Noname manuscript No.
(will be inserted by the editor)

A proximal-Newton method for unconstrained convex optimization in Hilbert spaces

M. Marques Alves · Benar F. Svaiter

Received: date / Accepted: date

Abstract We propose and study the iteration-complexity of a proximal-Newton method for finding approximate solutions of the problem of minimizing a twice continuously differentiable convex function on a (possibly infinite dimensional) Hilbert space. We prove global convergence rates for obtaining approximate solutions in terms of function/gradient values. Our main results follow from an iteration-complexity study of an (large-step) inexact proximal point method for solving convex minimization problems.

2000 Mathematics Subject Classification: 90C25, 90C30, 47H05.

Keywords Smooth convex optimization · Proximal-Newton method · Complexity · Proximal point methods

1 Introduction

In this paper we consider optimization problems of minimizing twice continuously differentiable convex functions on (possibly infinite dimensional) real Hilbert spaces. These problems appear in different fields of applied sciences and have been subject of intense research in the communities of numerical analysis and optimization. One of the most important numerical methods for finding approximate solutions of unconstrained optimization problems is the Newton method. In its simplest form, it depends, in each step, on the solution of a quadratic local model of the objective function which, on the other hand, leads to the solution of linear systems of equations defined by the Hessian of the objective function [2]. Since the Hessian can be degenerate at the current step, many different modifications of the Newton method have been proposed in the literature to ensure well-definedness and convergence (see, e.g., [8] for a discussion). In this work, we focus on the global performance of proximal-Newton methods, which are Newton-type methods where a proximal quadratic regularization term is added to the quadratic local model which has to be minimized in each iteration. Like in the Newton method, it leads, in each step, to the solution of linear systems, but now with an

The work of M. Marques Alves was partially supported by CNPq grants no. 306317/2014-1 and 406250/2013-8. The work of Benar F. Svaiter was partially supported by CNPq grants no. 474996/2013-1, 302962/2011-5, FAPERJ grant E-26/102.940/2011, and PRONEX-Optimization.

M. Marques Alves
Departamento de Matemática, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 88040-900.
Tel.: +55 48 37213678
E-mail: maicon.alves@ufsc.br

Benar F. Svaiter
IMPA, Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, Brazil (benar@impa.br).

additional regularization parameter which, in particular, guarantees the well-definedness of the method. Since proximal-Newton methods have a proximal nature, they can be analyzed in the setting of proximal point methods for optimization, which we briefly discuss in what follows.

The proximal point method is a classical scheme for solving monotone inclusion problems with point-to-set monotone operators, proposed by Martinet [3] and further developed by Rockafellar [9], which uses the current iteration to construct a regularized version of the original problem (the proximal subproblem) whose the solution is taken as the next iteration. In contrast to the Rockafellar's approach which relies on a summable error condition for solving each subproblem, the hybrid proximal extragradient (HPE) method of Solodov and Svaiter [11] requires for its convergence each proximal subproblem to be solved within a relative error condition.

In this paper we combine ideas from the HPE-theory and classical Newton method to propose a proximal-Newton method with global performance for solving smooth convex optimization problems. With this in mind, we first propose and study an inexact under-relaxed proximal point method (Algorithm 1) for (nonsmooth) convex optimization, which shares similar proprieties with the method recently proposed in [1], and after that we show how to obtain a proximal-Newton method as a special case. We obtain bounds on the iteration-complexity to find approximate solutions in terms of function/gradient values.

Contents. Section 2 presents an inexact under-relaxed proximal point method for convex optimization and its iteration-complexity analysis. Section 3 presents a proximal-Newton method for smooth convex optimization and discusses its main proprieties. Section 4 contains the main contributions of the paper, namely, the iteration-complexity of the method proposed in Section 3. Finally, the appendix contains the proofs of some results in Section 2 and an auxiliary lemma.

Notation. We denote by $\mathcal{H}$ a real Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and induced norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. The extended real line is denoted by $\overline{\mathbb{R}} := \mathbb{R}\cup\{-\infty,+\infty\}$ and we also use the notation $\log^+(t) = \max\{\log(t), 0\}$. Moreover, we use the standard notation and definitions of convex analysis for convex functions, subdifferentials, ε-subdifferentials, etc.

2 An inexact proximal point method for convex optimization

In this section we consider the minimization problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \mathcal{H}, \end{aligned} \tag{1}$$

where $f : \mathcal{H} \to \overline{\mathbb{R}}$ is a proper closed convex function. We also assume that the solution set of (1) is nonempty. One of the most important numerical schemes for finding approximate solutions of (1) is the *proximal point method* (PPM) [3,10]. For a given starting point $x_0 \in \mathcal{H}$, the PPM defines a sequence $\{x_k\}_{k\geq 1}$ of approximations to the solution of (1) according to

$$x_k \approx \operatorname*{argmin}_{x\in\mathcal{H}} f(x) + \frac{1}{2\lambda_k}\|x - x_{k-1}\|^2, \tag{2}$$

where $\lambda_k > 0$ is a sequence of stepsizes and x_{k-1} is the current iterate. Equivalently, the sequence $\{x_k\}_{k\geq 1}$ can be defined as an (approximate) solution of the monotone inclusion

$$0 \in \lambda_k \partial f(x) + x - x_{k-1}. \tag{3}$$

Whenever (2) is solved exactly then weak convergence of $\{x_k\}_{k\geq 1}$ to a solution of (1) is guaranteed under the assumption that $\lambda_k > 0$ is bounded away from zero [10]. Still under the latter condition on the sequence of stepsizes, weak convergence of the inexact PPM can

also be obtained under summable error criterion [10]: if, for all $k \geq 1$, $x_k^* \in \mathcal{H}$ is the exact solution of (2) (or (3)) then $\{x_k\}_{k\geq 1}$ must satisfy

$$\sum_{k=1}^{\infty} \|x_k - x_k^*\| < \infty. \tag{4}$$

Starting with [11,12], the last two decades have seen an intense research activity in the study and development of proximal point methods (for the more general problem of inclusions with monotone operators) which use relative error tolerance for finding approximate solutions of (3). Among these new methods, the *hybrid proximal extragradient* (HPE) *method* [11] has been shown to serve as a framework for the design and analysis of several first- and second-order methods in optimization, variational inequalities, saddle-point problems, etc. (see, e.g., [4–7].) A variant of the HPE method suitable for the analysis of second-order methods, called *large-step HPE method*, was proposed and studied in [7]. As one of its distinctive features, the latter method forces at each iteration a *large-step condition*, which plays a crucial role in obtaining superior rates of convergence (see, e.g., [6,7]).

An iteration of the large-step HPE method for solving, in particular, the monotone inclusion problem $0 \in \partial f(x)$ – which is clearly equivalent to (1) – consists of

$$\begin{aligned} &v_k \in (\partial f)^{\varepsilon_k}(y_k), \qquad \|\lambda_k v_k + y_k - x_{k-1}\|^2 + 2\lambda_k \varepsilon_k \leq \sigma^2 \|y_k - x_{k-1}\|^2, \\ &\lambda_k \|y_k - x_{k-1}\| \geq \eta > 0, \\ &x_k = x_{k-1} - \lambda_k v_k. \end{aligned} \tag{5}$$

Here $\sigma \in [0,1[$ is a relative error tolerance and $(\partial f)^{\varepsilon_k}$ denotes the ε_k-enlargement of ∂f (it has the property that $\partial_{\varepsilon_k} f(y) \subset (\partial f)^{\varepsilon_k}(y)$ for all $y \in \mathcal{H}$). Moreover, the second inequality in (5) is the large-step condition that we mentioned before. It should be emphasized that (a) if $\sigma = 0$, then the scheme in (5) reduces to the exact PPM, i.e., in this case y_k is the exact solution of (3), (b) the new iterate in (5) is defined as an extragradient step departing from x_{k-1}, (c) the large-step HPE method (5) has the same asymptotic behavior of the exact and inexact (under the condition (4)) PPM, namely, it converges weakly to a solution of (1), whenever there exists at least one [11]. Moreover, it has global rates of convergence [7]: (i) pointwise of order $(\|v_k\|, \varepsilon_k) = (\mathcal{O}(1/k), \mathcal{O}(1/k^{3/2})$ and (ii) ergodic of order $\max\{\|\bar{v}_k\|, \bar{\varepsilon}_k\} = \mathcal{O}(1/k^{3/2})$, where $\bar{v}_k$ and $\bar{\varepsilon}_k$ are constructed from all previous generated v_k and ε_k satisfying (5), (d) the analysis of [7] does not include rates of convergence for the scheme (5) in terms of function values of f, since [7] considers the more general problem of inclusion problems for maximal monotone operators.

In this section we will study global rates of convergence (iteration-complexity) in terms of both objective function values and $(\|v_k\|, \varepsilon_k)$ for the following variant of the large-step HPE method (5).

Algorithm 1 An inexact under-relaxed proximal point method for convex optimization

0) Let $x_0 \in \mathrm{dom}(f)$, $0 \le \sigma < 1$, $\eta > 0$, $0 < \tau \le 1$, $\lambda_1 > 0$ and set $k = 1$;
1) compute $(y_k, v_k, \varepsilon_k) \in \mathcal{H} \times \mathcal{H} \times \mathbb{R}_+$ satisfying

$$\begin{aligned} &v_k \in \partial_{\varepsilon_k} f(y_k), \\ &\|\lambda_k v_k + y_k - x_{k-1}\|^2 + 2\lambda_k \varepsilon_k \le \sigma^2 \|y_k - x_{k-1}\|^2, \\ &\lambda_k \|y_k - x_{k-1}\| \ge \eta \ \text{ or } \ v_k = 0\,; \end{aligned} \tag{6}$$

2) if $v_k = 0$, then **stop** and output y_k;
3) otherwise, choose $\lambda_{k+1} > 0$, $\tau_k \in [\tau, 1]$, and set

$$x_k = (1 - \tau_k) x_{k-1} + \tau_k y_k; \tag{7}$$

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

Remarks. 1) Algorithm 1 is an under-relaxed large-step proximal point method with relative error tolerance $\sigma > 0$; 2) the main difference between (5) and (6)–(7) is the inclusion, the one in (6) is stronger than the one in (5), and in the definition of the new iterate x_k. Rather than an extragradient step from x_{k-1}, Algorithm 1 defines the new iterate x_k in (7) as a convex combination between x_{k-1} and y_k; 3) in what follows, we will show how these distinctive features of Algorithm 1 pointed out in the previous remark allows one to prove superior convergence rates (when compared to the large-step HPE method) as well as convergence rates in terms of function values; 4) if we set $\tau = 1$ in Algorithm 1, then it reduces to Algorithm 1 in [1], for which iteration-complexity analysis was studied in the latter reference; 5) Since $x_0 \in \mathrm{dom}(f)$, an induction argument together with the inclusion in (6) shows that the same holds for every x_k and y_k generated by Algorithm 1.

From now one in this section we assume (w.l.o.g.) that Algorithm 1 generates infinite sequences $\{\lambda_k\}$, $\{y_k\}$, $\{v_k\}$, etc.

The following lemma is a direct consequence of the first inequality in (6) and the triangle inequality.

Lemma 2.1 *For every $k \ge 1$,*

$$(1 + \sigma)\|y_k - x_{k-1}\| \ge \|\lambda_k v_k\| \ge (1 - \sigma)\|y_k - x_{k-1}\|.$$

We mention that if $v_k = 0$ in step 2, then, in particular, it follows from Lemma 2.1 and (6) that $0 \in \partial f(y_k)$, i.e., y_k is a solution of (1).

Since the proofs of the next results follow the same outline of the ones in [1], we have included it in Appendix A.2.

Next proposition shows how does the function values decrease in Algorithm 1.

Proposition 2.1 *For every $k \ge 1$,*

$$\begin{aligned} f(x_{k-1}) - \max\{f(x_k), f(y_k)\} &\ge \tau \max\left\{ \sqrt{\eta(1-\sigma)}\|v_k\|^{3/2}, \frac{(1-\sigma^2)}{2\lambda_k}\|y_k - x_{k-1}\|^2 \right\}, \\ \|y_k - x_{k-1}\|^2 &\ge \frac{2\lambda_k \varepsilon_k}{\sigma^2}. \end{aligned} \tag{8}$$

Remark. From the first inequality in (8) and the assumption that $x_0 \in \mathrm{dom}(f)$ we have

$$\infty > f(x_0) \ge f(x_k) + \frac{\tau(1-\sigma^2)}{2} \sum_{j=1}^{k} \frac{\|y_j - x_{j-1}\|^2}{\lambda_j} \quad \forall k \ge 1. \tag{9}$$

As a consequence of the above inequality, we conclude that the sequence $\{\lambda_k^{-1}\|y_k - x_{k-1}\|^2\}$ converges to zero as $k \to \infty$, which combined with the large-step condition, i.e., the second inequality in (6), proves that $\lambda_k \to \infty$ as $k \to \infty$.

Let $\mathcal{D}_0$ denote the diameter of the level set $[f \leq f(x_0)]$, that is,

$$\mathcal{D}_0 = \sup\{\|x - y\| \ : \ \max\{f(x), f(y)\} \leq f(x_0)\}. \tag{10}$$

Lemma 2.2 *Assume that $0 < \mathcal{D}_0 < \infty$, let $\overline{x}$ be a solution of (1) and define*

$$\mathcal{D} := \frac{\tau\sqrt{\eta(1-\sigma)}}{\mathcal{D}_0^{3/2}\left(1 + \sigma^2/[2(1-\sigma)]\right)^{3/2}}. \tag{11}$$

Then, for all $k \geq 1$:

$$f(x_{k-1}) + (1-\tau)\mathcal{D}\left[f(x_{k-1}) - f(\overline{x})\right]^{3/2} \geq f(x_k) + \mathcal{D}\left[f(x_k) - f(\overline{x})\right]^{3/2}. \tag{12}$$

From now on in this section we will assume that x_0 is not a solution of (1).

The next proposition shows the first global rate of convergence for Algorithm 1 (in terms of objective function values and $(\|v_k\|, \varepsilon_k)$).

Theorem 2.1 *Assume that $0 < \mathcal{D}_0 < \infty$, let $\mathcal{D} > 0$ be defined in (11), let $\overline{x} \in \mathcal{H}$ be a solution of (1) and define*

$$\widehat{\mathcal{D}} := \frac{\tau\mathcal{D}}{2 + 3\mathcal{D}\sqrt{f(x_0) - f(\overline{x})}}. \tag{13}$$

Then, for every $k \geq 1$:

$$f(x_k) - f(\overline{x}) \leq \frac{f(x_0) - f(\overline{x})}{\left[1 + k\,\widehat{\mathcal{D}}\sqrt{f(x_0) - f(\overline{x})}\right]^2} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{14}$$

Moreover, for each $k \geq 2$ even, there exists $j \in \{k/2 + 1, \ldots, k\}$ such that

$$\|v_j\| \leq \frac{4}{\sqrt[3]{\eta(1-\sigma)}}\left(\frac{f(x_0) - f(\overline{x})}{\tau k\left[2 + k\,\widehat{\mathcal{D}}\sqrt{f(x_0) - f(\overline{x})}\right]^2}\right)^{2/3} = \mathcal{O}\left(\frac{1}{k^2}\right), \tag{15}$$

$$\varepsilon_j \leq \frac{4\sigma^2\left[f(x_0) - f(\overline{x})\right]}{\tau(1-\sigma^2)k\left[2 + k\widehat{\mathcal{D}}\sqrt{f(x_0) - f(\overline{x})}\right]^2} = \mathcal{O}\left(\frac{1}{k^3}\right). \tag{16}$$

Next we present a similar result to Theorem 2.1 under the assumption that $\varepsilon_k = 0$ for all $k \geq 1$ in Algorithm 1. To this end, first define (cf. (11) and (13))

$$\mathcal{E} := \frac{\tau\sqrt{\eta(1-\sigma)}}{\mathcal{D}_0^{3/2}}, \qquad \widehat{\mathcal{E}} := \frac{\tau\mathcal{E}}{2 + 3\mathcal{E}\sqrt{f(x_0) - f(\overline{x})}}, \tag{17}$$

where we have assumed that x_0 is not a solution of (1).

Theorem 2.2 *Assume that $0 < \mathcal{D}_0 < \infty$ and that $\varepsilon_k = 0$ for all $k \geq 1$ in* Algorithm 1. *Let $\mathcal{E}$ and $\widehat{\mathcal{E}}$ be defined in (17) and let $\overline{x}$ be a solution of (1). Then, for every $k \geq 1$:*

$$f(x_k) - f(\overline{x}) \leq \frac{f(x_0) - f(\overline{x})}{\left[1 + k\,\widehat{\mathcal{E}}\sqrt{f(x_0) - f(\overline{x})}\right]^2} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{18}$$

Moreover, for each $k \geq 2$ even, there exists $j \in \{k/2 + 1, \ldots, k\}$ such that

$$\|v_j\| \leq \frac{4}{\sqrt[3]{\eta(1-\sigma)}}\left(\frac{f(x_0) - f(\overline{x})}{\tau k\left[2 + k\,\widehat{\mathcal{E}}\sqrt{f(x_0) - f(\overline{x})}\right]^2}\right)^{2/3} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{19}$$

3 A proximal-Newton method for unconstrained convex optimization

In this section we consider the minimization problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \mathcal{H}, \end{aligned} \tag{20}$$

where $f : \mathcal{H} \to \mathbb{R}$ is assumed to have Lipschitz continuous second derivatives, i.e., it is twice differentiable and there exists $L > 0$ such that

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \qquad \forall\, x, y \in \mathcal{H}. \tag{21}$$

As in the previous section, we also assume that the solution set of (20) is nonempty. Clearly, problem (20) is a special case of (1) and, hence, Algorithm 1 can be applied to find approximate solutions of (20). As long as condition (21) holds, it is a matter of fact to prove that

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\| \leq \frac{L}{2}\|y - x\|^2 \quad \forall\, x, y \in \mathcal{H}. \tag{22}$$

For a given pair $(x, \lambda) \in \mathcal{H} \times \mathbb{R}_{++}$, an (exact) proximal step from x with parameter λ consists in finding the unique solution z_+ of

$$0 = \lambda \nabla f(z) + z - x, \tag{23}$$

because the latter equation with $z = z_+$ is clearly equivalent to $z_+ = (\lambda \nabla f + I)^{-1} x$. Since in this section we are interested in studding proximal-Newton methods for solving (20), in what follows we will show how to perform Newton steps to find approximate solutions of (23). To this end, first consider the following "system of neighborhoods", which we will show is "good" to perform Newton steps.

Given $0 < \theta < 1$, define, for each $x \in \mathcal{H}$ and $\lambda > 0$,

$$\mathcal{N}_\theta(x, \lambda) := \left\{ y \in \mathcal{H} \,:\, \frac{\lambda L}{2}\|\lambda \nabla f(y) + y - x\| \leq \theta \right\}. \tag{24}$$

Lemma 3.1 *For $x, y \in \mathcal{H}$ and $\lambda > 0$ define*

$$s := -\left(\lambda \nabla^2 f(y) + I\right)^{-1} (\lambda \nabla f(y) + y - x), \qquad y_+ := y + s.$$

Then, the following hold:

(a) $\|s\| \leq \|\lambda \nabla f(y) + y - x\|$;

(b) $\dfrac{\lambda L}{2}\|\lambda \nabla f(y_+) + y_+ - x\| \leq \left(\dfrac{\lambda L}{2}\|\lambda \nabla f(y) + y - x\|\right)^2$;

(c) *if $y \in \mathcal{N}_\theta(x, \lambda)$, then $y_+ \in \mathcal{N}_{\theta^2}(x, \lambda)$.*

Proof (a) Note first that since f is convex, then $\nabla^2 f(y) \geq 0$. Using the latter inequality and the definition of s we obtain

$$\|s\|^2 \leq \langle \left(\lambda \nabla^2 f(y) + I\right) s, s \rangle = -\langle (\lambda \nabla f(y) + y - x), s \rangle \leq \|\lambda \nabla f(y) + y - x\| \|s\|, \tag{25}$$

which gives the desired result.

(b) Using the definitions of s and y_+, and (22) we find that

$$\begin{aligned} \|\lambda \nabla f(y_+) + y_+ - x\| &= \|(\lambda \nabla f(y_+) + y_+ - x) - \left[(\lambda \nabla f(y) + y - x) + (\lambda \nabla^2 f(y) + I)s\right]\| \\ &= \lambda \|\nabla f(y_+) - \nabla f(y) - \nabla^2 f(y)(y_+ - y)\| \\ &\leq \frac{\lambda L}{2}\|y_+ - y\|^2 = \frac{\lambda L}{2}\|s\|^2. \end{aligned}$$

The desired result follows by multiplying both sides of the last displayed inequality by the term $\lambda L/2$ and using Item (a).

(c) This result is a direct consequence of Item (b) and (24). □

Lemma 3.2 *For $\lambda, \eta > 0$ and $0 < \theta < 1$, define*

$$\tau := \frac{2(1-\theta)}{2 + \frac{L\eta}{2\theta} + \sqrt{\left(2 + \frac{L\eta}{2\theta}\right)^2 - 4(1-\theta)}}, \quad \lambda_+ := (1-\tau)^{-1}\lambda. \tag{26}$$

Then,

$$\frac{1-\theta}{2 + L\eta/(2\theta)} < \tau < 1. \tag{27}$$

Moreover, if $z \in \mathcal{N}_{\theta^2}(x, \lambda)$ and $\lambda\|z - x\| \leq \eta$, then

$$z \in \mathcal{N}_{\theta}(x, \lambda_+). \tag{28}$$

Proof Define

$$q : [0,1] \to \mathbb{R}, \quad q(t) := (1-t)^2\theta - \left(\theta^2 + \frac{tL\eta}{2}\right) \tag{29}$$

and note that $q(0) = \theta(1-\theta) > 0$ and $q(1) = -(\theta^2 + L\eta/2) < 0$. Hence, using the definition of τ in (26), and (29) we conclude that τ is the smallest root of $q(t)$ and that $\tau \in]0,1[$, i.e.,

$$q(\tau) = 0, \quad 0 < \tau < 1. \tag{30}$$

Moreover,

$$\frac{1-\theta}{2 + L\eta/(2\theta)} = -\frac{q(0)}{q'(0)} < \tau < 1, \tag{31}$$

which gives (27).

Assume now that $z \in \mathcal{N}_{\theta^2}(x, \lambda)$ and $\lambda\|z - x\| \leq \eta$. Since $\lambda_+ > \lambda$, it follows from the triangle inequality that

$$\begin{aligned}\|\lambda_+\nabla f(z) + z - x\| &= \|(\lambda_+/\lambda)\,(\lambda\nabla f(z) + z - x) + (1 - (\lambda_+/\lambda))\,(z - x)\| \\ &\leq (\lambda_+/\lambda)\|\lambda\nabla f(z) + z - x\| + ((\lambda_+/\lambda) - 1)\,\|z - x\|.\end{aligned}$$

Multiplication of both sides of the latter inequality by $\lambda_+ L/2$, the assumptions on z, the fact that $\lambda_+/\lambda = (1-\tau)^{-1}$ and the identity in (30) yield

$$\begin{aligned}\frac{\lambda_+ L}{2}\|\lambda_+\nabla f(z) + z - x\| &\leq \left(\frac{\lambda_+}{\lambda}\right)^2 \frac{\lambda L}{2}\|\lambda\nabla f(z) + z - x\| + \frac{\lambda_+}{\lambda}\left(\frac{\lambda_+}{\lambda} - 1\right)\frac{L}{2}\,(\lambda\|z - x\|) \\ &\leq \left(\frac{\lambda_+}{\lambda}\right)^2 \theta^2 + \frac{\lambda_+}{\lambda}\left(\frac{\lambda_+}{\lambda} - 1\right)\frac{L\eta}{2} \\ &= \frac{\theta^2}{(1-\tau)^2} + \frac{\tau}{(1-\tau)^2}\frac{L\eta}{2} = \theta,\end{aligned}$$

which, combined with (24), gives (28). □

Corollary 3.1 *Let $y_+ = y + s$ be defined in Lemma 3.1 and λ_+ in (26). If $y \in \mathcal{N}_{\theta}(x, \lambda)$ and $\lambda\|y_+ - x\| \leq \eta$, then*

$$y_+ \in \mathcal{N}_{\theta}(x, \lambda_+). \tag{32}$$

Proof The result follows from Lemma 3.1(c) and the second statement in Lemma 3.2. □

Lemma 3.3 *Let $z \in \mathcal{N}_{\theta^2}(x,\lambda)$, τ as in (26), $0<\sigma<1$ and define*

$$w := (1-\tau)x + \tau z, \quad \lambda_+ := (1-\tau)\lambda. \tag{33}$$

If $\lambda\|z-x\| \geq \eta := \dfrac{2\theta^2}{\sigma L}$, then

(a) $\|\lambda\nabla f(z) + z - x\| \leq \sigma\|z-x\|$;
(b) $z \in \mathcal{N}_\theta(w,\lambda_+)$.

Proof (a) Using the assumptions that $z \in \mathcal{N}_{\theta^2}(x,\lambda)$, $\lambda\|z-x\| \geq \eta$, the definition of η and (24) we obtain

$$\|\lambda\nabla f(z) + z - x\| \leq \frac{2\theta^2}{\lambda L} = \frac{\sigma\eta}{\lambda} \leq \sigma\|z-x\|.$$

(b) Note that

$$\|\lambda_+\nabla f(z) + z - w\| = (1-\tau)\|\lambda\nabla f(z) + z - x\| \leq (1-\tau)\frac{2\theta^2}{\lambda L} = (1-\tau)^2\frac{2\theta^2}{\lambda_+ L} \leq \frac{2\theta^2}{\lambda_+ L}$$

and so, using (24), we conclude that $z \in \mathcal{N}_\theta(w,\lambda_+)$. □

Corollary 3.2 *Let $y_+ = y + s$ be defined in Lemma 3.1, τ and λ_+ in (26) and (33), respectively, and*

$$x_+ = (1-\tau)x + \tau y_+.$$

If $y \in \mathcal{N}_\theta(x,\lambda)$ and $\lambda\|y_+ - x\| \geq \eta := \dfrac{2\theta^2}{\sigma L}$, then

(a) $\|\lambda\nabla f(y_+) + y_+ - x\| \leq \sigma\|y_+ - x\|$;
(b) $y_+ \in \mathcal{N}_\theta(x_+,\lambda_+)$.

Proof The result is a direct consequence of Lemma 3.1(c) and Lemma 3.3. □

Next is the main algorithm of this paper.

Algorithm 2 A proximal-Newton method for convex optimization

0) Let $y_0 = x_0 \in \mathcal{H}$ such that $\nabla f(y_0) \neq 0$, let $0 < \sigma, \theta < 1$ and set

$$\eta := \frac{2\theta^2}{\sigma L}, \qquad \tau := \frac{2(1-\theta)}{(2+\theta/\sigma) + \sqrt{(2+\theta/\sigma)^2 - 4(1-\theta)}}, \quad \lambda_1 := \sqrt{\frac{2\theta}{L\|\nabla f(y_0)\|}}. \tag{34}$$

Set $i = 1$;
1) compute $s = s_i$ solving

$$\left(\lambda_i\nabla^2 f(y_{i-1}) + I\right) s = -\left(\lambda_i\nabla f(y_{i-1}) + y_{i-1} - x_{i-1}\right) \tag{35}$$

and set $y_i = y_{i-1} + s_i$;
2) if $\nabla f(y_i) = 0$, then **stop** and output y_i;
3) otherwise, **(3.a)** if

$$\lambda_i\|y_i - x_{i-1}\| \geq \eta \tag{36}$$

then set

$$x_i := (1-\tau)x_{i-1} + \tau y_i, \qquad \lambda_{i+1} := (1-\tau)\lambda_i; \tag{37}$$

(3.b) else

$$x_i = x_{i-1}, \quad \lambda_{i+1} := (1-\tau)^{-1}\lambda_i; \tag{38}$$

4) let $i \leftarrow i+1$ and go to step 1.

Remarks. 1) Using the definition of η and τ in (34) we can easily check that the latter parameter coincides with the one defined in (26) and so belongs to $]0,1[$; 2) In Section 7 of [1], for a given tolerance $\rho > 0$, a proximal-Newton method for solving (20) with iteration-complexity $\mathcal{O}(1/\sqrt{\rho})$ was proposed and analyzed. In each step $i \geq 1$, the latter algorithm computes $\lambda_i > 0$ and $s_i := (\lambda_i \nabla^2 f(x_{i-1}) + I)^{-1} \lambda_i \nabla f(x_{i-1})$ such that

$$\frac{2\sigma_\ell}{L} \leq \lambda_i \|s_i\| \leq \frac{2\sigma_u}{L},$$

where $0 < \sigma_\ell < \sigma_u < 1$. As was mentioned in [1], the procedure to find a pair (λ_i, s_i) as above depends on a binary search proposed in [7], and an improvement of this procedure would be a topic of future research. In this sense, note that Algorithm 2 does not depend on any procedure similar to the one discussed above, instead of that, in each iteration, it performs a fine tuning of the step-size $\lambda_i > 0$ (see (37) and (38) and Proposition 4.1).

4 Complexity analysis

In this section we will study the iteration-complexity of Algorithm 2 to find approximate solutions of (20). More precisely, for a given tolerance $\rho > 0$, we will estimate the number of iterations to find $x \in \mathcal{H}$ satisfying

$$f(x) - f(\bar{x}) \leq \rho \quad \text{or} \quad \|\nabla f(x)\| \leq \rho, \tag{39}$$

where $\bar{x}$ is a solution of (20). The main idea is to show that (for a suitable selection of indexes) Algorithm 2 is a special instance of Algorithm 1 and so the iteration-complexity will follow from Theorem 2.2. This is done in Proposition 4.2 and Theorems 4.2 and 4.3, where we show that the number of iterations to find $x \in \mathcal{H}$ satisfying (39) is bounded by $\mathcal{O}\left(1/\sqrt{\rho} + \log(1/\rho)\right)$.

From now one, w.l.o.g., we will assume that Algorithm 2 generates infinite sequences.

Proposition 4.1 *Let $\{x_i\}$, $\{y_i\}$ and $\{\lambda_i\}$ be generated by* Algorithm 2. *The following hold for every $i \geq 1$:*

(a) $y_{i-1} \in \mathcal{N}_\theta(x_{i-1}, \lambda_i)$;
(b) $y_i \in \mathcal{N}_{\theta^2}(x_{i-1}, \lambda_i)$.

Proof Let us proceed by induction on $i \geq 1$. (a) Using the definition of λ_1 in (34), (24) and the fact that $y_0 = x_0$ we conclude that (a) holds for $i = 1$. Assume now that it is true for some $i \geq 1$, i.e., $y_{i-1} \in \mathcal{N}_\theta(x_{i-1}, \lambda_i)$, for some $i \geq 1$. If (36) holds true, then we can use Algorithm 2's definition and Corollary 3.2 to conclude that $y_i \in \mathcal{N}_\theta(x_i, \lambda_{i+1})$. If this is not the case, then, in particular, it follows from (38) that $x_i = x_{i-1}$ and so by Algorithm 2's definition and Corollary 3.1 we have that $y_i \in \mathcal{N}_\theta(x_{i-1}, \lambda_{i+1}) = \mathcal{N}_\theta(x_i, \lambda_{i+1})$, which completes the induction argument.

(b) This result is direct consequence of item (a), Algorithm 2's definition and Lemma 3.1 (c). □

For every $j \geq 1$, define

$$\begin{aligned} A_j &:= \{1 \leq i \leq j \ : \ \text{step (3.a) is executed at iteration } i\}, \quad a_j := \#A_j\,, \\ B_j &:= \{1 \leq i \leq j \ : \ \text{step (3.b) is executed at iteration } i\}, \quad b_j := \#B_j\,, \end{aligned} \tag{40}$$

where $\#C$ stands for the number of elements of a set C. Moreover, define

$$A := \bigcup_{j=1}^{\infty} A_j, \qquad B := \bigcup_{j=1}^{\infty} B_j. \tag{41}$$

To further simplify the converge analysis, define

$$K = \{k \geq 1 \ : \ k \leq \#A\}, \quad i_0 = 0, \quad i_k = k\text{-th element of } A, \tag{42}$$

and note that

$$i_0 < i_1 < i_2 \cdots, \qquad A = \{i_k \ : \ k \in K\}. \tag{43}$$

Before proceeding to the iteration-complexity analysis of Algorithm 2, we observe that if $\{x_i\}$ is generated by Algorithm 2 then

$$x_{i_k - 1} = x_{i_{k-1}} \qquad \forall k \in K. \tag{44}$$

Indeed, for any $k \in K$, by (41) and (43) we have $\{i \geq 1 \ : \ i_{k-1} < i < i_k\} \subset B$. Consequently, by the definition of B in (41), and (38) we conclude that $x_i = x_{i_{k-1}}$ whenever $i_{k-1} \leq i < i_k$. As a consequence, we obtain that (44) follows from the fact that $i_{k-1} \leq i_k - 1 < i_k$.

Proposition 4.2 *Let $\{x_{i_k}\}_{k\in K}$, $\{y_{i_k}\}_{k\in K}$ and $\{\lambda_{i_k}\}_{k\in K}$ be generated by* Algorithm 2 *where, for every $k \in K$, i_k is defined in* (42)*. For every $k \in K$, define*

$$v_{i_k} := \nabla f(y_{i_k}), \quad \varepsilon_{i_k} := 0, \qquad \tau_{i_k} := \tau. \tag{45}$$

Then, for all $k \in K$:

$$\begin{aligned} &v_{i_k} \in \partial_{\varepsilon_{i_k}} f(y_{i_k}), \\ &\|\lambda_{i_k} v_{i_k} + y_{i_k} - x_{i_{k-1}}\| \leq \sigma \|y_{i_k} - x_{i_{k-1}}\|, \\ &\lambda_{i_k} \|y_{i_k} - x_{i_{k-1}}\| \geq \eta, \\ &x_{i_k} = (1 - \tau_{i_k}) x_{i_{k-1}} + \tau_{i_k} y_{i_k}. \end{aligned} \tag{46}$$

As a consequence, the sequences $\{x_{i_k}\}_{k\in K}$, $\{y_{i_k}\}_{k\in K}$, $\{v_{i_k}\}_{k\in K}$, $\{\varepsilon_{i_k}\}_{k\in K}$, $\{\tau_{i_k}\}_{k\in K}$ and $\{\lambda_{i_k}\}_{k\in K}$ are generated by Algorithm 1*, i.e.,* Algorithm 2 *with steps* **0)**–**2)**, **3.a)** *and* **4)** *is a special instance of* Algorithm 1.

Proof Note that the inclusion in (46) is a direct consequence of the convexity of f and the first definition in (45). Moreover, the remaining statements in (46) follow from Algorithm 2's definition, (42), (44), Proposition 4.1(a) and Corollary 3.2(a). The last statement of the proposition follows from (46) and the definitions of Algorithms 1 and 2. □

Analogously to the previous section, let $\mathcal{D}_0$ denote the diameter of the level set $[f \leq f(x_0)]$, that is,

$$\mathcal{D}_0 = \sup\{\|x - y\| \ : \ \max\{f(x), f(y)\} \leq f(x_0)\}. \tag{47}$$

and denote:

$$\mathcal{E} := \frac{\tau\sqrt{\eta(1-\sigma)}}{\mathcal{D}_0^{3/2}}, \qquad \widehat{\mathcal{E}} := \frac{\tau\mathcal{E}}{2 + 3\mathcal{E}\sqrt{f(x_0) - f(\overline{x})}}. \tag{48}$$

Theorem 4.1 *Assume that $0 < \mathcal{D}_0 < \infty$, let $\mathcal{E}$ and $\widehat{\mathcal{E}}$ be defined in* (48) *and let $\overline{x}$ be a solution of* (20)*. Let also $\{x_{i_k}\}_{k\in K}$ and $\{y_{i_k}\}_{k\in K}$ be generated by* Algorithm 2*. Then, for every $k \in K$:*

$$f(x_{i_k}) - f(\overline{x}) \leq \frac{f(x_0) - f(\overline{x})}{\left[1 + k\,\widehat{\mathcal{E}}\sqrt{f(x_0) - f(\overline{x})}\right]^2} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{49}$$

Moreover, for each $k \in K$ $(k \geq 2)$ *even, there exists* $j \in \{k/2+1, \ldots, k\}$ *such that*

$$\|\nabla f(y_{i_j})\| \leq \frac{4}{\sqrt[3]{\eta(1-\sigma)}} \left(\frac{f(x_0) - f(\overline{x})}{\tau k \left[2 + k\,\widehat{\mathcal{E}}\sqrt{f(x_0) - f(\overline{x})}\right]^2} \right)^{2/3} \tag{50}$$

$$\leq \frac{4}{\sqrt[3]{\eta(1-\sigma)\widehat{\mathcal{E}}^4\tau^2\,k^2}} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{51}$$

Proof The proof follows from the last statement in Proposition 4.2 and Theorem 2.2. □

Next we analyze the sequence generated by Algorithm 2 for the indexes $i \in B$. The following lemma is a direct consequence of Algorithm 2's definition.

Lemma 4.1 *For all* $i \geq 1$*:*

$$\lambda_{i+1} = (1-\tau)^{a_i - b_i}\lambda_1.$$

Proposition 4.3 *Let* $\{\lambda_i\}$ *and* $\{y_i\}$ *be generated by* Algorithm 2 *and let* $\bar{x}$ *be a solution of* (20)*. For any* $i \in B$*:*

(a) $\|\nabla f(y_i)\| \leq \dfrac{2\theta^2(1+\sigma)}{\sigma L\,\lambda_i^2}$;

(b) $f(y_i) - f(\bar{x}) \leq \dfrac{2\mathcal{D}_0\theta^2(1+\sigma)}{\sigma L\,\lambda_i^2}$.

Proof (a) Using Proposition 4.1(b) and (24) we have

$$\frac{\lambda_i L}{2}\|\lambda_i\nabla f(y_i) + y_i - x_{i-1}\| \leq \theta^2.$$

If $i \in B$, then $\lambda_i\|y_i - x_{i-1}\| < \eta$ and so, using the latter displayed equation and the triangle inequality, we obtain

$$\frac{\lambda_i L}{2}\|\lambda_i\nabla f(y_i)\| < \theta^2 + \frac{L\eta}{2},$$

which, in turn, combined with the definition of η in (34) gives Item (a).

(b) Since f is convex we have $f(\bar{x}) \geq f(y_i) + \langle\nabla f(y_i), \bar{x} - y_i\rangle$. If follows from the last statement in Proposition 4.2 and the first inequality in (8) that $f(y_i) \leq f(x_0)$. As a consequence, using (47) and the fact that $\bar{x}$ is a solution, we find $f(y_i) - f(\bar{x}) \leq \|\nabla f(y_i)\|\|y_i - \bar{x}\| \leq \|\nabla f(y_i)\|\mathcal{D}_0$. Now note that the desired result follows from the latter inequality and Item (a). □

The following corollary is a direct consequence of Lemma 4.1, Proposition 4.3 and the definition of λ_1 in (34).

Corollary 4.1 *If* $i \in B$*, then:*

(a) $\|\nabla f(y_{i+1})\| \leq \dfrac{\theta(1+\sigma)\|\nabla f(x_0)\|}{\sigma}(1-\tau)^{2(b_i - a_i)}$,

(b) $f(y_{i+1}) - f(\bar{x}) \leq \dfrac{\theta(1+\sigma)\mathcal{D}_0\|\nabla f(x_0)\|}{\sigma}(1-\tau)^{2(b_i - a_i)}$.

In what follows we present our main results, namely the iteration-complexity of Algorithm 2 to find approximate solutions of (20) in terms of function/gradient values.

Recall that $\mathcal{D}_0$ denotes the diameter of the sublevel determined by $f(x_0)$ and we are assuming in this paper that $0 < \mathcal{D}_0 < \infty$.

Theorem 4.2 *Let $\bar{x} \in \mathcal{H}$ be a solution of* (20) *and let $\rho > 0$ be a given tolerance. Let also $\widehat{\mathcal{E}}$ be defined in* (48)*. Then,* Algorithm 2 *finds a point $x \in \mathcal{H}$ such that*

$$f(x) - f(\bar{x}) \leq \rho \tag{52}$$

in at most

$$M := 2\left\lceil 1 + \frac{1}{\widehat{\mathcal{E}}}\left(\frac{1}{\sqrt{\rho}} - \frac{1}{\sqrt{f(x_0) - f(\bar{x})}}\right)^+\right\rceil + \left\lceil 2 + \frac{1}{2\tau}\log^+\left(\frac{\theta(1+\sigma)\mathcal{D}_0\|\nabla f(x_0)\|}{\sigma\rho}\right)\right\rceil \tag{53}$$

iterations.

Proof Define

$$M_1 := \left\lceil 1 + \frac{1}{\widehat{\mathcal{E}}}\left(\frac{1}{\sqrt{\rho}} - \frac{1}{\sqrt{f(x_0) - f(\bar{x})}}\right)^+\right\rceil, \quad M_2 := M - 2M_1. \tag{54}$$

We will consider two cases: (i) $\#A \geq M_1$ and (ii) $\#A < M_1$. In the first case, it follows from (49) and the definition of M_1 in (54) that Algorithm 2 finds $x \in \mathcal{H}$ satisfying (52) in at most M_1 iterations. Since $M \geq M_1$, we have the desired result.

Assume now that (ii) holds, i.e., $\#A < M_1$. Let $j^* \geq 1$ be such that $\#A = a_{j^*} = a_j$ for all $j \geq j^*$ (see (40)). Consequently, if $b_i \geq M_1 + M_2$, for some $i \geq j^*$, we have

$$\gamma_i := b_i - a_i = b_i - \#A > b_i - M_1 \geq M_2. \tag{55}$$

Using the definition of M_2 in (54), and (53), we have that the latter inequality gives

$$\gamma_i \geq \frac{1}{2\tau}\log\left(\frac{\theta(1+\sigma)\mathcal{D}_0\|\nabla f(x_0)\|}{\sigma\rho}\right),$$

which, in turn, combined with Corollary 4.1(b) and the definition of γ_i in (55) yields

$$\rho \geq f(y_{i+1}) - f(\bar{x}),$$

i.e., the point $x = y_{i+1}$ satisfies (52).

Since the index $i \geq j^*$ has been chosen to satisfy $b_i \geq M_1 + M_2$, and $a_i = \#A < M_1$, we conclude that, in this case, the total number of iterations to find a point $x \in \mathcal{H}$ satisfying (52) is at most $(M_1 + M_2) + M_2 = 2M_1 + M_2 = M$. □

Theorem 4.3 *Under the same assumptions of Theorem 4.2,* Algorithm 2 *finds a point $x \in \mathcal{H}$ such that*

$$\|\nabla f(x)\| \leq \rho \tag{56}$$

in at most

$$\widetilde{M} := 2\left\lceil 1 + \frac{2}{\sqrt{\rho}\,[\eta(1-\sigma)]^{1/6}\,\tau^{1/3}\widehat{\mathcal{E}}^{\,2/3}}\right\rceil + \left\lceil 2 + \frac{1}{2\tau}\log^+\left(\frac{\theta(1+\sigma)\|\nabla f(x_0)\|}{\sigma\rho}\right)\right\rceil \tag{57}$$

iterations.

Proof The proof follows the same outline of Theorem 4.2's proof, just using (51) and Corollary 4.1(a) instead of (49) and Corollary 4.1(b) , respectively. □

A Appendix

A.1 A rate of convergence result

Lemma A.1 *Let $\{\alpha_k\}_{k\geq 0}$ be a sequence of nonnegative real numbers such that*

$$\alpha_{k-1} + (1-\tau)\mathcal{D}\alpha_{k-1}^{3/2} \geq \alpha_k + \mathcal{D}\alpha_k^{3/2} \qquad \forall k \geq 1, \tag{58}$$

where $\mathcal{D} > 0$ and $0 < \tau \leq 1$. Then,

$$\alpha_k \leq \frac{\alpha_0}{\left[1 + \dfrac{\tau\mathcal{D}\sqrt{\alpha_0}}{2+3\mathcal{D}\sqrt{\alpha_0}}\, k\right]^2} \quad \forall k \geq 0. \tag{59}$$

Proof It follows from (58) that $\{\alpha_k\}_{k\geq 0}$ is nonincreasing. Hence, if $\alpha_j = 0$, then the same holds for for all $k \geq j$ and so that (59) is trivially true (whenever $k \geq j$). Assume now that $\alpha_j > 0$ for some $j \geq 1$. Define, for $1 \leq k \leq j$,

$$\beta_k := \sqrt{\alpha_k}, \qquad \varphi_k(t) := t^2 + \mathcal{D}t^3 - \left(\beta_{k-1}^2 + (1-\tau)\mathcal{D}\beta_{k-1}^3\right), \quad t \geq 0.$$

Since φ_k is convex and increasing, and $\varphi_k(\beta_{k-1}) = \tau\mathcal{D}\beta_{k-1}^3 > 0$ and, because of (58), $\varphi_k(\beta_k) \leq 0$ we obtain

$$\beta_k \leq \beta_{k-1} - \frac{\varphi_k(\beta_{k-1})}{\varphi_k'(\beta_{k-1})} = \beta_{k-1}\left(1 - \frac{\tau\mathcal{D}\beta_{k-1}}{2+3\mathcal{D}\beta_{k-1}}\right) \quad \forall k \leq j.$$

Thus, for $1 \leq k \leq j$,

$$\begin{aligned}
\frac{1}{\beta_k} \geq \frac{1}{\beta_{k-1}} \frac{1}{\left(1 - \dfrac{\tau\mathcal{D}\beta_{k-1}}{2+3\mathcal{D}\beta_{k-1}}\right)} &\geq \frac{1}{\beta_{k-1}}\left(1 + \frac{\tau\mathcal{D}\beta_{k-1}}{2+3\mathcal{D}\beta_{k-1}}\right) \\
&= \frac{1}{\beta_{k-1}} + \frac{\tau\mathcal{D}}{2+3\mathcal{D}\beta_{k-1}} \\
&\geq \frac{1}{\beta_{k-1}} + \frac{\tau\mathcal{D}}{2+3\mathcal{D}\beta_0} \\
&\geq \frac{1}{\beta_0} + \frac{\tau\mathcal{D}}{2+3\mathcal{D}\beta_0} k,
\end{aligned}$$

which after some trivial algebraic manipulations gives (59). □

A.2 Proof of results from Section 2

Proof of Proposition 2.1:

Proof The inclusion and the first inequality in (6) are, respectively, equivalent to

$$f(x) \geq f(y_k) + \langle v_k, x - y_k\rangle - \varepsilon_k \qquad \forall x \in \mathcal{H}, \tag{60}$$

$$\langle v_k, x_{k-1} - y_k\rangle - \varepsilon_k \geq \frac{\lambda_k}{2}\|v_k\|^2 + \frac{(1-\sigma^2)}{2\lambda_k}\|y_k - x_{k-1}\|^2. \tag{61}$$

Using (60) with $x = x_{k-1}$, (61) and the fact that $t + 1/t \geq 2$ for all $t > 0$ we find

$$\begin{aligned}
f(x_{k-1}) - f(y_k) &\geq \frac{\lambda_k}{2}\|v_k\|^2 + \frac{(1-\sigma^2)}{2\lambda_k}\|y_k - x_{k-1}\|^2 \\
&= \frac{\lambda_k}{2}\|v_k\|^2\left(1 + \frac{1-\sigma^2}{\mu^2}\right) \qquad \left[\mu := \frac{\|\lambda_k v_k\|}{\|y_k - x_{k-1}\|}\right] \\
&= \frac{\lambda_k}{2}\|v_k\|^2 \frac{\sqrt{1-\sigma^2}}{\mu}\left(\frac{\mu}{\sqrt{1-\sigma^2}} + \frac{\sqrt{1-\sigma^2}}{\mu}\right) \\
&\geq \lambda_k\|v_k\|^2 \frac{\sqrt{1-\sigma^2}}{\mu}.
\end{aligned} \tag{62}$$

Direct use of the definition of μ and the large-step condition, i.e., the second inequality in (6), yield $\sqrt{\|v_k\|} \geq \sqrt{\mu\eta}/\lambda_k$. Moreover, the definition of μ and Lemma 2.1 imply $1+\sigma \geq \mu$. The two last inequalities and (62) yield

$$\begin{aligned} f(x_{k-1}) - f(y_k) &\geq \lambda_k \|v_k\|^{3/2} \frac{\sqrt{\|v_k\|}\sqrt{1-\sigma^2}}{\mu} \\ &\geq \|v_k\|^{3/2} \sqrt{\frac{\eta(1-\sigma^2)}{\mu}} \geq \sqrt{\eta(1-\sigma)}\|v_k\|^{3/2}. \end{aligned} \tag{63}$$

Analogously, we have

$$f(x_{k-1}) - f(y_k) \geq \frac{\lambda_k}{2}\|v_k\|^2 + \frac{(1-\sigma^2)}{2\lambda_k}\|y_k - x_{k-1}\|^2 \geq \frac{(1-\sigma^2)}{2\lambda_k}\|y_k - x_{k-1}\|^2,$$

which, in turn, combined with (63) yields,

$$f(x_{k-1}) - f(y_k) \geq \max\left\{\sqrt{\eta(1-\sigma)}\|v_k\|^{3/2}, \frac{(1-\sigma^2)}{2\lambda_k}\|y_k - x_{k-1}\|^2\right\} =: M_k. \tag{64}$$

Note now that from (7) and the convexity of f we have $\tau_k f(y_k) + (1-\tau_k) f(x_{k-1}) \geq f(x_k)$. Multiplying both sides of (64) by τ_k, using the fifth remark after Algorithm 1 and summing the resulting inequality to the latter inequality we obtain

$$f(x_{k-1}) - f(x_k) \geq \tau_k M_k, \tag{65}$$

which together with (64) and the fact that $\tau \leq \tau_k \leq 1$ gives the first inequality in (8). To finish the proof note that the second inequality in (8) is a direct consequence of first inequality in (6). □

Proof of Lemma 2.2:

Proof From the inclusion in (6) and the definition of ε-subdifferential we have

$$f(y_k) - f(\overline{x}) \leq \|v_k\|\|y_k - \overline{x}\| + \varepsilon_k.$$

On the other hand, from the first inequality in (6) and Lemma 2.1 we obtain

$$\varepsilon_k \leq \frac{\sigma^2}{2(1-\sigma)}\|v_k\|\|y_k - x_{k-1}\|.$$

Using Proposition 2.1 and the fact that $\overline{x}$ is a solution of (1), we also obtain $\max\{f(y_k), f(\overline{x}), f(x_{k-1})\} \leq f(x_0)$. Using the latter inequality, (10) and the two above displayed equations we find

$$f(y_k) - f(\overline{x}) \leq \mathcal{D}_0\left(1 + \frac{\sigma^2}{2(1-\sigma)}\right)\|v_k\|.$$

The latter inequality, Proposition 2.1 (the first inequality in (8)) and (11) yields

$$f(x_{k-1}) - f(x_k) \geq \mathcal{D}\left(f(y_k) - f(\overline{x})\right)^{3/2}.$$

Using the above inequality together with (7), the convexity of f and of the scalar function $0 \leq t \mapsto t^{3/2}$, and the fact that $\tau \leq \tau_k \leq 1$ we obtain

$$\begin{aligned} [f(x_k) - f(\overline{x})]^{3/2} - (1-\tau)\left[f(x_{k-1}) - f(\overline{x})\right]^{3/2} &\leq [f(y_k) - f(\overline{x})]^{3/2} \\ &\leq \frac{[f(x_{k-1}) - f(x_k)]}{\mathcal{D}}, \end{aligned}$$

which is clearly equivalent to (12). □

Proof of Theorem 2.1:

Proof First note that (14) is a direct consequence of Lemma2.2, Lemma A.1 with $\alpha_k := f(x_k) - f(\overline{x})$ (for all $k \geq 1$) and (13).

Assume now that $k \geq 2$ is even. It follows from the latter assumption and Proposition 2.1 that there exists $j \in \{k/2+1, \ldots, k\}$ such that

$$\begin{aligned} f(x_{k/2}) - f(x_k) &= \sum_{i=k/2+1}^{k} [f(x_{i-1}) - f(x_i)] \\ &\geq \frac{k}{2}\tau \max\left\{\sqrt{\eta(1-\sigma)}\|v_j\|^{3/2}, \frac{(1-\sigma^2)}{\sigma^2}\varepsilon_j\right\}. \end{aligned}$$

After some trivial algebraic manipulations, the above inequality together with (13) and (14) gives (15) and (16). □

Proof of Theorem 2.2:

Proof First note that the assumption $\varepsilon_k = 0$ for all $k \geq 1$ (together with the definition of $\mathcal{E}$) implies that (12) holds with $\mathcal{E}$ replacing $\mathcal{D}$. The rest of the proof follows the same outline of Theorem 2.1's proof. □

References

1. Attouch, H., Marques Alves, M., Svaiter, B.F.: A dynamic approach to a proximal-Newton method for monotone inclusions in Hilbert spaces, with complexity $O(1/n^2)$. J. Convex Anal. 23, 139–180 (2016).
2. Dennis, Jr, J. E., Schnabel, R. B.: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1996). Corrected reprint of the 1983 original.
3. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle*, 4(Ser. R-3), 154–158 (1970).
4. Monteiro, R. D. C., Svaiter, B. F.: 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 J. Optim.* 21, 1688–1720 (2010).
5. Monteiro, R. D. C., Svaiter, B. F.: On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM J. Optim.* 20, 2755–2787 (2010).
6. Monteiro, R. D. C., Svaiter, B. F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM J. Optim.* 23, 1092–1125 (2013).
7. Monteiro, R. D. C., Svaiter, B. F.: Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems. *SIAM J. Optim.* 22(3), 914–935 (2012).
8. Nesterov, Yu., Polyak, B. T.: Cubic regularization of Newton method and its global performance. *Math. Program.* 108(1, Ser. A), 177–205 (2006).
9. Rockafellar, R. T.: Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization* 14(5), 877–898 (1976).
10. Rockafellar, R. T.: Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization* 14(5), 877–898 (1976).
11. Solodov, M. V., Svaiter, B. F.: A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Anal.* 7(4), 323–345 (1999).
12. Solodov, M. V., Svaiter, B. F.: A hybrid projection-proximal point algorithm. *J. Convex Anal.* 6(1), 59–70 (1999).