &= \left(\frac{(1+\mu A_k)A_{k+1}}{2a_{k+1}^2}\right)\|\widetilde{x} - z^k\|^2 + \left(\frac{\mu A_k}{2a_{k+1}}\right)\|\widetilde{x} - y^k\|^2 \\ &= \frac{1}{2\lambda_{k+1}}\left[p\|\widetilde{x} - z^k\|^2 + q\|\widetilde{x} - y^k\|^2\right] \\ &= \frac{1}{2\lambda_{k+1}}\left[\|\widetilde{x} - (pz^k + qy^k)\|^2 + pq\|y^k - z^k\|^2\right],\end{aligned} \tag{28}$$

where we also used the well-known identity $p\|z\|^2 + q\|w\|^2 = \|pz+qw\|^2 + pq\|z-w\|^2$, for $z, w \in \mathcal{H}$.

Using (22), the definitions of p, q, the fact that $p + q = 1$, (5) and (7), and some simple computations, we find

$$\begin{aligned} pz^k + qy^k &= (1-q)\left(\frac{a_{k+1}}{A_{k+1}}x^k + \frac{A_k}{A_{k+1}}y^k\right) + qy^k \\ &= (1-q)\frac{a_{k+1}}{A_{k+1}}x^k + \left(\frac{A_k}{A_{k+1}} + q\left(1 - \frac{A_k}{A_{k+1}}\right)\right)y^k \\ &= (1-q)\frac{a_{k+1}}{A_{k+1}}x^k + \left(\frac{A_k}{A_{k+1}} + q\frac{a_{k+1}}{A_{k+1}}\right)y^k \\ &= \left(1 - \frac{\mu A_k \lambda_{k+1}}{a_{k+1}}\right)\frac{a_{k+1}}{A_{k+1}}x^k + \left(\frac{A_k}{A_{k+1}} + \left(\frac{\mu A_k \lambda_{k+1}}{a_{k+1}}\right)\frac{a_{k+1}}{A_{k+1}}\right)y^k \\ &= \left(\frac{a_{k+1} - \mu A_k \lambda_{k+1}}{A_{k+1}}\right)x^k + \left(\frac{A_k + \mu A_k \lambda_{k+1}}{A_{k+1}}\right)y^k \\ &= \widetilde{x}^k. \end{aligned} \tag{29}$$

On the other hand, using again (22) and the definitions of p, q, we also obtain

$$\begin{aligned} pq\|y^k - z^k\|^2 &= \left(\frac{(1+\mu A_k)A_{k+1}\lambda_{k+1}}{a_{k+1}^2}\right)\left(\frac{\mu A_k \lambda_{k+1}}{a_{k+1}}\right)\frac{a_{k+1}^2}{A_{k+1}^2}\|x^k - y^k\|^2 \\ &= \left(\frac{\mu(1+\mu A_k)\lambda_{k+1}^2 A_k}{a_{k+1}A_{k+1}}\right)\|x^k - y^k\|^2. \end{aligned} \tag{30}$$

The desired result now follows directly from (28), (29) and (30).

(d) From items (b) and (c),

$$\begin{aligned} A_k h(y^k) + A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x - x^0\|^2 \geq \beta_k + A_{k+1}\Big[\gamma_{k+1}(\widetilde{x}) + \frac{1}{2\lambda_{k+1}}\|\widetilde{x} - \widetilde{x}^k\|^2\Big] \\ + \left(\frac{\mu(1+\mu A_k)\lambda_{k+1}A_k}{2a_{k+1}}\right)\|x^k - y^k\|^2. \end{aligned} \tag{31}$$

From (12),

$$\begin{aligned} \gamma_{k+1}(\widetilde{x}) + \frac{1}{2\lambda_{k+1}}\|\widetilde{x} - \widetilde{x}^k\|^2 = h(y^{k+1}) \\ \underbrace{+ \langle v^{k+1}, \widetilde{x} - y^{k+1}\rangle + \frac{\mu}{2}\|\widetilde{x} - y^{k+1}\|^2 - \varepsilon_{k+1} + \frac{1}{2\lambda_{k+1}}\|\widetilde{x} - \widetilde{x}^k\|^2}_{=:q_{k+1}(\widetilde{x})}. \end{aligned} \tag{32}$$

On the other hand, from Lemma A.2(c) applied to $q_{k+1}(\cdot)$ and (4),

$$q_{k+1}(\widetilde{x}) \geq \left(\frac{1-\sigma^2}{2\lambda_{k+1}}\right)\|y^{k+1} - \widetilde{x}^k\|^2,$$

which in turn combined with (32) gives

$$\gamma_{k+1}(\widetilde{x}) + \frac{1}{2\lambda_{k+1}}\|\widetilde{x} - \widetilde{x}^k\|^2 \geq h(y^{k+1}) + \left(\frac{1-\sigma^2}{2\lambda_{k+1}}\right)\|y^{k+1} - \widetilde{x}^k\|^2.$$

The desired result now follows by the substitution of the latter inequality in (31). □

Next is our first result on the iteration-complexity of Algorithm 1. Item (b) follows trivially from item (a), which will be derived from Lemmas 2.1, 2.2 and 2.3. The main results on the iteration-complexity of Algorithm 1 will then be presented in Theorem 2.6 below.

Proposition 2.4. *Consider the sequences evolved by* Algorithm 1*, let x^* denote the (unique) solution of* (2) *and let*

$$d_0 := \|x^* - x^0\|. \tag{33}$$

The following holds:

(a) *For all $k \geq 1$ and $x \in \mathcal{H}$,*

$$A_k\left[h(y^k) - h(x)\right] + \left(\frac{1-\sigma^2}{2}\right)\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|y^j - \widetilde{x}^{j-1}\|^2 + \sum_{j=1}^{k}\left(\frac{\mu(1+\mu A_{j-1})\lambda_j A_{j-1}}{2a_j}\right)\|x^{j-1} - y^{j-1}\|^2 + \left(\frac{1+\mu A_k}{2}\right)\|x - x^k\|^2 \leq \frac{1}{2}\|x - x^0\|^2.$$

(b) *If $\sigma < 1$, for all $k \geq 1$,*

$$\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|y^j - \widetilde{x}^{j-1}\|^2 \leq \frac{d_0^2}{1-\sigma^2}, \qquad \forall k \geq 1. \tag{34}$$

Proof. (a) From Lemma 2.3(d) and the definition of β_{k+1} – see (18) – we obtain, for all $k \geq 0$,

$$A_k h(y^k) + \beta_{k+1} \geq \beta_k + A_{k+1}h(y^{k+1}) + \left(\frac{1-\sigma^2}{2}\right)\left(\frac{A_{k+1}}{\lambda_{k+1}}\|y^{k+1} - \widetilde{x}^k\|^2\right) + \left(\frac{\mu(1+\mu A_k)\lambda_{k+1}A_k}{2a_{k+1}}\right)\|x^k - y^k\|^2,$$

and so, for all $k \geq 0$,

$$\underbrace{\sum_{j=0}^{k}\left[\beta_{j+1} - \beta_j\right]}_{\beta_{k+1}-\beta_0} \geq \underbrace{\sum_{j=0}^{k}\left[A_{j+1}h(y^{j+1}) - A_j h(y^j)\right]}_{A_{k+1}h(y^{k+1})-A_0h(y^0)} + \left(\frac{1-\sigma^2}{2}\right)\sum_{j=0}^{k}\frac{A_{j+1}}{\lambda_{j+1}}\|y^{j+1} - \widetilde{x}^j\|^2 + \sum_{j=0}^{k}\left(\frac{\mu(1+\mu A_j)\lambda_{j+1}A_j}{2a_{j+1}}\right)\|x^j - y^j\|^2.$$

which, since $\beta_0 = A_0 = 0$, yields, for all $k \geq 0$,

$$\beta_{k+1} \geq A_{k+1}h(y^{k+1}) + \left(\frac{1-\sigma^2}{2}\right)\sum_{j=1}^{k+1}\frac{A_j}{\lambda_j}\|y^j - \widetilde{x}^{j-1}\|^2 + \sum_{j=1}^{k+1}\left(\frac{\mu(1+\mu A_{j-1})\lambda_j A_{j-1}}{2a_j}\right)\|x^{j-1} - y^{j-1}\|^2.$$

By adding $\left(\frac{1+\mu A_{k+1}}{2}\right)\|x-x^{k+1}\|^2$ in both sides of the latter inequality, we obtain, for all $k \geq 0$,

$$\begin{aligned}\beta_{k+1} + \left(\frac{1+\mu A_{k+1}}{2}\right)\|x-x^{k+1}\|^2 \geq A_{k+1}h(y^{k+1}) + \left(\frac{1-\sigma^2}{2}\right)\sum_{j=1}^{k+1}\frac{A_j}{\lambda_j}\|y^j-\widetilde{x}^{j-1}\|^2 \\ + \sum_{j=1}^{k+1}\left(\frac{\mu(1+\mu A_{j-1})\lambda_j A_{j-1}}{2a_j}\right)\|x^{j-1}-y^{j-1}\|^2 \\ + \left(\frac{1+\mu A_{k+1}}{2}\right)\|x-x^{k+1}\|^2.\end{aligned}$$

Using Lemma 2.2(a) we then find, for all $k \geq 0$,

$$\begin{aligned}A_{k+1}\Gamma_{k+1}(x) + \frac{1}{2}\|x-x^0\|^2 \geq A_{k+1}h(y^{k+1}) + \left(\frac{1-\sigma^2}{2}\right)\sum_{j=1}^{k+1}\frac{A_j}{\lambda_j}\|y^j-\widetilde{x}^{j-1}\|^2 \\ + \sum_{j=1}^{k+1}\left(\frac{\mu(1+\mu A_{j-1})\lambda_j A_{j-1}}{2a_j}\right)\|x^{j-1}-y^{j-1}\|^2 \\ + \left(\frac{1+\mu A_{k+1}}{2}\right)\|x-x^{k+1}\|^2. \qquad (35)\end{aligned}$$

Note now that from (13) and Lemma 2.1(a) we obtain, for all $k \geq 0$,

$$A_{k+1}\Gamma_{k+1}(x) = \sum_{j=1}^{k+1}\lambda_j\gamma_j(x) \leq A_{k+1}h(x),$$

which combined with (35) yields, for all $k \geq 1$,

$$\begin{aligned}\frac{1}{2}\|x-x^0\|^2 \geq A_k\left[h(y^k)-h(x)\right] + \left(\frac{1-\sigma^2}{2}\right)\sum_{j=1}^{k}\frac{A_j}{\lambda_j}\|y^j-\widetilde{x}^{j-1}\|^2 \\ + \sum_{j=1}^{k}\left(\frac{\mu(1+\mu A_{j-1})\lambda_j A_{j-1}}{2a_j}\right)\|x^{j-1}-y^{j-1}\|^2 + \left(\frac{1+\mu A_k}{2}\right)\|x-x^k\|^2.\end{aligned}$$

(b) This follows trivially from item (a) and (33). □

Lemma 2.5. *For all $k \geq 0$,*

$$\left(1-\sigma\sqrt{1+\lambda_{k+1}\mu}\right)\|y^{k+1}-\widetilde{x}^k\| \leq \|\lambda_{k+1}v^{k+1}\| \leq \left(1+\sigma\sqrt{1+\lambda_{k+1}\mu}\right)\|y^{k+1}-\widetilde{x}^k\|. \qquad (36)$$

Proof. The proof follows from the inequality in (4), the fact that $\varepsilon_{k+1} \geq 0$ and a simple argument based on the triangle inequality. □

Since, under mild regularity assumptions on f and g, problem (2) is equivalent to the inclusion

$$0 \in \partial f(x) + \partial g(x), \tag{37}$$

it is natural to attempt to evaluate the residuals produced by Algorithm 1 in the light of (37), and this is exactly what Theorem 2.6(b) is about. Note that if we set $v^{k+1} = 0$ and $\varepsilon_{k+1} = 0$ in (38), then it follows that $x := y^{k+1}$ satisfies the inclusion (37).

As we mentioned before, Theorem 2.6 below is our main result on the iteration-complexity of Algorithm 1.

Theorem 2.6 (**Convergence rates for Algorithm 1).** *Consider the sequences evolved by* Algorithm 1*, let x^* be the (unique) solution of* (2) *and let d_0 be as in* (33)*. Then, the following holds:*

(a) *For all $k \geq 1$,*

$$h(y^k) - h(x^*) \leq \frac{d_0^2}{2A_k}, \qquad \|x^* - y^k\|^2 \leq \frac{d_0^2}{\mu A_k}, \qquad \|x^* - x^k\|^2 \leq \frac{d_0^2}{1 + \mu A_k}.$$

(b) *For all $k \geq 1$,*

$$\begin{cases} v^{k+1} \in \partial_{\varepsilon_{k+1}} f(y^{k+1}) + \partial g(y^{k+1}), \\ \|v^{k+1}\|^2 \leq \left(\dfrac{1 + \sigma\sqrt{1 + \mu\lambda_{k+1}}}{6^{-1/2}\lambda_{k+1}} \right)^2 \dfrac{d_0^2}{\mu A_k}, \\ \varepsilon_{k+1} \leq \left(\dfrac{3\sigma^2}{\lambda_{k+1}} \right) \dfrac{d_0^2}{\mu A_k}. \end{cases} \tag{38}$$

Proof. (a) Note that the bounds on $h(y^k) - h(x^*)$ and $\|x^* - x^k\|^2$ follow directly from Proposition 2.4(a) with $x = x^*$ and (33). Now, since $h(\cdot)$ is μ-strongly convex and $0 \in \partial h(x^*)$, one can use the inequality (see, e.g., [25, Proposition 6(c)]) $h(x) \geq h(x^*) + \frac{\mu}{2}\|x - x^*\|^2$, for all $x \in \mathcal{H}$, with $x = y^k$ and the bound on $h(y^k) - h(x^*)$ to conclude that $\|y^k - x^*\|^2 \leq \frac{2}{\mu}\left(h(y^k) - h(x^*)\right) \leq \frac{d_0^2}{\mu A_k}$.

(b) First, note that the inclusion in (38) follows from the inclusion in (4). Since we will use the second inequality in (36) to prove the inequality for $\|v^{k+1}\|^2$, it follows that we first have to bound the term $\|y^{k+1} - \widetilde{x}^k\|^2$. To this end, note that from the second inequality in item (a) with $k = k + 1$ and the fact that $A_{k+1} \geq A_k$,

$$\begin{aligned} \|y^{k+1} - \widetilde{x}^k\|^2 &\leq 2\left(\|x^* - y^{k+1}\|^2 + \|\widetilde{x}^k - x^*\|^2\right) \\ &\leq 2\left(\frac{d_0^2}{\mu A_k} + \|\widetilde{x}^k - x^*\|^2\right). \end{aligned} \tag{39}$$

We now have to bound the second term in (39). Since, from (5), $\widetilde{x}^k$ is a convex combination of x^k

and y^k, it follows that

$$\begin{aligned}\|\widetilde{x}^k - x^*\|^2 &\leq \|x^* - x^k\|^2 + \|x^* - y^k\|^2 \\ &\leq \frac{d_0^2}{1+\mu A_k} + \frac{d_0^2}{\mu A_k} \\ &\leq \frac{2d_0^2}{\mu A_k}, \end{aligned} \tag{40}$$

where in the second inequality we used the second and third inequalities in item (a). Now using (39) and (40), we find

$$\|y^{k+1} - \widetilde{x}^k\|^2 \leq 6\frac{d_0^2}{\mu A_k}. \tag{41}$$

To finish the proof of (b), note that using (41), we obtain the desired bounds on $\|v^{k+1}\|^2$ and ε_{k+1} as a consequence of the second inequality in (36) and the fact that $2\lambda_{k+1}\varepsilon_{k+1} \leq \sigma^2\|y^{k+1} - \widetilde{x}^k\|^2$ (see (4)), respectively. □

Next result is motivated by the fact that the rate of convergence of Algorithm 1 presented in Theorem 2.6 is given in terms of the sequence $\{A_k\}$. We also mention that the proof below (of Lemma 2.7) follows the same outline of an argument given in [5, Corollary 4.4].

Lemma 2.7. *The following holds:*

(a) *For all $k \geq 1$,*

$$A_{k+1} \geq \lambda_1 \prod_{j=2}^{k+1} \left(\frac{1}{1 - \sqrt{\dfrac{\mu\lambda_j}{1+\mu\lambda_j}}} \right). \tag{42}$$

(b) *For all $k \geq 1$,*

$$A_{k+1} \geq \lambda_1 \prod_{j=2}^{k+1} (1 + 2\mu\lambda_j). \tag{43}$$

Proof. (a) From (6),

$$\begin{aligned} a_{k+1} &= \frac{(1+2\mu A_k)\lambda_{k+1} + \sqrt{(1+2\mu A_k)^2\lambda_{k+1}^2 + 4(1+\mu A_k)A_k\lambda_{k+1}}}{2} \\ &\geq \frac{(2\mu A_k)\lambda_{k+1} + \sqrt{(2\mu A_k)^2\lambda_{k+1}^2 + 4(\mu A_k)A_k\lambda_{k+1}}}{2} \\ &= \frac{(2\mu A_k)\lambda_{k+1} + 2A_k\sqrt{\mu^2\lambda_{k+1}^2 + \mu\lambda_{k+1}}}{2} \\ &= A_k\left[\mu\lambda_{k+1} + \sqrt{\mu\lambda_{k+1}(1+\mu\lambda_{k+1})}\right]. \end{aligned}$$

Hence, from (7),

$$\begin{aligned} A_{k+1} &= A_k + a_{k+1} \\ &\geq A_k + A_k \left[\mu\lambda_{k+1} + \sqrt{\mu\lambda_{k+1}(1+\mu\lambda_{k+1})}\right] \\ &= A_k \left[1 + \mu\lambda_{k+1} + \sqrt{\mu\lambda_{k+1}(1+\mu\lambda_{k+1})}\right] \qquad (44) \\ &= A_k \left(\frac{1}{1-\sqrt{\frac{\mu\lambda_{k+1}}{1+\mu\lambda_{k+1}}}}\right), \qquad (45) \end{aligned}$$

where in the last equality we used the identity $1/\left(1-\sqrt{\frac{x}{1+x}}\right) = 1+x+\sqrt{x(1+x)}$ with $x = \mu\lambda_{k+1}$. Note now that (42) follows directly from (45) and the fact that $A_1 = \lambda_1$ – see (11).

(b) Using (44), the fact that $\sqrt{\mu\lambda_{k+1}(1+\mu\lambda_{k+1})} \geq \mu\lambda_{k+1}$ and a similiar reasoning to the proof of item (a), we obtain that (43) holds for all $k \geq 1$. □

Next is a corollary of Lemma 2.7(a) for the special case that the sequence $\{\lambda_k\}$ is bounded away from zero. Lemma 2.7(b) will be useful later in Section 3.

Corollary 2.8. *Assume that $\lambda_k \geq \underline{\lambda} > 0$, for all $k \geq 1$, and define $\alpha \in (0,1)$ as*

$$\alpha := \sqrt{\frac{\mu\underline{\lambda}}{1+\mu\underline{\lambda}}}. \qquad (46)$$

Then, for all $k \geq 1$,

$$A_k \geq \underline{\lambda}\left(\frac{1}{1-\alpha}\right)^{k-1}. \qquad (47)$$

Proof. Using the fact that the scalar function $(0,\infty) \ni t \mapsto \frac{\mu t}{1+\mu t} \in (0,1)$ is increasing, the assumption $\lambda_k \geq \underline{\lambda} > 0$, for all $k \geq 1$, and (46), we find

$$\frac{1}{1-\sqrt{\frac{\mu\lambda_j}{1+\mu\lambda_j}}} \geq \frac{1}{1-\alpha}, \qquad \forall j \geq 1.$$

Hence, from Lemma 2.7(a) and the assumption $\lambda_k \geq \underline{\lambda}$ with $k=1$ we obtain $A_{k+1} \geq \underline{\lambda}\left(\frac{1}{1-\alpha}\right)^k$, for all $k \geq 1$, which is clearly equivalent to $A_k \geq \underline{\lambda}\left(\frac{1}{1-\alpha}\right)^{k-1}$ for all $k \geq 2$. To finish the proof of item (a), note that the latter inequality holds trivialy for $k=1$ (because $A_1 = \lambda_1$ and $\lambda_1 \geq \underline{\lambda}$). □

Next we present convergence rate results for Algorithm 1 under the assumption that $\{\lambda_k\}$ is bounded away from zero.

Theorem 2.9 (Convergence rates for Algorithm 1 with $\{\lambda_k\}$ bounded below). *Consider the sequences evolved by* Algorithm 1 *and assume that $\lambda_k \geq \underline{\lambda} > 0$ for all $k \geq 1$. Let x^* be the (unique) solution of (2), let d_0 be as in (33) and let $\alpha \in (0,1)$ be as in (46). The following holds:*

(a) *For all $k \geq 1$,*

$$h(y^k) - h(x^*) \leq \frac{d_0^2}{2\underline{\lambda}}(1-\alpha)^{k-1},$$

$$\max\left\{\|x^* - y^k\|, \|x^* - x^k\|\right\} \leq \frac{d_0}{\sqrt{\mu\underline{\lambda}}}(1-\alpha)^{(k-1)/2}.$$

(b) *For all $k \geq 1$,*

$$\begin{cases} v^{k+1} \in \partial_{\varepsilon_{k+1}} f(y^{k+1}) + \partial g(y^{k+1}), \\ \|v^{k+1}\| \leq \left(\dfrac{1+\sigma\sqrt{1+\mu\underline{\lambda}}}{6^{-1/2}\mu^{1/2}\underline{\lambda}^{3/2}}\right) d_0\,(1-\alpha)^{(k-1)/2}, \\ \varepsilon_{k+1} \leq \left(\dfrac{3\sigma^2 d_0^2}{\mu\underline{\lambda}^2}\right)(1-\alpha)^{k-1}. \end{cases}$$

Proof. (a) This follows from Theorem 2.6(a) and Corollary 2.8.

(b) The result follows from Theorem 2.6(b), Corollary 2.8, the assumption $\lambda_k \geq \underline{\lambda}$ and the fact that, for $t > 0$, the scalar function $t \mapsto \frac{1+\sigma\sqrt{1+\mu t}}{t}$ is nonincreasing. □

3 A (high-order) large-step A-HPE algorithm for strongly convex problems

In this section, we also consider problem (2), i.e., $\min_{x\in\mathcal{H}}\{h(x) := f(x) + g(x)\}$, where the same assumptions as in Section 2 are assumed to hold on h, f and g.

For solving (2), we propose and study the iteration-complexity of a variant (Algorithm 2) of the large-step A-HPE algorithm of Monteiro and Svaiter [18], with a high-order large-step condition specially tailored for strongly convex objectives . Applications of this general framework to high-order tensor methods will be given in Section 4. The main results on convergence rates for Algorithm 2 are presented in Theorem 3.3 below.

Algorithm 2. A variant of the large-step A-HPE algorithm for (the strongly convex) problem (2)

0) Choose $x^0, y^0 \in \mathcal{H}$, $\sigma \in [0,1)$, $p \geq 2$ and $\theta > 0$; let $A_0 = 0$ and set $k = 0$.

1) Compute $\lambda_{k+1} > 0$ and $(y^{k+1}, v^{k+1}, \varepsilon_{k+1}) \in \mathcal{H} \times \mathcal{H} \times \mathbb{R}_+$ such that

$$v^{k+1} \in \partial_{\varepsilon_{k+1}} f(y^{k+1}) + \partial g(y^{k+1}),$$

$$\frac{\|\lambda_{k+1} v^{k+1} + y^{k+1} - \widetilde{x}^k\|^2}{1 + \lambda_{k+1}\mu} + 2\lambda_{k+1}\varepsilon_{k+1} \leq \sigma^2 \|y^{k+1} - \widetilde{x}^k\|^2, \tag{48}$$

$$\lambda_{k+1}\|y^{k+1} - \widetilde{x}^k\|^{p-1} \geq \theta,$$

where

$$\widetilde{x}^k = \left(\frac{a_{k+1} - \mu A_k \lambda_{k+1}}{A_k + a_{k+1}}\right) x^k + \left(\frac{A_k + \mu A_k \lambda_{k+1}}{A_k + a_{k+1}}\right) y^k, \tag{49}$$

$$a_{k+1} = \frac{(1 + 2\mu A_k)\lambda_{k+1} + \sqrt{(1 + 2\mu A_k)^2 \lambda_{k+1}^2 + 4(1 + \mu A_k) A_k \lambda_{k+1}}}{2}. \tag{50}$$

2) Let

$$A_{k+1} = A_k + a_{k+1}, \tag{51}$$

$$x^{k+1} = \left(\frac{1 + \mu A_k}{1 + \mu A_{k+1}}\right) x^k + \left(\frac{\mu a_{k+1}}{1 + \mu A_{k+1}}\right) y^{k+1} - \left(\frac{a_{k+1}}{1 + \mu A_{k+1}}\right) v^{k+1}. \tag{52}$$

3) Set $k = k + 1$ and go to step 1.

We now make a few remarks concerning Algorithm 2:

(i) By deleting the third inequality in (48) (the high-order large-step condition), we see that Algortihm 2 is a special instance of Algorithm 1. As a consequence, all results proved in Section 2 for Algorithm 1 also hold for Algorithm 2.

(ii) We mention that Algorithm 2 is a generalization of Algorithm 1 in [13] to strongly convex objectives. The authors of the latter work proved global $\mathcal{O}\left(k^{-\frac{3p+1}{2}}\right)$ and $\mathcal{O}\left(k^{-3p}\right)$ for function values and gradients/residuals, respectively. (see [13, Theorem 4].)

In what follows we will use remark (i) following Algorithm 2 to apply the results proved for Algorithm 1 in Section 2 to Algorithm 2.

The next two lemmas will be used to prove Theorem 3.3 below.

Lemma 3.1. *Consider the sequences evolved by* Algorithm 2 *and let* $d_0 := \|x^0 - x^*\|$, *where* x^* *is the (unique) solution of* (2). *Then, for all* $k \geq 1$,

$$\sum_{j=1}^{k} \frac{A_j}{\lambda_j^{\frac{p+1}{p-1}}} \leq \frac{d_0^2}{\theta^{\frac{2}{p-1}}(1-\sigma^2)}. \tag{53}$$

In particular, for all $k \geq 1$,

$$\lambda_k \geq C d_0^{-\frac{2(p-1)}{p+1}}, \qquad C := \lambda_1^{\frac{p-1}{p+1}} \theta^{\frac{2}{p+1}} (1-\sigma^2)^{\frac{p-1}{p+1}}. \tag{54}$$

Proof. Using (34) and third inequality in (48), we obtain

$$\left(\sum_{j=1}^{k} \frac{A_j}{\lambda_j^{\frac{p+1}{p-1}}}\right) \theta^{\frac{2}{p-1}} \leq \sum_{j=1}^{k} \frac{A_j}{\lambda_j^{\frac{p+1}{p-1}}} \left(\lambda_j \|y^j - \widetilde{x}^{j-1}\|^{p-1}\right)^{\frac{2}{p-1}} = \sum_{j=1}^{k} \frac{A_j}{\lambda_j} \|y^j - \widetilde{x}^{j-1}\|^2 \leq \frac{d_0^2}{1-\sigma^2},$$

which yields (53). To finish the proof of the lemma, note that (54) follows directly fom (53) and the fact that $A_k \geq \lambda_1$ for all $k \geq 1$ (see (7) and (11)). □

Lemma 3.2. *For all* $k \geq 0$,

$$A_{k+1} \geq \lambda_1 \left(1 + \frac{2\mu C}{d_0^{\frac{2(p-1)}{p+1}}} k^{\left(\frac{p-1}{p+1}\right)}\right)^k, \tag{55}$$

where $C > 0$ *is as in* (54).

Proof. First note that from (11) we have $A_1 = \lambda_1$, showing that (55) trivially holds for $k = 0$. Assume now that $k > 0$. From Lemma 3.1 we know, in particular, that

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

Since $A_j = A_{j-1} + a_j \geq A_{j-1} \geq \cdots \geq A_1$, for all $j \geq 2$, and $A_1 = \lambda_1$, we then obtain

$$\sum_{j=2}^{k+1} \frac{1}{\lambda_j^{\frac{p+1}{p-1}}} \leq \frac{d_0^2}{\lambda_1 \theta^{\frac{2}{p-1}}(1-\sigma^2)} =: c.$$

Now using Lemma A.1 with $c > 0$ as above, $q = \frac{p+1}{p-1}$ and $\lambda_j \leftarrow 2\mu\lambda_j$, we find

$$\begin{aligned}\prod_{j=2}^{k+1}(1+2\mu\lambda_j) &\geq \left(1+\left(\frac{(2\mu)^{\frac{p+1}{p-1}}}{c}k\right)^{\frac{p-1}{p+1}}\right)^k \\ &= \left(1+\frac{2\mu}{c^{\frac{p-1}{p+1}}}k^{\frac{p-1}{p+1}}\right)^k \\ &= \left(1+\left(\frac{2\mu\lambda_1^{\frac{p-1}{p+1}}\theta^{\frac{2}{p+1}}(1-\sigma^2)^{\frac{p-1}{p+1}}}{d_0^{\frac{2(p-1)}{p+1}}}\right)k^{\frac{p-1}{p+1}}\right)^k,\end{aligned}$$

which, in turn, combined with (43) and the definition of C in (54) finishes the proof of the lemma. □

Next is the main result on global convergence rates for Algorithm 2. As we mentioned before, it provides a global *superlinear* $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ convergence, where $p-1 \geq 1$ is the power in the high-order large-step condition (third inequality in (48)).

Theorem 3.3 (**Convergence rates for Algorithm 2).** *Consider the sequences evolved by* Algorithm 2*, let x^* denote the (unique) solution of* (2) *and let $C > 0$ be as in* (54)*. Then the following holds:*

(a) *For all $k \geq 0$,*

$$h(y^{k+1}) - h(x^*) \leq \frac{d_0^2}{2\lambda_1\left(1+\frac{2\mu C}{d_0^{\frac{2(p-1)}{p+1}}}k^{\left(\frac{p-1}{p+1}\right)}\right)^k} = \mathcal{O}\left(\frac{1}{k^{k\left(\frac{p-1}{p+1}\right)}}\right),$$

$$\max\left\{\|x^*-x^{k+1}\|^2, \|x^*-y^{k+1}\|^2\right\} \leq \frac{d_0^2}{\mu\lambda_1\left(1+\frac{2\mu C}{d_0^{\frac{2(p-1)}{p+1}}}k^{\left(\frac{p-1}{p+1}\right)}\right)^k} = \mathcal{O}\left(\frac{1}{k^{k\left(\frac{p-1}{p+1}\right)}}\right).$$

(b) *For all $k \geq 1$,*

$$\begin{cases} v^{k+1} \in \partial_{\varepsilon_{k+1}} f(y^{k+1}) + \partial g(y^{k+1}), \\ \|v^{k+1}\|^2 \leq \left(1+\sigma\sqrt{1+\mu C d_0^{-\frac{2(p-1)}{p+1}}}\right)^2 \frac{6d_0^{\frac{2(3p-1)}{p+1}}}{\mu C^2\lambda_1\left(1+\frac{2\mu C}{d_0^{\frac{2(p-1)}{p+1}}}(k-1)^{\left(\frac{p-1}{p+1}\right)}\right)^{k-1}} = \mathcal{O}\left(\frac{1}{(k-1)^{(k-1)\left(\frac{p-1}{p+1}\right)}}\right), \\ \varepsilon_{k+1} \leq \frac{3\sigma^2 d_0^{\frac{4p}{p+1}}}{\mu C\lambda_1\left(1+\frac{2\mu C}{d_0^{\frac{2(p-1)}{p+1}}}(k-1)^{\left(\frac{p-1}{p+1}\right)}\right)^{k-1}} = \mathcal{O}\left(\frac{1}{(k-1)^{(k-1)\left(\frac{p-1}{p+1}\right)}}\right). \end{cases}$$

Proof. Both items follow from Theorem 2.6 and Lemmas 3.1 and 3.2. To prove the inequalities in item (b), one also has to use the fact that the scalar function $t \mapsto \frac{1+\sigma\sqrt{1+\mu t}}{t}$ is nonincreasing as well as the lower bound on λ_k given in (54). □

4 Applications to accelerated high-order tensor methods for strongly convex objectives

In this section, we consider the problem

$$\min_{x\in\mathcal{H}} \{h(x) := f(x) + g(x)\}, \tag{56}$$

where $f, g : \mathcal{H} \to (-\infty, \infty]$ are proper, closed and convex functions, $\operatorname{dom} h \neq \emptyset$, and g is *μ-strongly convex* on $\mathcal{H}$ and $p \geq 2$ *times continuously differentiable* on $\Omega \supseteq \operatorname{Dom}(\partial f)$ with $D^p g(\cdot)$ being *L_p-Lipschitz continuous on* Ω: $0 < L_p < +\infty$ and

$$\|D^p g(x) - D^p g(y)\| \leq L_p \|x - y\|, \qquad \forall x, y \in \Omega. \tag{57}$$

Define

$$g_{x,p}(y) := g(x) + \sum_{k=1}^{p} \frac{1}{k!} D^k g(x)[y-x]^k + \frac{M}{(p+1)!}\|y-x\|^{p+1}, \qquad (x,y) \in \Omega \times \mathcal{H}, \tag{58}$$

where $M > 0$ is such that $M \geq pL_p$.

As observed by Nesterov in [20], the function $g_{x,p}(\cdot)$ is convex whenever $M \geq pL_p$ and, moreover,

$$\|\nabla g(y) - \nabla g_{x,p}(y)\| \leq \frac{L_p + M}{p!}\|y-x\|^p, \qquad \forall (x,y) \in \Omega \times \mathcal{H}. \tag{59}$$

At each iteration of the (exact) Proximal-Tensor method for solving (56) one has to find $y \in \mathcal{H}$ solving an inclusion of the form

$$0 \in \lambda\Big(\partial f(y) + \nabla g_{z,p}(y)\Big) + y - x, \tag{60}$$

where $z = P_\Omega(x)$ and $\lambda > 0$. Note also that (60) is equivalent to solving the convex problem

$$\min_{y\in\mathcal{H}} \left\{ f(y) + g_{z,p}(y) + \frac{1}{2\lambda}\|y-x\|^2 \right\}. \tag{61}$$

Next we introduce a notion *relative-error* inexact solution for (60) (or, equivalently, (61)). It will be used in step 2 (see (66)) of Algorithm 3.

Definition 4.1. *The triple $(y, u, \varepsilon) \in \mathcal{H} \times \mathcal{H} \times \mathbb{R}_+$ is a $\hat{\sigma}$-approximate* Tensor solution *of* (60) *at $(x, \lambda) \in \mathcal{H} \times \mathbb{R}_{++}$ if $\hat{\sigma} \geq 0$ and*

$$u \in \partial_\varepsilon f(y) + \nabla g_{z,p}(y), \qquad \frac{\|\lambda u + y - x\|^2}{1+\lambda\mu} + 2\lambda\varepsilon \leq \hat{\sigma}^2\|y-x\|^2, \tag{62}$$

where $z = P_\Omega(x)$.

Note that if $\hat{\sigma} = 0$ in (62), then it follows that $\varepsilon = 0$, $u \in \partial f(y) + \nabla g_{z,p}(y)$ and $\lambda u + y - x = 0$, which implies that y is the solution of (60). We also mention that if we set $\mu = 0$ in Definition 4.1 then we recover [11, Definition 2.1] (see also [13, Definition 1]).

Next proposition shows that $\hat{\sigma}$-approximate solutions of (60) provide relative-error appoximate solutions in the sense of (48).

Proposition 4.2. *Let (u, y, ε) be a $\hat{\sigma}$-approximate Tensor solution of* (60) *at $(x, \lambda) \in \mathcal{H} \times \mathbb{R}_{++}$ (in the sense of* Definition 4.1) *and define*

$$v = u - \nabla g_{z,p}(y) + \nabla g(y), \qquad \sigma = \frac{\lambda(L_p + M)\|y - x\|^{p-1}}{p!\sqrt{1 + \lambda\mu}} + \hat{\sigma}. \tag{63}$$

Then,

$$v \in \partial_\varepsilon f(y) + \nabla g(y), \qquad \frac{\|\lambda v + y - x\|^2}{1 + \lambda\mu} + 2\lambda\varepsilon \leq \sigma^2 \|y - x\|^2. \tag{64}$$

Proof. Note that the inclusion in (64) follows from the definition of v in (63) and the inclusion in (62). To prove the inequality in (64), note that from the definition of v in (63), the triangle inequality and property (59), we find

$$\begin{aligned}
\|\lambda v + y - x\|^2 &= \|\lambda u + y - x + \lambda\big(\nabla g(y) - \nabla g_{z,p}(y)\big)\|^2 \\
&\leq \Big(\|\lambda u + y - x\| + \lambda\|\nabla g(y) - \nabla g_{z,p}(y)\|\Big)^2 \\
&\leq \left(\|\lambda u + y - x\| + \frac{\lambda(L_p + M)}{p!}\|y - z\|^p\right)^2 \\
&\leq \left(\|\lambda u + y - x\| + \frac{\lambda(L_p + M)}{p!}\|y - x\|^p\right)^2,
\end{aligned}$$

where in the last inequality we also used the fact that $\|y - z\| \leq \|y - x\|$. (because $y \in \mathrm{Dom}(\partial_\varepsilon f) \subset \overline{\mathrm{Dom}(\partial f)} \subset \Omega$ and $z = P_\Omega(x)$.)

Hence,

$$\frac{\|\lambda v + y - x\|^2}{1 + \lambda\mu} + 2\lambda\varepsilon \leq \left(\frac{\|\lambda u + y - x\|}{\sqrt{1 + \lambda\mu}} + \frac{\lambda(L_p + M)}{p!\sqrt{1 + \lambda\mu}}\|y - x\|^p\right)^2 + 2\lambda\varepsilon.$$

Using now the elementar inequality $(a + b)^2 + c \leq \left(b + \sqrt{a^2 + c}\right)^2$ with $a = \|\lambda u + y - x\|/\sqrt{1 + \lambda\mu}$, $b = \lambda(L_p + M)\|y - x\|^p/(p!\sqrt{1 + \lambda\mu})$ and $c = 2\lambda\varepsilon$, we find

$$\begin{aligned}
\frac{\|\lambda v + y - x\|^2}{1 + \lambda\mu} + 2\lambda\varepsilon &\leq \left(\frac{\lambda(L_p + M)}{p!\sqrt{1 + \lambda\mu}}\|y - x\|^p + \sqrt{\frac{\|\lambda u + y - x\|^2}{1 + \lambda\mu} + 2\lambda\varepsilon}\right)^2 \\
&\leq \left(\frac{\lambda(L_p + M)}{p!\sqrt{1 + \lambda\mu}}\|y - x\|^p + \hat{\sigma}\|y - x\|\right)^2 \\
&= \left(\frac{\lambda(L_p + M)}{p!\sqrt{1 + \lambda\mu}}\|y - x\|^{p-1} + \hat{\sigma}\right)^2 \|y - x\|^2 \\
&= \sigma^2\|y - x\|^2,
\end{aligned}$$

where in the second inequality we used the inequality in (64) and in the identity we used the second equality (63). □

Next we present our p-th order inexact (relative-error) accelerated tensor algorithm for solving (56).

Algorithm 3. An accelerated inexact high-order tensor method for solving (56)

0) Choose $x^0, y^0 \in \mathcal{H}$ and $p \geq 2$, $\hat{\sigma} \geq 0$, $0 < \sigma_\ell < \sigma_u < 1$ such that

$$\sigma := \sigma_u + \hat{\sigma} < 1, \qquad \sigma_\ell (1+\hat{\sigma})^{p-1} < \sigma_u (1-\hat{\sigma})^{p-1}; \tag{65}$$

let $A_0 = 0$ and set $k = 0$.

1) Compute $\lambda_{k+1} > 0$ and a $\hat{\sigma}$-approximate Tensor solution $(u^{k+1}, y^{k+1}, \varepsilon_{k+1})$ (in the sense of Definition 4.1) of (60) at $(\widetilde{x}^k, \lambda_{k+1})$ satisfying

$$\frac{p!\,\sigma_\ell}{L_p + M} \leq \lambda_{k+1} \|y^{k+1} - \widetilde{x}^k\|^{p-1} \leq \frac{p!\,\sigma_u \sqrt{1+\lambda_{k+1}\mu}}{L_p + M}, \tag{66}$$

where

$$\widetilde{x}^k = \left(\frac{a_{k+1} - \mu A_k \lambda_{k+1}}{A_k + a_{k+1}}\right) x^k + \left(\frac{A_k + \mu A_k \lambda_{k+1}}{A_k + a_{k+1}}\right) y^k, \tag{67}$$

$$a_{k+1} = \frac{(1+2\mu A_k)\lambda_{k+1} + \sqrt{(1+2\mu A_k)^2 \lambda_{k+1}^2 + 4(1+\mu A_k) A_k \lambda_{k+1}}}{2}. \tag{68}$$

2) Let

$$A_{k+1} = A_k + a_{k+1}, \tag{69}$$

$$v^{k+1} = u^{k+1} - \nabla g_{z^k,p}(y^{k+1}) + \nabla g(y^{k+1}), \quad z^k = P_\Omega(\widetilde{x}^k), \tag{70}$$

$$x^{k+1} = \left(\frac{1+\mu A_k}{1+\mu A_{k+1}}\right) x^k + \left(\frac{\mu a_{k+1}}{1+\mu A_{k+1}}\right) y^{k+1} - \left(\frac{a_{k+1}}{1+\mu A_{k+1}}\right) v^{k+1}. \tag{71}$$

3) Set $k = k+1$ and go to step 1.

We now make two remarks concerning Algorithm 3:

(i) Algorithm 3 is a generalization of [13, Algorithm 3] for strongly convex problems. Global $\mathcal{O}\left(k^{-\frac{3p+1}{2}}\right)$ and $\mathcal{O}\left(k^{-3p}\right)$ convergence rates for function values and gradients/residuals, respectively, were proved in [13]. In contrast to this, here we obtained, see Theorem 4.4, the fast global $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ convergence rate.

(ii) We also mention that a $\hat{\sigma}$-approximate Tensor solution satisfying (66) can be computed using bisection schemes (see [2] and [11]).

Proposition 4.3. Algorithm 3 *is a special instance of* Algorithm 2 *for solving* (56)*, where*

$$\theta := \frac{p!\sigma_\ell}{L_p + M}. \tag{72}$$

Proof. It follows from the definitions of Algorithms 2 and 3 that we only have to prove that (48) holds. Note that the inclusion and the first inequality in (48) follow from step 2 of Algorithm 3 – the fact that $(u^{k+1}, y^{k+1}, \varepsilon_{k+1})$ is a $\hat{\sigma}$-approximate Tensor solution of (60)–, the second inequality in (66), the definition of σ in (65) and Proposition 4.2. To finish the proof of the proposition, note that the last inequality in (48) (the large-step condition) is a direct consequence of the first inequality in (66) and (72). □

Next theorem states the fast global $\mathcal{O}\left(k^{-k\left(\frac{p-1}{p+1}\right)}\right)$ convergence rate for Algorithm 3.

Theorem 4.4 (Convergence rates for Algorithm 3). *Consider the sequences generated by* Algorithm 3*, let $\theta > 0$ be as in* (72) *and let $C > 0$ be as in* (54)*, where $d_0 := \|x^0 - x^*\|$ and x^* is the (unique) solution of* (56)*.*

Then all the conclusions of Theorem 3.3 *hold.*

Proof. The proof follows from Proposition 4.3 and Theorem 3.3. □

5 Applications to first-order methods for strongly convex problems

Consider the convex optimization problem

$$\min_{x\in\mathcal{H}} \{h(x) := f(x) + g(x)\}, \tag{73}$$

where $f, g : \mathcal{H} \to (-\infty, \infty]$ are proper, closed and convex functions, $\operatorname{dom} h \neq \emptyset$, and, additionally, g is *μ-strongly convex* on $\mathcal{H}$ and *differentiable* on $\Omega \supseteq \operatorname{dom} f$ with ∇g being *L-Lipschitz continuous on* Ω

An iteration of the proximal-gradient (forward-backward) method for solving (73) can be written as follows:

$$y = (\lambda\partial f + I)^{-1}(x - \lambda\nabla g(z)), \tag{74}$$

where $z = P_\Omega(x)$ and $\lambda > 0$. Using the definition of $(\lambda\partial f + I)^{-1}$, it is easy to see that (74) is equivalent to solving the inclusion

$$0 \in \lambda\Big(\partial f(y) + \nabla g(z)\Big) + y - x. \tag{75}$$

Next we define a notion of approximate solution for (75) within a *relative-error* criterion.

Definition 5.1. *The triple* $(y, u, \varepsilon) \in \mathcal{H} \times \mathcal{H} \times \mathbb{R}_+$ *is a* $\hat{\sigma}$*-approximate Proximal-Gradient (PG) solution of* (75) *at* $(x, \lambda) \in \mathcal{H} \times \mathbb{R}_{++}$ *if* $\hat{\sigma} \geq 0$ *and*

$$u \in \partial_\varepsilon f(y), \qquad \frac{\|\lambda\,(u + \nabla g(z)) + y - x\|^2}{1 + \lambda\mu} + 2\lambda\varepsilon \leq \hat{\sigma}^2 \|y - x\|^2, \tag{76}$$

where $z = P_\Omega(x)$*. We also write*

$$(y, u, \varepsilon) \approx (\lambda \partial f + I)^{-1}\,(x - \lambda \nabla g(z))$$

to mean that (y, u, ε) *is a* $\hat{\sigma}$*-approximate PG solution of* (75) *at* (x, λ)*.*

Note that if $\hat{\sigma} = 0$ in (76), then it follows that $\varepsilon = 0$, $u \in \partial f(y)$ and $\lambda\,[u + \nabla g(z)] + y - x = 0$, which implies that y is the (exact) solution of (75). In particular, in this case, y satisfies (74).

Proposition 5.2. *Let* (u, y, ε) *be a* $\hat{\sigma}$*-approximate PG solution of* (75) *at* $(x, \lambda) \in \mathcal{H} \times \mathbb{R}_{++}$ *as in* Definition 5.1 *and define*

$$v = u + \nabla g(y), \qquad \sigma = \frac{\lambda L}{\sqrt{1 + \lambda\mu}} + \hat{\sigma}. \tag{77}$$

Then,

$$v \in \partial_\varepsilon f(y) + \nabla g(y), \qquad \frac{\|\lambda v + y - x\|^2}{1 + \lambda\mu} + 2\lambda\varepsilon \leq \sigma^2 \|y - x\|^2. \tag{78}$$

Proof. The proof follows the same outline of Proposition 4.2's proof. □

For solving (73), we propose the following inexact (relative-error) accelerated first-order algorithm.

Algorithm 4. An accelerated inexact proximal-gradient algorithm for solving (73)

0) Choose $x^0, y^0 \in \mathcal{H}$ and $\hat{\sigma} \geq 0$, $0 < \sigma_u \leq 1$ such that $\sigma := \sigma_u + \hat{\sigma} < 1$ and let

$$\lambda = \frac{\sigma_u}{\sqrt{\left(\frac{\sigma_u \mu}{2}\right)^2 + L^2} - \frac{\sigma_u \mu}{2}} > \frac{\sigma_u}{L}; \tag{79}$$

let $A_0 = 0$ and set $k = 0$.

1) Compute $z^k = P_\Omega(\widetilde{x}^k)$ and

$$(y^{k+1}, u^{k+1}, \varepsilon_{k+1}) \approx (\lambda \partial f + I)^{-1} \left(\widetilde{x}^k - \lambda \nabla g(z^k)\right), \tag{80}$$

i.e., compute a $\hat{\sigma}$-approximate PG solution $(u^{k+1}, y^{k+1}, \varepsilon_{k+1})$ at $(\widetilde{x}^k, \lambda)$ (in the sense of Definition 5.1), where

$$\widetilde{x}^k = \left(\frac{a_{k+1} - \mu A_k \lambda}{A_k + a_{k+1}}\right) x^k + \left(\frac{A_k + \mu A_k \lambda}{A_k + a_{k+1}}\right) y^k, \tag{81}$$

$$a_{k+1} = \frac{(1 + 2\mu A_k)\lambda + \sqrt{(1 + 2\mu A_k)^2 \lambda^2 + 4(1 + \mu A_k) A_k \lambda}}{2}. \tag{82}$$

2) Let

$$A_{k+1} = A_k + a_{k+1}, \tag{83}$$

$$v^{k+1} = u^{k+1} + \nabla g(y^{k+1}), \tag{84}$$

$$x^{k+1} = \left(\frac{1 + \mu A_k}{1 + \mu A_{k+1}}\right) x^k + \left(\frac{\mu a_{k+1}}{1 + \mu A_{k+1}}\right) y^{k+1} - \left(\frac{a_{k+1}}{1 + \mu A_{k+1}}\right) v^{k+1}. \tag{85}$$

3) Set $k = k + 1$ and go to step 1.

We now make the following remark concerning Algorithm 4:

(i) From the definition of $\lambda > 0$ in (79) we obtain

$$\frac{\lambda^2 L^2}{1 + \lambda \mu} = \sigma_u^2. \tag{86}$$

Indeed, it is easy to check that $\lambda > 0$ is the largest root of $L^2 \lambda^2 - (\sigma_u^2 \mu)\lambda - \sigma_u^2 = 0$, which is clearly equivalent to (86). Now using (86), we find

$$\sigma = \sigma_u + \hat{\sigma} = \frac{\lambda L}{\sqrt{1 + \lambda \mu}} + \hat{\sigma}. \tag{87}$$

Next proposition shows that Algorithm 4 is a special instance of Algorithm 1 for solving (73).

Proposition 5.3. *Consider the sequences evolved by* Algorithm 4 *and let* $\lambda_{k+1} \equiv \lambda$. *Then,* $\lambda_{k+1} > 0$ *and the triple* $(y^{k+1}, v^{k+1}, \varepsilon_{k+1})$ *satisfy condition* (4) *in* Algorithm 1 *with* $\sigma = \hat{\sigma} + \sigma_u$. *As a consequence,* Algorithm 4 *is a special instance of* Algorithm 1 *for solving* (73).

Proof. The proof follows from (80), (87), Proposition 5.2 and the definitions of Algorithms 1 and 4. □

Next we summarize the results on *linear convergence* rates for Algorithm 4.

Theorem 5.4 (**Convergence rates for Algorithm 4**). *Consider the sequences evolved by* Algorithm 4 *and let* $\sigma = \hat{\sigma} + \sigma_u$. *Let also* x^* *be the unique solution of* (73), *let* d_0 *be as in* (33) *and denote* $\gamma = \sqrt{(1+\sigma_u)^{-1}\sigma_u}$. *The following holds:*

(a) *For all* $k \geq 1$,

$$h(y^k) - h(x^*) \leq \frac{L d_0^2}{2\sigma_u}\left(1 - \gamma\sqrt{\frac{\mu}{L}}\right)^{k-1},$$

$$\max\left\{\|x^* - y^k\|, \|x^* - x^k\|\right\} \leq \sqrt{\frac{L}{\sigma_u \mu}}\, d_0 \left(1 - \gamma\sqrt{\frac{\mu}{L}}\right)^{(k-1)/2}.$$

(b) *For all* $k \geq 1$,

$$\begin{cases} v^{k+1} \in \partial_{\varepsilon_{k+1}} f(y^{k+1}) + \partial g(y^{k+1}), \\ \|v^{k+1}\| \leq \dfrac{6 d_0 L^{3/2}}{\mu^{1/2}\sigma_u^{3/2}}\left(1 + \sigma\sqrt{1 + \frac{\sigma_u \mu}{L}}\right)\left(1 - \gamma\sqrt{\frac{\mu}{L}}\right)^{(k-1)/2}, \\ \varepsilon_{k+1} \leq \dfrac{3\sigma^2 d_0^2 L^2}{\sigma_u^2 \mu}\left(1 - \gamma\sqrt{\frac{\mu}{L}}\right)^{k-1}. \end{cases}$$

Proof. (a) First note that simple computations using (46) with $\underline{\lambda} = \lambda$, the inequality in (79), the definition of $\gamma > 0$ and the fact that $L \geq \mu$ show that

$$\alpha > \sqrt{(1+\sigma_u)^{-1}\sigma_u}\sqrt{\frac{\mu}{L}} = \gamma\sqrt{\frac{\mu}{L}}, \qquad \lambda > \frac{\sigma_u}{L}, \tag{88}$$

which combined with Proposition 5.3 and Theorem 2.9(a) gives the proof of (a).

(b) The result follows from (88), Proposition 5.3 and Theorem 2.9(b). □

Acknowledgments

The author would like to thank Dr. Benar F. Svaiter for the fruitful discussions related to the first draft of this work.

A Some auxiliary results

Lemma A.1. *For all $k \geq 1$, the optimal value of the minimization problem, over $\lambda_1, \ldots, \lambda_k > 0$,*

$$\begin{aligned} &\min \prod_{j=1}^{k} (1+\lambda_j) \\ &\text{s.t.} \ \sum_{j=1}^{k} \frac{1}{\lambda_j^q} \leq c, \end{aligned} \tag{89}$$

where $c > 0$ and $q \geq 1$, is given by

$$\left(1 + \left(\frac{k}{c}\right)^{1/q}\right)^k.$$

Proof. First consider the convex problem

$$\begin{aligned} &\min \sum_{j=1}^{k} \log(1+e^{t_j}) \\ &\text{s.t.} \ \sum_{j=1}^{k} \frac{1}{e^{qt_j}} \leq c. \end{aligned} \tag{90}$$

Since the objective and constraint functions in (90) are convex and invariant under permutations on $(t_1, \ldots, t_k)$, it follows that one of its solutions takes the form $(t, \ldots, t)$. It is also clear that at any solution the inequality in (90) must hold as an equality. Hence, $k\frac{1}{e^{qt}} = c$, i.e., $e^t = \left(\frac{k}{c}\right)^{1/q}$. As a consequence, for all $(t_1, \ldots, t_k)$ such that $\sum_{j=1}^{k} \frac{1}{e^{qt_j}} \leq c$,

$$\sum_{j=1}^{k} \log(1+e^{t_j}) \geq k \log(1+e^t) = k \log\left(1 + \left(\frac{k}{c}\right)^{1/q}\right) = \log\left(\left(1 + \left(\frac{k}{c}\right)^{1/q}\right)^k\right). \tag{91}$$

Now let $\lambda_1, \ldots, \lambda_k > 0$ be such that $\sum_{j=1}^{k} \frac{1}{\lambda_j^q} \leq c$ and define $t_j := \log(\lambda_j)$, for $j \in \{1, \ldots, k\}$. Then, since in this case $\sum_{j=1}^{k} \frac{1}{e^{qt_j}} \leq c$, using (91) and some basic properties of logarithms we find

$$\begin{aligned} \prod_{j=1}^{k} (1+\lambda_j) = \prod_{j=1}^{k} (1+e^{t_j}) &= e^{\log\left(\prod_{j=1}^{k} (1+e^{t_j})\right)} \\ &= e^{\sum_{j=1}^{k} \log(1+e^{t_j})} \\ &\geq e^{\log\left(\left(1+\left(\frac{k}{c}\right)^{1/q}\right)^k\right)} \\ &= \left(1 + \left(\frac{k}{c}\right)^{1/q}\right)^k, \end{aligned}$$

which concludes the proof of the lemma. □

Lemma A.2. *The following holds for $q(\cdot)$ defined by*

$$q(x) = \langle v, x - y\rangle + \frac{\mu}{2}\|x - y\|^2 - \varepsilon + \frac{1}{2\lambda}\|x - z\|^2 \qquad (x \in \mathcal{H}) \tag{92}$$

where $v, y, z \in \mathcal{H}$ and $\mu, \varepsilon, \lambda > 0$.

(a) *The (unique) global minimizer of $q(\cdot)$ is given by*

$$x^* = \frac{1}{1+\lambda\mu}z + \frac{\lambda\mu}{1+\lambda\mu}y - \frac{\lambda}{1+\lambda\mu}v.$$

(b) *We have,*

$$\min_x q(x) = \frac{1}{2\lambda}\left[\|y - z\|^2 - \left(\frac{\|\lambda v + y - z\|^2}{1+\lambda\mu} + 2\lambda\varepsilon\right)\right].$$

(c) *We have,*

$$q(x) = \frac{1}{2\lambda}\left[\|y - z\|^2 - \left(\frac{\|\lambda v + y - z\|^2}{1+\lambda\mu} + 2\lambda\varepsilon\right)\right] + \frac{1+\lambda\mu}{2\lambda}\|x - x^*\|^2, \qquad \forall x \in \mathcal{H}.$$

Proof. (a) This follows directly from (92) and some simple calculus.

(b) Note first that

$$\min_x q(x) = q(x^*) = \langle v, x^* - y\rangle + \frac{\mu}{2}\|x^* - y\|^2 - \varepsilon + \frac{1}{2\lambda}\|x^* - z\|^2. \tag{93}$$

Using the well-known identity $a\|z\|^2 + b\|w\|^2 = \frac{1}{a+b}\left[\|az + bw\|^2 + ab\|z - w\|^2\right]$ with $a = \mu$, $b = 1/\lambda$, $z = x^* - y$ and $w = x^* - z$, and (a) we find

$$\begin{aligned}\mu\|x^* - y\|^2 + \frac{1}{\lambda}\|x^* - z\|^2 &= \frac{\lambda}{1+\lambda\mu}\left[\left\|\underbrace{\frac{1+\lambda\mu}{\lambda}x^* - \mu y - \frac{1}{\lambda}z}_{-v}\right\|^2 + \frac{\mu}{\lambda}\|z - y\|^2\right] \\ &= \frac{\lambda}{1+\lambda\mu}\left[\|v\|^2 + \frac{\mu}{\lambda}\|z - y\|^2\right]. \end{aligned} \tag{94}$$

On the other hand, we also have $x^* - y = \frac{1}{1+\lambda\mu}(z - y) - \frac{\lambda}{1+\lambda\mu}v$, which in turn gives

$$\langle v, x^* - y\rangle = \frac{1}{1+\lambda\mu}\left[\langle v, z - y\rangle - \lambda\|v\|^2\right]. \tag{95}$$

Direct use of (93), (94) and (95) yields

$$\begin{aligned}\min_x q(x) + \varepsilon &= \frac{1}{1+\lambda\mu}\left[\langle v, z - y\rangle - \lambda\|v\|^2\right] + \frac{\lambda}{2(1+\lambda\mu)}\left[\|v\|^2 + \frac{\mu}{\lambda}\|z - y\|^2\right] \\ &= \frac{1}{2\lambda(1+\lambda\mu)}\left[2\langle\lambda v, z - y\rangle - \|\lambda v\|^2 + \lambda\mu\|z - y\|^2\right] \\ &= \frac{1}{2\lambda(1+\lambda\mu)}\left[(1+\lambda\mu)\|y - z\|^2 - \|\lambda v + y - z\|^2\right] \\ &= \frac{1}{2\lambda}\left[\|y - z\|^2 - \frac{\|\lambda v + y - z\|^2}{1+\lambda\mu}\right], \end{aligned}$$

which then yields

$$\begin{aligned}\min_x q(x) &= \frac{1}{2\lambda}\left[\|y-z\|^2 - \frac{\|\lambda v + y - z\|^2}{1+\lambda\mu}\right] - \varepsilon \\ &= \frac{1}{2\lambda}\left[\|y-z\|^2 - \left(\frac{\|\lambda v + y - z\|^2}{1+\lambda\mu} + 2\lambda\varepsilon\right)\right].\end{aligned}$$

(c) This follows from (b) and Taylor's theorem applied to $q(\cdot)$. □

References

[1] M. M. Alves, R. D. C. Monteiro, and B. F. Svaiter. Regularized HPE-type methods for solving monotone inclusions with improved pointwise iteration-complexity bounds. *SIAM J. Optim.*, 26(4):2730–2743, 2016.

[2] M. M. Alves, R.D.C. Monteiro, and B.F. Svaiter. Primal-dual regularized SQP and SQCQP type methods for convex programming and their complexity analysis. Technical report (optimization-online 4353), Sep, 2014.

[3] Y. Arjevani, O. Shamir, and R. Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Math. Program.*, 178(1-2, Ser. A):327–360, 2019.

[4] H. Attouch, M. M. Alves, and B. F. Svaiter. A dynamic approach to a proximal-Newton method for monotone inclusions in Hilbert spaces, with complexity $\mathcal{O}(1/n^2)$. *J. Convex Anal.*, 23(1):139–180, 2016.

[5] M. Barré, A. Taylor, and F. Bach. Principled analyses and design of first-order methods with inexact proximal operators. Technical report (arxiv preprint arxiv:2006.06041), Jun, 2020.

[6] S. Bubeck, Q. Jiang, Y. Lee, Y. Yuanzhil, and A. Sidford. Near-optimal method for highly smooth convex optimization. *Proceedings of Machine Learning Research*, 99:1–16, 2019.

[7] N. Doikov and Y. Nesterov. Local convergence of tensor methods. Technical report (arxiv:1912.02516v2 [math.oc]), Dec, 2019.

[8] J. Eckstein and W. Yao. Relative-error approximate versions of Douglas-Rachford splitting and special cases of the ADMM. *Math. Program.*, 170(2, Ser. A):417–444, 2018.

[9] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and C. Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. *Proceedings of Machine Learning Research*, 99:1–18, 2019.

[10] G. N. Grapiglia and Y. Nesterov. Tensor methods for finding approximate stationary points of convex functions. *Optimization Methods and Software*, pages 1–34, 2020.

[11] B. Jiang, H. Wang, and S. Zhang. An optimal high-order tensor method for convex optimization. Technical report (arxiv:1812.06557 [math.oc]), Apr, 2020.

[12] G. Kornowski and O. Shamir. High-order oracle complexity of smooth and strongly convex optimization. Technical report (arxiv:2010.06642v1 [math.oc]), Oct, 2020.

[13] T. Lin and M. I. Jordan. A control-theoretic perspective on optimal high-order optimization. Technical report (arxiv:1912.07168 [math.oc]), Dec 2019.

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

[15] 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.

[16] 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 J. Optim.*, 20(6):2755–2787, 2010.

[17] R. D. C. Monteiro and B. F. Svaiter. Iteration-Complexity of a Newton Proximal Extragradient Method for Monotone Variational Inequalities and Inclusion Problems. *SIAM J. Optim.*, 22(3):914–935, 2012.

[18] R. D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM J. Optim.*, 23(2):1092–1125, 2013.

[19] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[20] Y. Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 2019.

[21] Y. Nesterov. Inexact accelerated high-order proximal-point methods. Technical report, Feb, 2020.

[22] Y. Nesterov. Inexact accelerated high-order proximal-point methods with auxiliary search procedure. Technical report, Feb, 2020.

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

[24] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

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

[26] 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.

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

[28] 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.

[29] 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.