2018/10

DP

Geovani N. Grapiglia and Yurii Nesterov

Accelerated regularized Newton methods for minimizing composite convex functions

CORE
CENTER FOR OPERATIONS RESEARCH AND ECONOMETRICS

CORE
Voie du Roman Pays 34, L1.03.01
Tel (32 10) 47 43 04
Fax (32 10) 47 43 01
Email: immaq-library@uclouvain.be
https://uclouvain.be/en/research-institutes/immaq/core/discussion-papers.html

CORE DISCUSSION PAPER
2018/10

ACCELERATED REGULARIZED NEWTON METHODS FOR MINIMIZING COMPOSITE CONVEX FUNCTIONS

G.N. GRAPIGLIA* AND YU. NESTEROV†

Abstract. In this paper, we study accelerated Regularized Newton Methods for minimizing objectives formed as a sum of two functions: one is convex and twice differentiable with Hölder-continuous Hessian, and the other is a simple closed convex function. For the case in which the Hölder parameter $\nu \in [0,1]$ is known, we propose methods that take at most $\mathcal{O}\left(\frac{1}{\epsilon^{1/(2+\nu)}}\right)$ iterations to reduce the functional residual below a given precision $\epsilon > 0$. For the general case, in which the ν is not known, we propose a universal method that ensures the same precision in at most $\mathcal{O}\left(\frac{1}{\epsilon^{2/[3(1+\nu)]}}\right)$ iterations.

Key words. unconstrained minimization, second-order methods, Hölder condition, worst-case global complexity bounds

AMS subject classifications. 49M15, 49M37, 58C15, 90C25, 90C30

*Departamento de Matemática, Universidade Federal do Paraná, Centro Politécnico, Cx. postal 19.081, 81531-980, Curitiba, Paraná, Brazil (grapiglia@ufpr.br). This author was supported by the National Council for Scientific and Technological Development (CNPq), Brazil, under grant 401288/2014-5.

†Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium (Yurii.Nesterov@uclouvain.be) an National Research University Higher School of Economics (Moscow). This author was partially supported by the National Council for Scientific and Technological Development (CNPq), Brazil, grants 401288/2014-5, 314908/2014-5, and by RSF grant 17-11-01027.

1. Introduction.

1.1. Motivation. Following the worst-case complexity analysis presented in [11] for a cubic regularization of Newton method, several variants of this method have been considered (see, for example, [1], [2], [4], [5] [7], [8]). Recently, in [6], regularized Newton methods were proposed for unconstrained minimization of twice-differentiable function with Hölder-continuous Hessians. Some of these methods are "universal", in the sense that they do not require the prior knowledge of the Hölder parameter $\nu \in [0,1]$ for the Hessian. When the objective is convex, it was shown that these schemes take at most $\mathcal{O}\left(\frac{1}{\epsilon^{1/(1+\nu)}}\right)$ iterations to reduce the functional residual below a given precision $\epsilon > 0$. These complexity results generalize the bound of $\mathcal{O}\left(\frac{1}{\epsilon^{1/2}}\right)$ iterations proved in [11] for the cubic regularization of Newton's method, which is applicable to functions with Lipschitz continuous Hessians ($\nu = 1$). Generalizations of these methods using high-order models were proposed in [3, 9].

As a natural step, in this paper we investigate the possibility of acceleration of regularized Newton methods in the context of composite minimization [13]. That is, we suppose that the objective is formed as a sum of two functions: one is a convex twice differentiable with Hölder-continuous Hessian, and the other is a simple closed convex function. For the case with known Hölder parameter $\nu \in [0,1]$, we propose methods with worst-case complexity of $\mathcal{O}\left(\frac{1}{\epsilon^{1/(2+\nu)}}\right)$ iterations. These complexity results generalize the bound of $\mathcal{O}\left(\frac{1}{\epsilon^{1/3}}\right)$ proved by in [12] for the accelerated cubic regularization of Newton's method with $\nu = 1$. For the general case, in which the ν is not known, we propose a universal method that ensures the same precision in at most $\mathcal{O}\left(\frac{1}{\epsilon^{2/[3(1+\nu)]}}\right)$ iterations.

1.2. Contents. The paper is organized as follows. In Section 2, we define our problem and derive the main inequalities related to the Hölder-continuity of the Hessians of the first term in the objective. In Section 3, we present complexity results for the accelerated schemes that require perfect knowledge of the Hölder parameter. Finally, in Section 4, we present an accelerated universal second-order method and establish its complexity bound for achieving small residual in the function value[1].

1.3. Notations and Generalities. In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space, and by $\mathbb{E}^*$ its *dual* space, composed by linear functions on $\mathbb{E}$. The value of function $s \in \mathbb{E}^*$ at point $x \in \mathbb{E}$ is denoted by $\langle s, x\rangle$. Important elements of the dual space are the *gradients* of a differentiable function $f : \mathbb{E} \to \mathbb{R}$:

$$\nabla f(x) \quad \in \quad \mathbb{E}^*, \quad x \in \mathbb{E}.$$

For operator $A : \mathbb{E} \to \mathbb{E}^*$, denote by A^* its *adjoint* operator defined by the identity

$$\langle Ax, y\rangle \quad = \quad \langle A^* y, x\rangle, \quad x, y \in \mathbb{E}.$$

Thus, $A^* : \mathbb{E} \to \mathbb{E}^*$. It is called self-adjoint if $A = A^*$. Important examples of such operators are *Hessians* of a twice differentiable function $f : \mathbb{E} \to \mathbb{R}$:

$$\langle \nabla^2 f(x)u, v\rangle \quad = \quad \langle \nabla^2 f(x)v, u\rangle, \quad x, u, v \in \mathbb{E}.$$

Operator $B : \mathbb{E} \to \mathbb{E}^*$ is *positive-definite* if

$$\langle Bx, x\rangle \quad > \quad 0, \quad x \in \mathbb{E} \setminus \{0\},$$

[1] Sections 3 and 4 are independent. Thus, the reader interested in the universal scheme and its implementation details can go directly to Section 4 right after reading Section 2.

(notation $B \succ 0$; we use notation $B \succeq 0$ if the above inequality is not strict). In what follows, we fix some self-adjoint positive-definite operator $B \succ 0$ for defining Euclidean norms in the primal and dual spaces:

$$\|x\| \;=\; \langle Bx, x\rangle^{1/2},\; x \in \mathbb{E}, \quad \|s\|_* \;=\; \langle s, B^{-1}s\rangle^{1/2},\; s \in \mathbb{E}^*.$$

In our analysis, we shall use some properties of uniformly convex functions.

DEFINITION 1.1. *Function $f : \mathbb{E} \to \mathbb{R}$ is called uniformly convex of degree $p \geq 2$ if for some $\sigma_p = \sigma_p(f) > 0$ and all $x, y \in \mathbb{E}$, $\theta \in [0,1]$ we have*

$$f((1-\theta)x + \theta y) \;\leq\; (1-\theta)f(x) + \theta f(y) - \tfrac{\sigma_p \theta(1-\theta)}{p}\|y - x\|^p.$$

Pair (p, σ_p) is called the pair of parameters of the uniformly convex function f. Note that, adding such a function to an arbitrary convex function gives a uniformly convex function with the same pair of parameters.

Next lemma gives a guarantee for the rate of growth of uniformly convex function.

LEMMA 1.2. *Let $\psi : \mathbb{E} \to \mathbb{R}$ be a uniformly convex function of degree $p \geq 2$. Denote $\bar{x} = \arg \min\limits_{x \in \mathrm{dom}\, \psi} \psi(x)$. Then,*

$$\psi(y) \;\geq\; \psi(\bar{x}) + \tfrac{\sigma_p}{p}\|y - \bar{x}\|^p, \quad \forall y \in \mathbb{E},$$

where (p, σ_p) is the pair of parameters of function ψ.

Proof. Given $\alpha \in (0,1]$, we have

$$\begin{aligned} \psi(\bar{x}) &\leq \psi((1-\alpha)\bar{x} + \alpha y) \\ &\leq (1-\alpha)\psi(\bar{x}) + \alpha\psi(y) - \frac{\sigma_p \alpha(1-\alpha)}{p}\|y - \bar{x}\|^p \end{aligned}$$

and so

$$\psi(y) \geq \psi(\bar{x}) + \frac{\sigma_p(1-\alpha)}{p}\|y - \bar{x}\|^p.$$

The conclusion follows by making $\alpha \to 0$. □

LEMMA 1.3. *For any $h \in \mathbb{E}$, $s \in \mathbb{E}^*$, $p \geq 2$, and $\omega > 0$, we have*

$$\langle s, h\rangle + \tfrac{\omega}{p}\|h\|^p \;\geq\; -\tfrac{(p-1)}{p}\left(\tfrac{1}{\omega}\right)^{\frac{1}{p-1}}\|s\|_*^{\frac{p}{p-1}}.$$

Proof. See Lemma 2 in [12]. □

The next lemma gives us some lower bounds for the rate of the growth of a sequence satisfying certain conditions. It will be crucial for establishing the complexity results for our accelerated schemes.

LEMMA 1.4. *Let $\alpha \in [0,1)$, and suppose that $\{B_t\}_{t \geq 0}$ is a sequence of nonnegative numbers with $B_t > 0$, $t \geq 1$, and*

$$B_{t+1} - B_t \geq B_{t+1}^{\alpha}, \quad \forall t \geq 0.$$

Then, $B_t \geq \left[(1-\alpha)\left(\frac{B_1^{1-\alpha}}{B_1^{1-\alpha}+1}\right)^{\alpha}\right]^{1/(1-\alpha)}(t-1)^{\frac{1}{1-\alpha}}$ for all $t \geq 2$.

Proof. Indeed, from the assumption on $\{B_t\}$ we have

$$B_{t+1}^{1-\alpha} \geq (B_t + B_{t+1}^{\alpha})^{1-\alpha}.$$

Then, subtracting $B_t^{1-\alpha}$ on both sides, we obtain

$$B_{t+1}^{1-\alpha} - B_t^{1-\alpha} \geq (B_t + B_{t+1}^{\alpha})^{1-\alpha} - B_t^{1-\alpha}. \tag{1.1}$$

Since $0 < 1-\alpha \leq 1$, function $g(u) = u^{1-\alpha}$ is concave on $(0, +\infty)$. Therefore,

$$u^{1-\alpha} \leq v^{1-\alpha} + (1-\alpha)v^{-\alpha}(u-v), \ \forall u, v \in (0, +\infty).$$

In particular, considering $v = B_t + B_{t+1}^{\alpha}$ and $u = B_t$, we get

$$B_t^{1-\alpha} \leq (B_t + B_{t+1}^{\alpha})^{1-\alpha} + (1-\alpha)(B_t + B_{t+1}^{\alpha})^{-\alpha}(-B_{t+1}^{\alpha}).$$

Hence,

$$(B_t + B_{t+1}^{\alpha})^{1-\alpha} - B_t^{1-\alpha} \geq (1-\alpha)(B_t + B_{t+1}^{\alpha})^{-\alpha} B_{t+1}^{\alpha}. \tag{1.2}$$

Combining (1.1) and (1.2) we obtain

$$B_{t+1}^{1-\alpha} - B_t^{1-\alpha} \geq (1-\alpha)(B_t + B_{t+1}^{\alpha})^{-\alpha} B_{t+1}^{\alpha}.$$

Thus, since sequence $\{B_t\}$ is nondecreasing, it follows that

$$\begin{aligned} (B_{t+1}^{1-\alpha} - B_t^{1-\alpha})^{\frac{1}{\alpha}} \quad &\geq \quad (1-\alpha)^{\frac{1}{\alpha}} \tfrac{B_{t+1}}{(B_t + B_{t+1}^{\alpha})} \ \geq \ (1-\alpha)^{\frac{1}{\alpha}} \tfrac{B_{t+1}}{(B_{t+1} + B_{t+1}^{\alpha})} \\ &= \quad (1-\alpha)^{\frac{1}{\alpha}} \tfrac{1}{1 + B_{t+1}^{\alpha-1}} \ \geq \ (1-\alpha)^{\frac{1}{\alpha}} \tfrac{1}{1 + B_1^{\alpha-1}}, \end{aligned} \tag{1.3}$$

where the the last inequality follows from the fact that $B_{t+1} \geq B_1 > 0$. Therefore,

$$B_{t+1}^{1-\alpha} - B_t^{1-\alpha} \quad \geq \quad (1-\alpha)\left(\tfrac{B_1^{1-\alpha}}{B_1^{1-\alpha}+1}\right)^{\alpha}, \quad \forall t \geq 1. \tag{1.4}$$

Finally, it follows from (1.4) that, for all $t \geq 2$,

$$B_t^{1-\alpha} - B_1^{1-\alpha} \quad = \quad \textstyle\sum_{i=1}^{t-1} [B_{i+1}^{1-\alpha} - B_i^{1-\alpha}] \geq (t-1)(1-\alpha)\left(\tfrac{B_1^{1-\alpha}}{B_1^{1-\alpha}+1}\right)^{\alpha},$$

and we conclude that $B_t \geq \left[(1-\alpha)\left(\frac{B_1^{1-\alpha}}{B_1^{1-\alpha}+1}\right)^{\alpha}\right]^{\frac{1}{1-\alpha}} (t-1)^{\frac{1}{1-\alpha}}$. □

Finally, we need the following lower bound on the size of subgradients of convex functions.

LEMMA 1.5. *Let $\tilde{f}$ be a closed convex function attaining its minimum at some point $x^* \in \operatorname{dom} \tilde{f}$. Given $\epsilon > 0$, let*

$$R(\epsilon) = \max_{x \in \operatorname{dom} \varphi} \left\{ \|x - x^*\| \, : \, \tilde{f}(x) \leq \tilde{f}(x^*) + \epsilon \right\}.$$

If $R(\epsilon) < +\infty$, then $\|\tilde{g}\|_* \geq \frac{\epsilon}{R(\epsilon)}$ for all $\tilde{g} \in \partial \tilde{f}(x)$ with $\tilde{f}(x) \geq \tilde{f}(x^*) + \epsilon$.

Proof. Indeed, let $\tilde{f}(x) \geq \tilde{f}(x^*) + \epsilon$. Since

$$\tilde{f}(x) \geq \tilde{f}(x) + \epsilon > \tilde{f}(x^*),$$

it follows from the Intermediate Value Theorem that there exists $\alpha \in (0,1]$ such that

$$\tilde{f}(\alpha x + (1-\alpha)x^*) = \tilde{f}(x^*) + \epsilon.$$

Then, by the convexity of $\tilde{f}$, we obtain

$$\tilde{f}(x^*) + \epsilon \leq \alpha \tilde{f}(x) + (1-\alpha)\tilde{f}(x^*),$$

which gives

$$\frac{\epsilon}{\alpha} \leq \tilde{f}(x) - \tilde{f}(x^*).$$

On the other hand,

$$R(\epsilon) \geq \|(\alpha x + (1-\alpha)x^*) - x^*\| = \alpha \|x - x^*\|,$$

and so

$$\frac{1}{\alpha} \geq \frac{\|x - x^*\|}{R(\epsilon)}.$$

Thus, if $\tilde{g} \in \partial \tilde{f}(x)$, it follows from the definition of subgradient, the Cauchy-Schwartz inequality and the above inequalities, that

$$\|\tilde{g}\|_* \|x - x^*\| \geq \tilde{f}(x) - \tilde{f}(x^*) \geq \frac{\epsilon}{R(\epsilon)} \|x - x^*\|.$$

□

2. Problem statement and auxiliary results. In this paper we consider methods for solving the following composite minimization problem:

$$\min_{x \in \mathbb{E}} \left\{ \tilde{f}(x) \equiv f(x) + \varphi(x) \right\}, \tag{2.1}$$

where $f : \mathbb{E} \to \mathbb{R}$ is a convex twice differentiable function and $\varphi : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is a simple closed convex function. Our assumption on simplicity of φ means that all subproblems appearing in our methods and involving this function are easily solvable. We assume that there exists at least one optimal solution $x^* \in \mathbb{E}$ for problem (2.1).

Let us characterize the level of smoothness of function f in problem (2.1) by the system of Hölder constants

$$H_f(\nu) \equiv \sup_{x,y \in \mathrm{dom}\,\varphi} \left\{ \tfrac{\|\nabla^2 f(x) - \nabla^2 f(y)\|}{\|x-y\|^\nu} \; : \; x \neq y \right\}, \; 0 \leq \nu \leq 1. \tag{2.2}$$

It follows from (2.2) and from an integral form of the Mean-Value Theorem that

$$\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y - x \rangle \right| \;\; \leq \;\; \tfrac{H_f(\nu)\|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)} \tag{2.3}$$

and

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\| \;\; \leq \;\; \tfrac{H_f(\nu)\|y-x\|^{1+\nu}}{1+\nu}. \tag{2.4}$$

Let $H_f(\nu) < +\infty$ for some $\nu \in [0,1]$. Consider the following model of the objective function $\tilde{f}$ around some point $x \in \mathbb{E}$:

$$\begin{aligned} Q(x;y) &= f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle, \\ M_{\nu,H}(x;y) &= Q(x;y) + \tfrac{H\|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(y),\ y \in \operatorname{dom}\varphi, \end{aligned}$$

where the parameter $H > 0$ is an estimate for the Hölder constant $H_f(\nu)$. Clearly, if $H \geq H_f(\nu)$, it follows from (2.3) that

$$\tilde{f}(y) \leq M_{\nu,H}(x;y),\ y \in \operatorname{dom}\varphi. \tag{2.5}$$

This observation suggests computation of the point

$$T_{\nu,H}(x) = \arg\min_{y \in \operatorname{dom}\varphi} M_{\nu,H}(x;y). \tag{2.6}$$

Note that, point $T = T_{\nu,H}(x)$ satisfies the following first-order optimality condition:

$$\langle \nabla f(x) + \nabla^2 f(x)(T-x) + \tfrac{H\|T-x\|^{\nu}}{1+\nu}B(T-x), y-T\rangle + \varphi(y) \geq \varphi(T) \tag{2.7}$$

for all points $y \in \operatorname{dom}\varphi$. If we denote

$$g_{\varphi}(T) = -\left(\nabla f(x) + \nabla^2 f(x)(T-x) + \tfrac{H\|T-x\|^{\nu}}{1+\nu}B(T-x)\right), \tag{2.8}$$

then by the above inequality we have

$$\langle -g_{\varphi}(T), y-T\rangle + \varphi(y) \geq \varphi(T), \quad \forall y \in \operatorname{dom}\varphi.$$

Hence, $g_{\varphi}(T) \in \partial\varphi(T)$. Moreover,

$$\nabla f(x) + \nabla^2 f(x)(T-x) + \tfrac{H\|T-x\|^{\nu}}{1+\nu}B(T-x) + g_{\varphi}(T) = 0. \tag{2.9}$$

In what follows, we use

$$\nabla\tilde{f}(T) \equiv \nabla f(T) + g_{\varphi}(T) \in \partial\tilde{f}(T),$$

with $g_{\varphi}(T)$ given by (2.8).

The following result ensures a descent condition and forms the basis for our backtracking strategies in the schemes where ν is known but $H_f(\nu)$ is unknown.

LEMMA 2.1. *Let $x_+ = T_{\nu,H}(\bar{x})$ for some $\bar{x} \in \operatorname{dom}\varphi$. If $H \geq (1+\nu)H_f(\nu)$, then*

$$\langle \nabla\tilde{f}(x_+), \bar{x} - x_+\rangle \geq \left(\tfrac{1}{2H}\right)^{\frac{1}{1+\nu}} \|\nabla\tilde{f}(x_+)\|_*^{\frac{2+\nu}{1+\nu}}.$$

Proof. Denote $r = \|x_+ - \bar{x}\|$. Then, by (2.4) we have

$$\|\nabla f(x_+) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x_+ - \bar{x})\|_*^2 \leq \tfrac{H_f(\nu)^2 r^{2(1+\nu)}}{(1+\nu)^2}. \tag{2.10}$$

On the other hand, by (2.9)

$$\nabla f(\bar{x}) + \nabla^2 f(\bar{x})(x_+ - x) + \tfrac{H}{1+\nu}r^{\nu}B(x_+ - x) + g_{\varphi}(x_+) = 0. \tag{2.11}$$

Thus, combining (2.10) and (2.11), we get

$$\begin{aligned}
\tfrac{H_f(\nu)^2 r^{2(1+\nu)}}{(1+\nu)^2} &\geq \|\nabla f(x_+) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x_+ - \bar{x})\|_*^2 \\
&= \|\nabla f(x_+) + g_\varphi(x_+) + \tfrac{1}{1+\nu} H r^\nu B(x_+ - \bar{x})\|_*^2 \\
&= \|\nabla \tilde{f}(x_+) + \tfrac{1}{1+\nu} H r^\nu B(x_+ - \bar{x})\|_*^2 \\
&= \|\nabla \tilde{f}(x_+)\|_*^2 + \tfrac{2}{(1+\nu)} H r^\nu \langle \nabla \tilde{f}(x_+), x_+ - \bar{x}\rangle + \tfrac{H^2 r^{2(1+\nu)}}{(1+\nu)^2}.
\end{aligned}$$

Hence,

$$\langle \nabla \tilde{f}(x_+), \bar{x} - x_+\rangle \geq \tfrac{(1+\nu)}{2Hr^\nu}\|\nabla \tilde{f}(x_+)\|_*^2 + \tfrac{1}{2(1+\nu)H}(H^2 - H_f(\nu)^2) r^{2+\nu}. \tag{2.12}$$

For $\nu = 0$, this inequality leads to the desired relation. Let us assume that $\nu > 0$. Denote $g = \|\nabla \tilde{f}(x_+)\|_*$ and $\Delta^2 = 1 - \left(\frac{H_f(\nu)}{H}\right)^2 \geq \frac{\nu(2+\nu)}{(1+\nu)^2}$. Consider the right-hand side of inequality (2.12) as a function of r:

$$h(r) = \tfrac{(1+\nu)}{2Hr^\nu} g^2 + \tfrac{H\Delta^2 r^{2+\nu}}{2(1+\nu)}.$$

Let us find the optimal r_* as a solution to the first-order optimality condition for function h:

$$\tfrac{\nu(1+\nu)g^2}{Hr^{1+\nu}} = \tfrac{(2+\nu)H\Delta^2 r^{1+\nu}}{1+\nu}.$$

Thus, $r_*^{1+\nu} = \frac{(1+\nu)g}{H\Delta}\sqrt{\frac{\nu}{2+\nu}}$. Consequently,

$$\begin{aligned}
h(r_*) &= \tfrac{r_*}{2H}\left[\tfrac{(1+\nu)g^2}{r_*^{1+\nu}} + \tfrac{H^2\Delta^2 r_*^{1+\nu}}{1+\nu}\right] \\
&= \tfrac{r_*}{2H}\left[(1+\nu)g^2 \tfrac{H\Delta}{(1+\nu)g}\sqrt{\tfrac{2+\nu}{\nu}} + \tfrac{H^2\Delta^2}{1+\nu}\tfrac{(1+\nu)g}{H\Delta}\sqrt{\tfrac{\nu}{2+\nu}}\right] \\
&= \tfrac{(1+\nu)g\Delta r_*}{\sqrt{\nu(2+\nu)}} = \tfrac{(1+\nu)g\Delta}{\sqrt{\nu(2+\nu)}}\left[\tfrac{(1+\nu)g}{H\Delta}\sqrt{\tfrac{\nu}{2+\nu}}\right]^{\frac{1}{1+\nu}} \\
&= \tfrac{(1+\nu)g^{\frac{2+\nu}{1+\nu}}\Delta^{\frac{\nu}{1+\nu}}}{\sqrt{\nu(2+\nu)}}\left[\tfrac{(1+\nu)}{H}\sqrt{\tfrac{\nu}{2+\nu}}\right]^{\frac{1}{1+\nu}} \\
&\geq \tfrac{(1+\nu)g^{\frac{2+\nu}{1+\nu}}}{\sqrt{\nu(2+\nu)}}\left[\tfrac{(1+\nu)}{H}\sqrt{\tfrac{\nu}{2+\nu}}\right]^{\frac{1}{1+\nu}}\left(\tfrac{\nu(2+\nu)}{(1+\nu)^2}\right)^{\frac{\nu}{2(1+\nu)}} \\
&= \left(\tfrac{1}{H}\right)^{\frac{1}{1+\nu}} g^{\frac{2+\nu}{1+\nu}} \tfrac{(1+\nu)^{\frac{2}{1+\nu}}}{(2+\nu)^{\frac{1}{1+\nu}}} \geq \left(\tfrac{1}{2H}\right)^{\frac{1}{1+\nu}} g^{\frac{2+\nu}{1+\nu}}.
\end{aligned}$$

□

The next lemma allows us to overestimate the objective function $\tilde{f}$ by a model with cubic regularization, when H and $\|\nabla \tilde{f}(x_+)\|$ are sufficiently large. This provides us with a basis for universal methods.

LEMMA 2.2. *Let $x_+ = T_{1,H}(\bar{x})$ for some $\bar{x} \in \mathbb{E}$ and $H > 0$. If for some $\delta > 0$ and $\nu \in [0,1]$ we have*

$$\|\nabla \tilde{f}(x_+)\|_* \geq \delta \quad \text{and} \quad H \geq \left[\tfrac{CH_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}} \left(\tfrac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}}, \tag{2.13}$$

with constant $C \geq 6$, then

$$\|x_+ - \bar{x}\|^{1-\nu} \geq \frac{CH_f(\nu)}{(1+\nu)(2+\nu)H} \tag{2.14}$$

and, consequently,

$$\tilde{f}(x_+) \leq M_{1,H}(\bar{x}, x_+). \tag{2.15}$$

Proof. For $\nu = 1$ the statement is trivial. Assume that $\nu \in [0,1)$. Denote $r = \|x_+ - \bar{x}\|$. Then, the first inequality in (2.13) and inequalities (2.4) and (2.9) imply that

$$\begin{aligned} \delta &\leq \|\nabla \tilde{f}(x_+)\|_* = \|\nabla f(x_+) + g_\varphi(x_+)\|_* \\ &\leq \|\nabla f(x_+) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x_+ - \bar{x})\|_* \\ &\quad + \|\nabla f(\bar{x}) + \nabla^2 f(\bar{x})(x_+ - \bar{x}) + g_\varphi(x_+)\|_* \\ &\leq \tfrac{H_f(\nu) r^{1+\nu}}{1+\nu} + \tfrac{1}{2} H r^2 = r^{1+\nu}\left[\tfrac{H_f(\nu)}{1+\nu} + \tfrac{1}{2} H r^{1-v}\right]. \end{aligned}$$

For the purpose of reaching a contradiction, assume that $Hr^{1-\nu} < \frac{CH_f(\nu)}{(1+\nu)(2+\nu)}$. Then

$$\begin{aligned} \delta &< r^{1+\nu}\left[\tfrac{H_f(\nu)}{1+\nu} + \tfrac{1}{2}\tfrac{CH_f(\nu)}{(1+\nu)(2+\nu)}\right] = \tfrac{r^{1+\nu}}{1+\nu} \cdot H_f(\nu) \cdot \left(1 + \tfrac{C}{2(2+\nu)}\right) \\ &< \tfrac{H_f(\nu)}{1+\nu}\left(1 + \tfrac{C}{2(2+\nu)}\right)\left[\tfrac{CH_f(\nu)}{(1+\nu)(2+\nu)H}\right]^{\frac{1+\nu}{1-\nu}}. \end{aligned}$$

Since $C \geq 6$, we have $1 + \frac{C}{2(2+\nu)} \leq \frac{C}{2+\nu}$. Therefore, $\delta < \left[\frac{CH_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1-\nu}} \left(\frac{1}{H}\right)^{\frac{1+\nu}{1-\nu}}$. This contradicts the second inequality in (2.13). Therefore, (2.14) holds. Note that if H satisfies the second inequality in (2.13), then $H \geq H_f(\nu)$. Thus, combining (2.5) and (2.14), we obtain (2.15):

$$\begin{aligned} \tilde{f}(x_+) &\leq Q(\bar{x}; x_+) + \frac{Hr^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(x_+) \\ &\leq Q(\bar{x}; x_+) + \frac{Hr^3}{6} + \varphi(x_+) \\ &= M_{1,H}(\bar{x}, x_+). \end{aligned}$$

□

Using Lemma 2.2, we can modify Lemma 2.1 in the following way.

LEMMA 2.3. *Let $x_+ = T_{1,H}(\bar{x})$ for some $\bar{x} \in \mathbb{E}$ and $H > 0$. If for some $\delta > 0$ and $\nu \in [0,1]$ we have*

$$\|\nabla \tilde{f}(x_+)\|_* \geq \delta \quad \text{and} \quad H \geq \left[\tfrac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}} \left(\tfrac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}},$$

then

$$\langle \nabla \tilde{f}(x_+), \bar{x} - x_+ \rangle \quad \geq \quad \sqrt{\tfrac{4}{3H}} \|\nabla \tilde{f}(x_+)\|_*^{\frac{3}{2}}. \tag{2.16}$$

Proof. Denote $r = \|x_+ - \bar{x}\|$. Then, by Lemma 2.2 (with $C = 12$),

$$\|\nabla f(x_+) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x_+ - x)\|_* \quad \leq \quad \tfrac{H_f(\nu) r^{1+\nu}}{1+\nu} \leq \tfrac{H}{4} r^2. \tag{2.17}$$

On the other hand, as $x_+ = T_{1,H}(\bar{x})$ we have

$$\nabla f(\bar{x}) + \nabla^2 f(\bar{x})(x_+ - x) + \frac{H}{2} r B(x_+ - x) + g_\varphi(x_+) = 0. \tag{2.18}$$

Thus, combining (2.17) and (2.18), we get

$$\begin{aligned}
\tfrac{H^2 r^4}{16} \quad &\geq \quad \|\nabla f(x_+) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x_+ - \bar{x})\|_*^2 \\
&= \quad \|\nabla f(x_+) + g_\varphi(x_+) + \tfrac{H}{2} r B(x_+ - \bar{x})\|_*^2 \\
&= \quad \|\nabla \tilde{f}(x_+) + \tfrac{H}{2} r B(x_+ - \bar{x})\|_*^2 \\
&= \quad \|\nabla \tilde{f}(x_+)\|_*^2 + Hr\langle \nabla \tilde{f}(x_+), x_+ - \bar{x} \rangle + \tfrac{H^2 r^4}{4}.
\end{aligned}$$

Hence, $\langle \nabla \tilde{f}(x_+), \bar{x} - x_+ \rangle \geq \frac{g^2}{Hr} + \frac{3Hr^3}{16}$, where $g = \|\nabla \tilde{f}(x_+)\|_*$. The minimum of the right-hand side in the last inequality is attained at $r_*^2 = \frac{4g}{3H}$. Thus,

$$\langle \nabla \tilde{f}(x_+), \bar{x} - x_+ \rangle \quad \geq \quad r_* \left[\tfrac{g^2}{H r_*^2} + \tfrac{3 H r_*^2}{16} \right] \; = \; r_* g \left[\tfrac{3}{4} + \tfrac{1}{4} \right] \; = \; r_* g.$$

□

3. Numerical Schemes for ν known. In this section we consider minimization schemes to solve problem (2.1) when the Hölder parameter ν is not known. We also assume that function $\varphi(\,.\,)$ is uniformly convex of degree $p = 2 + \nu$ and that its convexity parameter $\sigma_p = \sigma_p(\varphi) \geq 0$ is known[2]. In the spirit of estimating sequences [10], our accelerated schemes update recursively sequence of points $\{x_t\}_{t=0}^\infty$ and functions $\{\psi_t(\cdot)\}_{t=0}^\infty$ in such a way that they satisfy the following relation

$$A_t \tilde{f}(x_t) \leq \min_{x \in \mathbb{E}} \psi_t(x), \;\; \forall t \geq 0, \tag{3.1}$$

where $A_t = \sum_{i=0}^t a_t$ with $\{a_t\}_{t=0}^\infty$ being positive stepsize parameters, and the estimating functions being recursively updated as

$$\psi_{t+1}(x) = \psi_t(x) + a_t \left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle \right]$$

with $\psi_0(x) = \frac{1}{2+\nu}\|x - x_0\|^{2+\nu}$. Recall that from inequality (3.1) we conclude that $A_t \tilde{f}(x_t) \leq A_t \tilde{f}(x^*) + \frac{1}{2}\|x^* - x_0\|^2$. Thus, the rate of growth of coefficients $\{A_t\}_{t=0}^\infty$ defines the rate of convergence of the method.

[2]Note that $\sigma_p = 0$ implies only convexity of function φ.

Let us start with a generic framework to deal with the case in which ν is not known.

Algorithm 1. Accelerated RNM for known parameter ν

Initialization Choose $x_0 \in \mathrm{dom}\,\varphi$ and $\gamma \geq 1$. Set $v_0 = x_0$ and $A_0 = 0$.

Iteration $t \geq 0$:

a) Find $M_t \in (0, \gamma H_f(\nu))$, such that

$$\langle \nabla \tilde{f}(x_{t+1}), y_t - x_{t+1} \rangle \quad \geq \quad \left(\tfrac{1}{2M_t}\right)^{\frac{1}{1+\nu}} \|\nabla \tilde{f}(x_{t+1})\|_*^{\frac{2+\nu}{1+\nu}}, \tag{3.2}$$

where

$$\begin{aligned} x_{t+1} \quad &= \quad T_{\nu,M_t}(y_t) \;\equiv\; \arg\min_{x \in \mathrm{dom}\,\varphi} \Big\{ f(y_t) + \langle \nabla f(y_t), x - y_t \rangle \\ &\quad + \frac{1}{2}\langle \nabla^2 f(y_t)(x - y_t), x - y_t \rangle + \tfrac{M_t \|x - y_t\|^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(x) \Big\}, \end{aligned} \tag{3.3}$$

and $y_t = (1 - \alpha_t)x_t + \alpha_t v_t$, with $\alpha_t = \frac{a_t}{A_t + a_t}$ and coefficient $a_t > 0$ computed from the equation

$$a_t^{2+\nu} \quad = \quad \tfrac{(1+2^{\nu}\sigma_p A_t)}{2M_t}(A_t + a_t)^{1+\nu}. \tag{3.4}$$

b) Set

$$\begin{aligned} \psi_{t+1}(x) \quad &= \quad \psi_t(x) + a_t \left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \varphi(x) \right], \\ v_{t+1} \quad &= \quad \arg\min_{x \in \mathrm{dom}\,\varphi} \psi_{t+1}(x), \quad A_{t+1} = A_t + a_t. \end{aligned} \tag{3.5}$$

The next result establishes the relationship between the estimating functions $\psi_t(x)$ and the objective function $\tilde{f}(x)$. It will be crucial to prove global complexity rates for Algorithm 1.

LEMMA 3.1. *For all $t \geq 0$,*

$$\psi_t(x) \quad \leq \quad A_t \tilde{f}(x) + \tfrac{1}{(2+\nu)}\|x - x_0\|^{2+\nu}, \;\; \forall x \in \mathbb{E}. \tag{3.6}$$

Proof. Indeed, since $A_0 = 0$, for all $x \in \mathbb{E}$, we have

$$\psi_0(x) \quad = \quad \tfrac{1}{(2+\nu)}\|x - x_0\|^{2+\nu} = A_0 \tilde{f}(x) + \tfrac{1}{(2+\nu)}\|x - x_0\|^{2+\nu}.$$

Thus, (3.6) is true for $t = 0$. Suppose that (3.6) is true for some $t \geq 0$. Then, (3.5) and convexity of f imply that, for all $x \in \mathbb{E}$,

$$\begin{aligned}\psi_{t+1}(x) &= \psi_t(x) + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right] \\ &\leq \psi_t(x) + a_t\left[f(x) + \varphi(x)\right] \; = \psi_t(x) + a_t \tilde{f}(x) \\ &\leq A_t \tilde{f}(x) + \tfrac{\|x-x_0\|^{2+\nu}}{(2+\nu)} + a_t \tilde{f}(x) \\ &= (A_t + a_t)\tilde{f}(x) + \tfrac{\|x-x_0\|^{2+\nu}}{(2+\nu)} = A_{t+1}\tilde{f}(x) + \tfrac{\|x-x_0\|^{2+\nu}}{(2+\nu)}.\end{aligned}$$

Thus, (3.6) is also true for $t+1$. □

Now we are in position to prove that the sequences in Algorithm 1 satisfy (3.1). By combining (3.1) with (3.6) we also obtain global complexity rates for Algorithm 1.

THEOREM 3.2. *Assume that $H_f(\nu) < +\infty$ for some $\nu \in [0,1]$. If sequence $\{x_t\}_{t=0}^{\infty}$ is generated by Algorithm 1, then for all $t \geq 0$ we have*

$$A_t \tilde{f}(x_t) \leq \psi_t^* \equiv \min_{x\in\mathbb{E}} \psi_t(x). \tag{3.7}$$

Moreover,

$$A_t \geq \begin{cases} \frac{1}{2\gamma H_f(\nu)}\left[\frac{1}{(2+\nu)}\left(\frac{1}{2}\right)^{\frac{1+\nu}{2+\nu}}\right]^{2+\nu}(t-1)^{2+\nu} & \forall t \geq 2, \quad \text{if } \sigma_p = 0, \\ \left(\frac{1}{2\gamma H_f(\nu)}\right)\left[1 + \left(\frac{\sigma_p}{8\gamma H_f(\nu)}\right)^{\frac{1}{2+\nu}}\right]^{2(t-1)} & \forall t \geq 0, \quad \text{if } \sigma_p > 0. \end{cases} \tag{3.8}$$

Consequently, we have

$$\tilde{f}(x_t) - \tilde{f}^* \leq \begin{cases} \frac{(2\gamma H_f(\nu))(4+2\nu)^{1+\nu}\|x^*-x_0\|^{2+\nu}}{(t-1)^{2+\nu}} \;\; \forall t \geq 2, & \text{if } \sigma_p = 0, \\ \frac{2\gamma H_f(\nu)\|x^*-x_0\|^{2+\nu}}{(2+\nu)}\left[1 + \left(\frac{\sigma_p}{8\gamma H_f(\nu)}\right)^{\frac{1}{2+\nu}}\right]^{-2(t-1)} \;\; \forall t \geq 0, & \text{if } \sigma_p > 0, \end{cases} \tag{3.9}$$

where $\tilde{f}^* = \tilde{f}(x^*)$ and x^* is an optimal solution to the problem.

Proof. Let us prove relation (3.7) by induction over t. Since $A_0 = 0$, for $t = 0$ it is evident:

$$A_0 \tilde{f}(x_0) = 0 = \min_{x\in\mathbb{E}} \psi_0(x).$$

Assume that (3.7) is true for some $t \geq 0$. Note that, for any $x \in \mathbb{E}$,

$$\begin{aligned}\psi_t(x) &= \textstyle\sum_{i=0}^{t-1} a_i\left[f(x_{i+1}) + \langle \nabla f(x_{i+1}), x - x_{i+1}\rangle + \varphi(x)\right] + \frac{\|x-x_0\|}{2+\nu}^{2+\nu} \\ &= \textstyle\sum_{i=0}^{t-1} a_t\left[f(x_{i+1}) + \langle \nabla f(x_{i+1}), x - x_{i+1}\rangle\right] + \sum_{i=0}^{t-1} a_t \varphi(x) + \frac{\|x-x_0\|}{2+\nu}^{2+\nu} \\ &\equiv \ell_t(x) + A_t\varphi(x) + \tfrac{1}{2+\nu}\|x-x_0\|^{2+\nu}, \text{ for all } t \geq 1.\end{aligned}$$

Note that $\ell_t(x)$ is a linear function. Moreover, by Lemma 4 in [12], function $A_t\varphi(x) + \frac{1}{(2+\nu)}\|x-x_0\|^{2+\nu}$ is uniformly convex of degree $p = 2+\nu$ with parameter $2^{-\nu} + \sigma_p A_t$.

Thus, $\psi_t(x)$ is also a uniformly convex function of degree $p = 2+\nu$ with parameter $2^{-\nu} + \sigma_p A_t$. Therefore, Lemma 1.2 and the induction assumption imply that

$$
\begin{aligned}
\psi_t(x) &\geq \psi_t^* + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu} \\
&\geq A_t \tilde{f}(x_t) + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu}.
\end{aligned} \tag{3.10}
$$

Therefore,

$$
\begin{aligned}
\psi_{t+1}^* &= \min_{x\in\mathrm{dom}\,\varphi}\{\psi_t(x) + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right]\} \\
&\geq \min_{x\in\mathrm{dom}\,\varphi}\{A_t\tilde{f}(x_t) + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu} \\
&\quad + a_t[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)]\} \\
&= \min_{x\in\mathrm{dom}\,\varphi}\{A_t f(x_t) + A_t\varphi(x_t) + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu} \\
&\quad + a_t[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)]\}.
\end{aligned}
$$

Now, using the convexity and differentiability of f and the fact that $g_{\varphi}(x_{t+1}) \in \partial\varphi(x_{t+1})$ we obtain

$$
\begin{aligned}
f(x_t) &\geq f(x_{t+1}) + \langle \nabla f(x_{t+1}), x_t - x_{t+1}\rangle, \\
\varphi(x_t) &\geq \varphi(x_{t+1}) + \langle g_{\varphi}(x_{t+1}), x_t - x_{t+1}\rangle,
\end{aligned}
$$

and $\varphi(x) \geq \varphi(x_{t+1}) + \langle g_{\varphi}(x_{t+1}), x - x_{t+1}\rangle$. Substituting these inequalities above, it follows that

$$
\begin{aligned}
\psi_{t+1}^* \geq \min_{x\in\mathrm{dom}\,\varphi}\{&A_{t+1}\tilde{f}(x_{t+1}) + \langle \nabla\tilde{f}(x_{t+1}), A_t x_t - A_t x_{t+1}\rangle \\
&+ a_t\langle \nabla\tilde{f}(x_{t+1}), x - x_{t+1}\rangle + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu}\}.
\end{aligned}
$$

Note that $y_t = (1-\alpha_t)x_t + \alpha_t v_t = \frac{A_t}{A_{t+1}}x_t + \frac{a_t}{A_{t+1}}v_t$. Hence, $A_t x_t = A_{t+1}y_t - a_t v_t$, and

$$
\begin{aligned}
\psi_{t+1}^* \geq \min_{x\in\mathrm{dom}\,\varphi}\{&A_{t+1}\tilde{f}(x_{t+1}) + \langle \nabla\tilde{f}(x_{t+1}), A_{t+1}y_t - a_t v_t - A_t x_{t+1}\rangle \\
&+ a_t\langle \nabla\tilde{f}(x_{t+1}), x - x_{t+1}\rangle + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu}\}.
\end{aligned}
$$

Further, $A_{t+1}x_{t+1} = A_t x_{t+1} + a_t x_{t+1}$. Hence,

$$
\begin{aligned}
\psi_{t+1}^* &\geq \min_{x\in\mathrm{dom}\,\varphi}\{A_{t+1}\tilde{f}(x_{t+1}) + A_{t+1}\langle \nabla\tilde{f}(x_{t+1}), y_t - x_{t+1}\rangle \\
&\quad + a_t\langle \nabla\tilde{f}(x_{t+1}), x - v_t\rangle + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu}\} \\
&\geq A_{t+1}\tilde{f}(x_{t+1}) + \min_{x\in\mathrm{dom}\,\varphi}\{A_{t+1}\left(\tfrac{1}{2M_t}\right)^{\frac{1}{1+\nu}}\|\nabla\tilde{f}(x_{t+1})\|_*^{\frac{2+\nu}{1+\nu}} \\
&\quad + a_t\langle \nabla\tilde{f}(x_{t+1}), x - v_t\rangle + \tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu}\},
\end{aligned}
$$

where the last inequality is due to (3.2). Thus, to prove that (3.7) is true for $t+1$, it is enough to show that

$$A_{t+1}\left(\tfrac{1}{2M_t}\right)^{\frac{1}{1+\nu}}\|\nabla\tilde{f}(x_{t+1})\|_*^{\frac{2+\nu}{1+\nu}}+a_t\langle\nabla\tilde{f}(x_{t+1}),x-v_t\rangle \tag{3.11}$$
$$+\tfrac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x-v_t\|^{2+\nu}\geq 0$$

for all $x\in\mathbb{E}$. Using Lemma 1.3 with $p=2+\nu$, $s=a_t\nabla\tilde{f}(x_{t+1})$ and $\omega=2^{-\nu}+\sigma_p A_t$, we see that a sufficient condition for (3.11) is

$$A_{t+1}\left(\tfrac{1}{2M_t}\right)^{\frac{1}{1+\nu}}\|\nabla\tilde{f}(x_{t+1})\|_*^{\frac{2+\nu}{1+\nu}}\quad\geq\quad\tfrac{(1+\nu)}{(2+\nu)}\left(\tfrac{1}{2^{-\nu}+\sigma_p A_t}\right)^{\frac{1}{1+\nu}}a_t^{\frac{2+\nu}{1+\nu}}\|\nabla\tilde{f}(x_{t+1})\|_*^{\frac{2+\nu}{1+\nu}},$$

that is,

$$A_{t+1}\left(\tfrac{1}{2M_t}\right)^{\frac{1}{1+\nu}}\quad\geq\quad\tfrac{(1+\nu)}{(2+\nu)}\left(\tfrac{1}{2^{-\nu}+\sigma_p A_t}\right)^{\frac{1}{1+\nu}}a_t^{\frac{2+\nu}{1+\nu}}, \tag{3.12}$$

which is equivalent to

$$a_t^{2+\nu}\quad\leq\quad\left(\tfrac{2+\nu}{1+\nu}\right)^{1+\nu}\tfrac{(2^{-\nu}+\sigma_p A_t)}{2M_t}A_{t+1}^{1+\nu}\;=\;\left(\tfrac{2+\nu}{1+\nu}\right)^{1+\nu}\tfrac{(2^{-\nu}+\sigma_p A_t)}{2M_t}(A_t+a_t)^{1+\nu}.$$

Note that, $\left(\frac{2+\nu}{1+\nu}\right)^{1+\nu}=2^{1+\nu}\left(1-\frac{\nu}{2(1+\nu)}\right)^{1+\nu}\geq 2^{\nu}$. Therefore, by (3.4) we have

$$a_t^{2+\nu}\quad=\quad\tfrac{(1+2^{\nu}\sigma_p A_t)}{2M_t}(A_t+a_t)^{1+\nu}\;\leq\;\left(\tfrac{2+\nu}{1+\nu}\right)^{1+\nu}\tfrac{(2^{-\nu}+\sigma_p A_t)}{2M_t}(A_t+a_t)^{1+\nu}.$$

Thus (3.7) is true for $t+1$, completing the induction argument.

Let us now estimate the growth of the coefficients A_t. Recall that, by assumption,

$$0<M_t\leq\gamma H_f(\nu),\quad\forall t\geq 0$$

for some constant $\gamma\geq 1$. Thus, if $\sigma_p=0$, it follows from (3.4) that $a_t^{2+\nu}\geq\frac{1}{2\gamma H_f(\nu)}(A_t+a_t)^{1+\nu}$. Hence,

$$A_{t+1}-A_t\quad=\quad a_t\;\geq\;\left(\tfrac{1}{2\gamma H_f(\nu)}\right)^{\frac{1}{2+\nu}}A_{t+1}^{\frac{1+\nu}{2+\nu}}. \tag{3.13}$$

Now, denoting $B_t=2\gamma H_f(\nu)A_t$ for all $t\geq 0$, it follows from (3.13) that,

$$B_{t+1}-B_t\quad\geq\quad B_{t+1}^{\frac{1+\nu}{2+\nu}}.$$

Then, by Lemma 1.4, with $\alpha=\frac{1+\nu}{2+\nu}$, we have

$$B_t\quad\geq\quad\left[\left(\tfrac{1}{2+\nu}\right)\left(\frac{B_1^{\frac{1}{2+\nu}}}{B_1^{\frac{1}{2+\nu}}+1}\right)^{\frac{1+\nu}{2+\nu}}\right]^{2+\nu}(t-1)^{2+\nu}\quad\forall t\geq 2.$$

Note that $A_1\geq\frac{1}{2\gamma H_f(\nu)}$. Thus, $B_1\geq 1$ and consequently

$$B_t\quad\geq\quad\left[\tfrac{1}{(2+\nu)}\left(\tfrac{1}{2}\right)^{\frac{1+\nu}{2+\nu}}\right]^{2+\nu}(t-1)^{2+\nu}.$$

Therefore, for all $t \geq 2$, $A_t \geq \frac{1}{2\gamma H_f(\nu)}\left[\frac{1}{(2+\nu)}\left(\frac{1}{2}\right)^{\frac{1+\nu}{2+\nu}}\right]^{2+\nu}(t-1)^{2+\nu}$.

On the other hand, if $\sigma_p > 0$, it follows from (3.4) that

$$(A_{t+1}-A_t)^{2+\nu} \;=\; a_t^{2+\nu} \;\geq\; \frac{(1+2^{\nu}\sigma_p A_t)}{2\gamma H_f(\nu)}(A_t+a_t)^{1+\nu}.$$

Thus,

$$\begin{aligned} 2^{\nu}\sigma_p A_t A_{t+1}^{1+\nu} &\leq A_{t+1}^{1+\nu}(1+2^{\nu}\sigma_p A_t) \;\leq\; 2\gamma H_f(\nu)(A_{t+1}-A_t)^{2+\nu} \\ &= 2\gamma H_f(\nu)\left[A_{t+1}^{\frac{1}{2}}-A_t^{\frac{1}{2}}\right]^{2+\nu}\left[A_{t+1}^{\frac{1}{2}}+A_t^{\frac{1}{2}}\right]^{2+\nu} \\ &\leq 2^{3+\nu}\gamma H_f(\nu)A_{t+1}^{\frac{2+\nu}{2}}\left[A_{t+1}^{\frac{1}{2}}-A_t^{\frac{1}{2}}\right]^{2+\nu}. \end{aligned}$$

Therefore, $\sigma_p A_t^{\frac{2+\nu}{2}} \leq \sigma_p A_t A_{t+1}^{\frac{\nu}{2}} \leq 8\gamma H_f(\nu)\left[A_{t+1}^{\frac{1}{2}}-A_t^{\frac{1}{2}}\right]^{2+\nu}$. Consequently,

$$\left(\frac{\sigma_p}{8\gamma H_f(\nu)}\right)^{\frac{1}{2+\nu}} A_t^{\frac{1}{2}} \;\leq\; A_{t+1}^{\frac{1}{2}} - A_t^{\frac{1}{2}}.$$

Hence, $A_{t+1} \geq A_t\left[1+\left(\frac{\sigma_p}{8\gamma H_f(\nu)}\right)^{\frac{1}{2+\nu}}\right]^2$. Since $A_1 \geq \frac{1}{2\gamma H_f(\nu)}$, it follows that

$$A_t \;\geq\; \left(\frac{1}{2\gamma H_f(\nu)}\right)\left[1+\left(\frac{\sigma_p}{8\gamma H_f(\nu)}\right)^{\frac{1}{2+\nu}}\right]^{2(t-1)},$$

and so, (3.8) holds.

Finally, by (3.7) and Lemma 3.1, for $t \geq 0$, we have

$$A_t\tilde{f}(x_t) \;\leq\; \psi_t^* \;\leq\; A_t\tilde{f}(x^*) + \frac{1}{2+\nu}\|x^*-x_0\|^{2+\nu}.$$

Hence, $A_t(\tilde{f}(x_t)-\tilde{f}(x^*)) \leq \frac{1}{2+\nu}\|x^*-x_0\|^{2+\nu}$, and (3.9) follows immediately from inequality (3.8). □

Algorithm 1 can be equipped with an implementable stopping criterion. Assume that $\frac{1}{(2+\nu)}\|x^*-x_0\|^{2+\nu} \leq D$ and that the constant D is known. Denote

$$\ell_t(y) = \sum_{i=0}^{t-1} a_i\left[f(x_{i+1}) + \langle \nabla f(x_{i+1}), x - x_{i+1}\rangle + \varphi(y)\right]$$

and $\hat{f}_t = \min\limits_{y\in\mathbb{E}}\left\{\frac{1}{A_t}\ell_t(y) : \; \frac{1}{(2+\nu)}\|y-x_0\|^{2+\nu} \leq D\right\}$. Then

$$\tilde{f}(x_t) \;\leq\; \frac{1}{A_t}\psi_t^* \;\leq\; \hat{f}_t + \frac{D}{A_t} \;\leq\; \tilde{f}(x^*) + \frac{D}{A_t}.$$

Thus, if $\frac{D}{A_t} \leq \epsilon$, then $\tilde{f}(x_t) - \tilde{f}(x^*) \leq \epsilon$, and we can use inequality

$$\tilde{f}(x_t) - \hat{f}_t \;\leq\; \epsilon$$

as a stopping criterion[3].

[3]We emphasize that the use of this stopping criterion depends strongly on the knowledge of a good upper bound D. Of course, if one takes D very large it is very likelly that $\frac{1}{(2+\nu)}\|x^*-x_0\|^{2+\nu} \leq D$ will be satisfied. However, with such a choice, the running time of the algorithm will be big.

Note that the key point in Algorithm 1 is how to compute M_t such that

$$0 < M_t \leq \gamma H_f(\nu) \tag{3.14}$$

for some constant $\gamma \geq 1$ independent of t, and for which condition (3.2) is satisfied. Let us look now at possible strategies for finding such values.

3.1. Constant $H_f(\nu)$ is known. If we assume that $H_f(\nu)$ is known, then in Algorithm 1 we can take

$$M_t = M \equiv (1+\nu)H_f(\nu) \quad \text{for all } t \geq 0,$$

which gives (3.14) with $\gamma = (1+\nu)$. Therefore, in view of the estimate (3.9), the corresponding scheme can find δ-solution of problem (2.1) in at most $\mathcal{O}(\delta^{-\frac{1}{2+\nu}})$ iterations if $\sigma_p = 0$, and in at most $\mathcal{O}(\log(\delta^{-1}))$ if $\sigma_p > 0$.

Note that for $\sigma_p = 0$, the computation of a_t and A_{t+1} in Algorithm 1 can be simplified. Indeed, note that in this method the equation (3.4) can be replaced by condition

$$a_t^{2+\nu} \quad \leq \quad \tfrac{1}{2M_t} A_{t+1}^{1+\nu}.$$

Denoting $B_t = 2M_t A_t$, we can see that the latter inequality is equivalent to the following:

$$B_{t+1} - B_t \leq B_{t+1}^{\frac{1+\nu}{2+\nu}} \quad \Longleftrightarrow \quad 1 - \tfrac{B_t}{B_{t+1}} \;\leq\; \left(\tfrac{1}{B_{t+1}}\right)^{\frac{1}{2+\nu}}.$$

It is clear that this inequality is valid for $B_t = \left(\frac{t}{2+\nu}\right)^{2+\nu}$. Indeed, in this case

$$\tfrac{B_t}{B_{t+1}} \quad = \quad \left(1 - \tfrac{1}{t+1}\right)^{2+\nu} \;\geq\; 1 - \tfrac{2+\nu}{t+1} \;=\; 1 - \left(\tfrac{1}{B_{t+1}}\right)^{\frac{1}{2+\nu}}.$$

Thus, we can take $A_t = \frac{1}{2M_t}\left(\frac{t}{2+\nu}\right)^{2+\nu}$ and define $a_t = A_{t+1} - A_t$.

Let us present now the corresponding version of Algorithm 1, which becomes a generalization of scheme (4.8) in [12].

Algorithm 2. Accelerated RNM with known $H_f(\nu)$ and $\sigma_p = 0$.

Initialization Choose $x_0 \in \mathrm{dom}\,\varphi$. Set $v_0 = x_0$ and $M = (1+\nu)H_f(\nu)$. Define $A_t = \frac{1}{2M}\left(\frac{t}{2+\nu}\right)^{2+\nu}$, $t \geq 0$.

Iteration $t \geq 0$**:**

a) Compute

$$
\begin{aligned}
x_{t+1} \;&=\; T_{\nu,M}(y_t) \;\equiv\; \arg\min_{x \in \mathrm{dom}\,\varphi} \Big\{ f(y_t) + \langle \nabla f(y_t), x - y_t \rangle \\
&\quad + \frac{1}{2}\langle \nabla^2 f(y_t)(x - y_t), x - y_t \rangle + \tfrac{M\|x - y_t\|^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(x) \Big\},
\end{aligned} \tag{3.15}
$$

where $y_t = v_t + \dfrac{A_t}{A_{t+1}}(x_t - v_t)$.

b) Set

$$
\psi_{t+1}(x) = \psi_t(x) + (A_{t+1} - A_t)\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \varphi(x)\right].
$$

and

$$
v_{t+1} = \arg\min_{x \in \mathrm{dom}\,\varphi} \psi_{t+1}(x).
$$

3.2. Adaptive estimate of $H_f(\nu)$. For real-life problems, usually we don't know the constant $H_f(\nu)$. In this case, we can consider the following adaptive strategy for estimating the unknown constant $H_f(\nu)$.

Algorithm 3. Accelerated RNM with adaptive estimate of $H_f(\nu)$

Initialization Choose $x_0 \in \operatorname{dom} \varphi$ and $H_0 \in (0, (1+\nu)H_f(\nu)]$. Set $v_0 = x_0$ and $A_0 = 0$.

Iteration $t \geq 0$:

a) Find the smallest integer $i_t \geq 0$ such that

$$\langle \nabla \tilde{f}(x_{t+1,i_t}), y_{t,i_t} - x_{t+1,i_t} \rangle \geq \left(\tfrac{1}{2(2^{i_t} H_t)} \right)^{\frac{1}{1+\nu}} \| \nabla \tilde{f}(x_{t+1,i_t}) \|_*^{\frac{2+\nu}{1+\nu}}. \tag{3.16}$$

where

(3.17)

$$x_{t+1} = T_{\nu, 2^{i_t} H_t}(y_t) \equiv \arg \min_{x \in \operatorname{dom} \varphi} \Big\{ f(y_{t,i_t}) + \langle \nabla f(y_{t,i_t}), x - y_{t,i_t} \rangle$$

$$+ \frac{1}{2} \langle \nabla^2 f(y_{t,i_t})(x - y_{t,i_t}), x - y_{t,i_t} \rangle + \tfrac{2^{i_t} H_t \|x - y_{t,i_t}\|^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(x) \Big\},$$

and $y_{t,i_t} = (1-\alpha_{t,i_t}) x_t + \alpha_{t,i_t} v_t$, with $\alpha_{t,i_t} = \frac{a_{t,i_t}}{A_t + a_{t,i_t}}$ and coefficient $a_{t,i_t} > 0$ computed from the equation

$$a_{t,i_t}^{2+\nu} = \tfrac{(1+2^\nu \sigma_p A_t)}{2(2^{i_t} H_t)} (A_t + a_{t,i_t})^{1+\nu} \tag{3.18}$$

b) Set $x_{t+1} = x_{t+1,i_t}$, $y_t = y_{t,i_t}$, $a_t = a_{t,i_t}$, $\alpha_t = \alpha_{t,i_t}$. Define $A_{t+1} = A_t + a_t$, $H_{t+1} = 2^{i_t - 1} H_t$,

$$\psi_{t+1}(x) = \psi_t(x) + a_t \left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \varphi(x) \right]. \tag{3.19}$$

and

$$v_{t+1} = \arg \min_{x \in \operatorname{dom} \varphi} \psi_{t+1}(x).$$

REMARK 3.3. *Although $H_f(\nu)$ appears in the Initialization step of Algorithm 3, in practice it is not used. In fact, even without knowing $H_f(\nu)$, if we compute*

$$H_0 = \|\nabla^2 f(x) - \nabla^2 f(y)\|$$

for $x, y \in \mathbb{E}$ with $\|x - y\| = 1$, then we have $0 < H_0 \leq H_f(\nu) \leq (1+\nu) H_f(\nu)$.

The next result gives convergence rates for Algorithm 3.

THEOREM 3.4. *Assume that $H_f(\nu) < +\infty$. Then, the scaling coefficients in Algorithm 3 satisfy condition*

$$0 < 2^{i_t} H_t \leq 2(1+\nu) H_f(\nu), \quad t \geq 0. \tag{3.20}$$

Consequently, we have

(3.21)

$$\tilde{f}(x_t) - \tilde{f}(x^*) \leq \begin{cases} \dfrac{4(1+\nu) H_f(\nu) (4+2\nu)^{1+\nu} \|x^* - x_0\|^{2+\nu}}{(t-1)^{2+\nu}}, & \forall t \geq 2 \text{ if } \sigma_p = 0, \\ \dfrac{4(1+\nu) H_f(\nu) \|x^* - x_0\|^{2+\nu}}{(2+\nu)} \left[1 + \left(\dfrac{\sigma_p}{16(1+\nu) H_f(\nu)} \right)^{\frac{1}{2+\nu}} \right]^{-2(t-1)}, & \forall t \geq 0 \text{ if } \sigma_p > 0. \end{cases}$$

Furthermore, the total numbers N_t of calls of oracle[4] *after t iterations of Algorithm 3 is bounded as follows:*

$$N_t \leq 2t + \log_2 \tfrac{2(1+\nu)H_f(\nu)}{H_0}. \tag{3.22}$$

Proof. The upper bound (3.20) follows from Lemma 2.1 and the backtracking strategy of the algorithm. Therefore, the rate of convergence (3.21) can be obtained by Theorem 3.2 with $\gamma = 2(1+\nu)$. Since $i_t = \log_2 \frac{2H_{t+1}}{H_t}$, we get the upper bound (3.22) for the total number of calls of oracle. □

REMARK 3.5. *From Theorem 3.3 we see that Algorithm 3 has the same rates of convergence as Algorithms 1 and 2, which use the exact value of the Hölder constant $H_f(\nu)$. However, by (3.22), Algorithm 3 needs on average twice the number of computations of the oracle per iteration.*

4. Universal accelerated scheme. As we saw, Algorithms 1-3 require the knowledge of the Hölder parameter ν. In this section we describe a universal scheme that works for any $\nu \in [0,1]$ without using it explicitly in the algorithm. The key to this "universal property" is Lemma 2.3, which garantees that even if we use the possible wrong value $\nu = 1$ in our regularized model for $\tilde{f}$, we still can obtain a descent condition. Regarding the estimating functions, now we shall start from

$$\psi_0(x) = \frac{1}{3}\|x - x_0\|^3.$$

Given an accuracy $\epsilon > 0$, from Lemma 1.5 recall the function

$$R(\epsilon) = \max_{x \in \operatorname{dom} \varphi} \{ \|x - x^*\| : \tilde{f}(x) \leq \tilde{f}(x^*) + \epsilon \}. \tag{4.1}$$

[4]By *calls of oracle* we mean the joint computation of $f(x)$, $\nabla f(x)$ and $\nabla^2 f(x)$.

Let us assume that $R(\epsilon) < +\infty$. Denote $\gamma_\nu(\epsilon) = \left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}} \left(\frac{R(\epsilon)}{\epsilon}\right)^{\frac{1-\nu}{1+\nu}}$.

Algorithm 4. Accelerated Universal CNM

Initialization Choose $x_0 \in \operatorname{dom}\varphi$ and $0 < H_0 \leq \inf\limits_{\nu\in[0,1]} \gamma_\nu(\epsilon)$.
Set $v_0 = x_0$ and $A_0 = 0$.

Iteration $t \geq 0$**:**

a) Find the smallest integer $i_t \geq 0$ such that

$$\langle \nabla \tilde{f}(x_{t+1,i_t}), y_{t,i_t} - x_{t+1,i_t}\rangle \quad \geq \quad \left(\frac{4}{3(2^{i_t}H_t)}\right)^{\frac{1}{2}} \|\nabla \tilde{f}(x_{t+1,i_t})\|_*^{\frac{3}{2}}. \tag{4.2}$$

where
(4.3)

$$\begin{aligned} x_{t+1,i_t} \quad &= \quad T_{1,2^{i_t}H_t}(y_t) \;\equiv\; \arg\min_{x\in\operatorname{dom}\varphi} \Big\{ f(y_{t,i_t}) + \langle \nabla f(y_{t,i_t}), x - y_{t,i_t}\rangle \\ &\quad + \frac{1}{2}\langle \nabla^2 f(y_{t,i_t})(x - y_{t,i_t}), x - y_{t,i_t}\rangle + \frac{2^{i_t}H_t\|x-y_{t,i_t}\|^3}{6} + \varphi(x)\Big\}, \end{aligned}$$

and $y_{t,i_t} = (1-\alpha_{t,i_t})x_t + \alpha_{t,i_t}v_t$, with $\alpha_{t,i_t} = \frac{a_{t,i_t}}{A_t + a_{t,i_t}}$ and coefficient $a_{t,i_t} > 0$ computed from the equation

$$a_{t,i_t}^3 \quad = \quad \frac{3}{4(2^{i_t}H_t)}(A_t + a_{t,i_t})^2, \tag{4.4}$$

b) Set $x_{t+1} = x_{t+1,i_t}$, $y_t = y_{t,i_t}$, $a_t = a_{t,i_t}$. Define $A_{t+1} = A_t + a_t$,
$H_{t+1} = 2^{i_t-1}H_t$,

$$\psi_{t+1}(x) = \psi_t(x) + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right]. \tag{4.5}$$

and

$$v_{t+1} = \arg\min_{x\in\operatorname{dom}\varphi} \psi_{t+1}(x).$$

To obtain convergence rates for Algorithm 4, we need the following corollary of Lemma 3.1.

LEMMA 4.1. *For all $t \geq 0$ and $x \in \operatorname{dom}\varphi$, we have*

$$\psi_t(x) \quad \leq \quad A_t\tilde{f}(x) + \tfrac{1}{3}\|x - x_0\|^3. \tag{4.6}$$

Proof. It can accomplished as the proof of Lemma 3.1 with $\nu = 1$. □

THEOREM 4.2. *Assume that $H_f(\nu) < +\infty$ for some $\nu \in [0,1]$. Let the sequence $\{x_t\}_{t=0}^T$ be generated by Algorithm 4 and suppose that for $i = 0,\ldots,i_t$ and $t = 0,\ldots,T$ we have:*

$$\tilde{f}(T_{1,2^iH_t}(y_{t,i})) - \tilde{f}(x^*) \quad \geq \quad \epsilon \tag{4.7}$$

Then, for $t = 2,\ldots,T$, we have $H_t \leq \gamma_\nu(\epsilon)$ and

$$\tilde{f}(x_t) - \tilde{f}(x^*) \quad \leq \quad \frac{96\gamma_\nu(\epsilon)\|x_0 - x^*\|^3}{(t-1)^3}. \tag{4.8}$$

Therefore,

$$T \quad \leq \quad 1 + \tfrac{14}{3}\|x_0 - x^*\| \inf_{\nu\in[0,1]} \left[\tfrac{12H_f(\nu)R(\epsilon)^{\frac{1-\nu}{2}}}{(1+\nu)(2+\nu)\epsilon}\right]^{\frac{2}{3(1+\nu)}}. \tag{4.9}$$

Proof. Firstly, let us prove that the sequence $\{x_t\}_{t=0}^T$ is well defined. In view of Lemma 1.5, at any test point x of the algorithm the norm of the gradient is big enough:

$$\|\nabla \tilde{f}(x)\|_* \quad \geq \quad \tfrac{\epsilon}{R(\epsilon)}.$$

Thus, by Lemma 2.3, the search procedure at each iteration of Algorithm 4 is finite. In particular, we can guarantee that $2^{i_t}H_t \leq 2\gamma_\nu(\epsilon)$. Consequently, inequality $H_t \leq \gamma_\nu(\epsilon)$ can be justified by induction.

Now, let us prove by induction that

$$A_t\tilde{f}(x_t) \quad \leq \quad \psi_t^* \equiv \min_{x\in\operatorname{dom}\varphi} \psi_t(x). \tag{4.10}$$

For $t = 0$ this is evident: $A_0\tilde{f}(x_0) = 0 = \min\limits_{x\in\operatorname{dom}\varphi} \psi_0(x)$. Assume that (4.10) is true for some $t \geq 0$. Note that, for any $x \in \operatorname{dom}\varphi$ we have

$$\begin{aligned} \psi_t(x) \quad &= \quad \textstyle\sum_{i=0}^{t-1} a_i\left[f(x_{i+1}) + \langle \nabla f(x_{i+1}), x - x_{i+1}\rangle + \varphi(x)\right] + \tfrac{1}{3}\|x - x_0\|^3 \\ &= \quad \ell_t(x) + \tfrac{1}{3}\|x - x_0\|^3 \end{aligned}$$

for all $t = 1,\ldots,T$. Note that $\ell_t(x)$ is a linear function. Moreover, by Lemma 4 in [12], $\frac{1}{3}\|x - x_0\|^3$ is a uniformly convex function of degree $p = 3$ with parameter $\sigma_p = \frac{1}{2}$. Thus, $\psi_t(x)$ is also a uniformly convex function of degree $p = 3$ with parameter $\sigma_p = \frac{1}{2}$. Therefore, Lemma 1.2 and the induction assumption imply that

$$\psi_t(x) \quad \geq \quad \psi_t^* + \tfrac{1}{6}\|x - v_t\|^3 \;\geq\; A_t\tilde{f}(x_t) + \tfrac{1}{6}\|x - v_t\|^3. \tag{4.11}$$

Therefore,

$$\begin{aligned} \psi_{t+1}^* \quad &= \quad \min_{x\in\operatorname{dom}\varphi} \left\{\psi_t(x) + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right]\right\} \\ &\geq \quad \min_{x\in\operatorname{dom}\varphi} \{A_t\tilde{f}(x_t) + \tfrac{1}{6}\|x - v_t\|^3 \\ &\qquad + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right]\} \\ &= \quad \min_{x\in\operatorname{dom}\varphi} \{A_t f(x_t) + A_t\varphi(x_t) + \tfrac{1}{6}\|x - v_t\|^3 \\ &\qquad + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right]\}. \end{aligned}$$

Now, using the convexity and differentiability of f and the fact that $g_\varphi(x_{t+1}) \in \partial\varphi(x_{t+1})$, we obtain

$$\begin{aligned} f(x_t) \quad &\geq \quad f(x_{t+1}) + \langle \nabla f(x_{t+1}), x_t - x_{t+1}\rangle, \\ \varphi(x_t) \quad &\geq \quad \varphi(x_{t+1}) + \langle g_\varphi(x_{t+1}), x_t - x_{t+1}\rangle, \\ \varphi(x) \quad &\geq \quad \varphi(x_{t+1}) + \langle g_\varphi(x_{t+1}), x - x_{t+1}\rangle. \end{aligned}$$

Substituting these inequalities in the above relation, we get

$$\begin{aligned}\psi^*_{t+1} \;\geq\; & \min_{x\in \mathrm{dom}\,\varphi}\{A_{t+1}\tilde f(x_{t+1}) + \langle \nabla \tilde f(x_{t+1}), A_t x_t - A_t x_{t+1}\rangle \\ & + a_t\langle \nabla \tilde f(x_{t+1}), x - x_{t+1}\rangle + \tfrac{1}{6}\|x - v_t\|^3\}.\end{aligned}$$

Note that $y_t = (1-\alpha_t)x_t + \alpha_t v_t = \frac{A_t}{A_{t+1}}x_t + \frac{a_t}{A_{t+1}}v_t$. Hence $A_t x_t = A_{t+1}y_t - a_t v_t$ and

$$\begin{aligned}\psi^*_{t+1} \;\geq\; & \min_{x\in \mathrm{dom}\,\varphi}\{A_{t+1}\tilde f(x_{t+1}) + \langle \nabla \tilde f(x_{t+1}), A_{t+1} y_t - a_t v_t - A_t x_{t+1}\rangle \\ & + a_t\langle \nabla \tilde f(x_{t+1}), x - x_{t+1}\rangle + \tfrac{1}{6}\|x - v_t\|^3\}.\end{aligned}$$

Now, note that $A_{t+1}x_{t+1} = A_t x_{t+1} + a_t x_{t+1}$. Hence,

$$\begin{aligned}\psi^*_{t+1} \;\geq\; & \min_{x\in \mathrm{dom}\,\varphi}\{A_{t+1}\tilde f(x_{t+1}) + A_{t+1}\langle \nabla \tilde f(x_{t+1}), y_t - x_{t+1}\rangle \\ & + a_t\langle \nabla \tilde f(x_{t+1}), x - v_t\rangle + \tfrac{1}{6}\|x - v_t\|^3\} \\ \geq\; & A_{t+1}\tilde f(x_{t+1}) + \min_{x\in \mathrm{dom}\,\varphi}\{A_{t+1}\left(\tfrac{2}{3(2^{i_t}H_t)}\right)^{\frac{1}{2}}\|\nabla \tilde f(x_{t+1})\|_*^{\frac{3}{2}} \\ & + a_t\langle \nabla \tilde f(x_{t+1}), x - v_t\rangle + \tfrac{1}{6}\|x - v_t\|^3\},\end{aligned}$$

where the last inequality is due to (4.2). Thus, for proving that (4.10) is true for $t+1$, it is enough to show that for all $x \in \mathbb{E}$ we have

$$A_{t+1}\left(\tfrac{2}{3(2^{i_t}H_t)}\right)^{\frac{1}{2}}\|\nabla \tilde f(x_{t+1})\|_*^{\frac{3}{2}} + a_t\langle \nabla \tilde f(x_{t+1}), x - v_t\rangle + \tfrac{\|x-v_t\|}{6}^3 \;\geq\; 0. \tag{4.12}$$

Using Lemma 1.3 with $p=3$, $s = a_t\nabla f(x_{t+1})$, and $\omega = \frac{1}{2}$, we see that necessary and sufficient condition for (4.10) is

$$A_{t+1}\left(\tfrac{2}{3(2^{i_t}H_t)}\right)^{\frac{1}{2}}\|\nabla \tilde f(x_{t+1})\|_*^{\frac{3}{2}} \;\geq\; \tfrac{2\sqrt{2}}{3}a_t^{\frac{3}{2}}\|\nabla \tilde f(x_{t+1})\|_*^{\frac{3}{2}}.$$

That is $A_{t+1}\left(\frac{2}{3(2^{i_t}H_t)}\right)^{\frac{1}{2}} \geq \frac{2\sqrt{2}}{3}a_t^{\frac{3}{2}}$, which is equivalent to $a_t^3 \leq \frac{3}{4(2^{i_t}H_t)}A_{t+1}^2$. Therefore, (4.10) is true for $t+1$ due to (4.4), completing our proof by induction.

Let us now estimate the growth of the coefficients A_t. By (4.4) and the bound $2^{i_t}H_t \leq 2\gamma_\nu(\epsilon)$, we have

$$a_t^3 \;=\; \tfrac{3}{4(2^{i_t}H_t)}A_{t+1}^2 \;\geq\; \tfrac{3}{8\gamma_\nu(\epsilon)}A_{t+1}^2.$$

Consequently,

$$A_{t+1} - A_t \;\geq\; \left(\tfrac{3}{8\gamma_\nu(\epsilon)}\right)^{\frac{1}{3}} A_{t+1}^{\frac{2}{3}}. \tag{4.13}$$

Now, denoting $B_t = \frac{8}{3}\gamma_\nu(\epsilon)A_t$ it follows from (4.13) that $B_{t+1} - B_t \geq B_{t+1}^{\frac{2}{3}}$ for $t \geq 0$. As $A_0 = 0$, we have $B_0 = 0$, which in the previous inequality implies that $B_1 \geq 1$. Then, by Lemma 1.4, with $\alpha = 2/3$, we have

$$B_t \;\geq\; \left[\tfrac{1}{3}\left(\tfrac{1}{2}\right)^{\frac{2}{3}}\right]^3 (t-1)^3, \quad t \geq 1.$$

Therefore, for all $t \geq 2$, we have $A_t \geq \frac{3}{8\gamma_{\nu}(\epsilon)}\frac{(t-1)^3}{108} = \frac{(t-1)^3}{288}$. Recall that from Lemma 4.1 and (4.10) it follows that

$$A_t\tilde{f}(x_t) \;\leq\; \psi_t^* \quad\leq\quad A_t\tilde{f}(x^*) + \tfrac{1}{3}\|x^* - x_0\|^3$$

Therefore, for $t \geq 2$ we have

$$\tilde{f}(x_t) - \tilde{f}(x^*) \quad\leq\quad \frac{96\gamma_{\nu}(\epsilon)\|x_0 - x^*\|^3}{(t-1)^3}. \tag{4.14}$$

Finally, by (4.7) and (4.14) we have

$$\epsilon \quad\leq\quad \tilde{f}(x_t) - \tilde{f}(x^*) \;\leq\; \frac{96\gamma_{\nu}(\epsilon)\|x_0 - x^*\|^3}{(t-1)^3}, \qquad t = 2, \ldots, T.$$

Therefore,

$$\begin{aligned}(T-1)^3 \quad&\leq\quad \frac{96}{\epsilon}\left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}}\left(\frac{R(\epsilon)}{\epsilon}\right)^{\frac{1-\nu}{1+\nu}}\|x_0 - x^*\|^3\\ &=\quad 96\left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}}\left(\frac{1}{\epsilon}\right)^{\frac{2}{1+\nu}} R(\epsilon)^{\frac{1-\nu}{1+\nu}}\|x_0 - x^*\|^3,\end{aligned}$$

which implies (4.9). We can put inf there since the scheme of Algorithm 4 does not depend on ν. □

REMARK 4.3. *From Theorem 4.1 it follows that Algorithm 4 can find an ϵ-solution of problem (2.1) in at most $\mathcal{O}\left(\frac{1}{\epsilon^{2/[3(1+\nu)]}}\right)$ iterations, which is slightly worse than the bound of $\mathcal{O}\left(\frac{1}{\epsilon^{1/(1+\nu)}}\right)$ iterations obtained for Algorithms 1 to 3. This is a moderate price to pay for the absence of perfect information about ν.*

COROLLARY 4.4. *Let function $\tilde{f}$ be uniformly convex of degree p with constant $\sigma_p > 0$. Then the number of iterations in Algorithm 4 is bounded as follows:*

$$T \quad\leq\quad 1 + \tfrac{14}{3}\|x_0 - x^*\| \inf_{\nu\in[0,1]}\left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\left(\frac{p}{\sigma_p}\right)^{\frac{1-\nu}{2p}}\right]^{\frac{2}{3(1+\nu)}}\left(\frac{1}{\epsilon}\right)^{\frac{2p+\nu-1}{3p(1+\nu)}}. \tag{4.15}$$

Proof. Indeed, in view of Lemma 1.2, for any $x \in \operatorname{dom}\varphi$ with $\tilde{f}(x) - \tilde{f}(x^*) \leq \epsilon$ we have $\epsilon \geq \frac{\sigma_p}{p}\|x - x^*\|^p$. Therefore, in this case $R(\epsilon) \leq \left(\frac{\epsilon p}{\sigma_p}\right)^{\frac{1}{p}}$. It remains to use the upper bound (4.9). □

4.1. Computational Issues. For starting Algorithm 4, it is necessary to ensure the initial condition

$$0 < H_0 \quad\leq\quad \gamma_{\nu}^*(\epsilon) \;\stackrel{\text{def}}{=}\; \inf_{\nu\in[0,1]}\gamma_{\nu}(\epsilon).$$

Usually this is not difficult since typically the values $\gamma_{\nu}(\epsilon)$ are big. However, we can use a more sophisticated procedure.

Using (4.1), define $D = R(\tilde{f}(x_0) - \tilde{f}(x^*))$ and $\hat{\gamma}_{\nu}(\epsilon) = \left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}}\left(\frac{D}{\epsilon}\right)^{\frac{1-\nu}{1+\nu}}$. Let $\hat{\gamma}_{\nu}^*(\epsilon) \stackrel{\text{def}}{=} \inf_{\nu\in[0,1]}\hat{\gamma}_{\nu}(\epsilon)$. This is an upper bound for $\gamma_{\nu}^*(\epsilon)$, which could be used in the right-hand side of inequality (4.8). For that, we need to start Algorithm 4 with $H_0 < \hat{\gamma}_{\nu}^*(\epsilon)$.

Let us show how this can be done. Take a point $y_0 \neq x_0$ such that $f(y_0) \leq f(x_0)$ and $\nabla^2 f(y_0) \neq \nabla^2 f(x_0)$. Then, define

$$\Delta \;=\; \|\nabla^2 f(y_0) - \nabla^2 f(x_0)\|, \quad r = \|y_0 - x_0\|.$$

Now we can choose

$$H_0 \;=\; \min_{\nu \in [0,1]} \left[\tfrac{12\Delta}{(1+\nu)(2+\nu) r^\nu} \right]^{\frac{2}{1+\nu}} \left(\tfrac{r}{\epsilon}\right)^{\frac{1-\nu}{1+\nu}} \;\leq\; \hat{\gamma}_\nu^*(\epsilon).$$

Taking the logarithm of the objective function in this minimization problem, we get

$$\tfrac{1}{1+\nu} \left[2\left(\ln(12\Delta) - \ln(1+\nu) - \ln(2+\nu) - \nu \ln r \right) + (1-\nu) \ln \tfrac{r}{\epsilon} \right].$$

This is a ratio of convex function and a positive linear function in ν. Thus, it is quasi-convex and its global minimum can be easily approximated by bisection algorithm.

Finally, as in Algorithm 1, we can also consider a proper stopping criterion in Algorithm 4. Denote

$$\ell_t(y) \;=\; \sum_{i=0}^{t-1} a_t \left[f(x_{i+1}) + \langle \nabla f(x_{i+1}), x - x_{i+1} \rangle + \varphi(y) \right].$$

Assume that, $\frac{1}{3}\|x^* - x_0\|^3 \leq D$ and that constant D is known. Denote

$$\hat{f}_t \;=\; \min_{y \in \operatorname{dom} \varphi} \left\{ \tfrac{1}{A_t} \ell_t(y) : \; \tfrac{1}{3}\|y - x^*\|^3 \leq D \right\}.$$

Then, as in Section 2, we can see that

$$\tilde{f}(x_t) \;\leq\; \tfrac{1}{A_t}\psi_t^* \;\leq\; \hat{f}_t + \tfrac{D}{A_t} \;\leq\; \tilde{f}(x^*) + \tfrac{D}{A_t}.$$

So, if $A_t \geq \frac{D}{\epsilon}$, then $\tilde{f}(x_t) - \tilde{f}(x^*) \leq \epsilon$, and we can use inequality

$$\tilde{f}(x_t) - \hat{f}_t \;\leq\; \epsilon$$

as a stopping criterion for Algorithm 4.

5. Conclusion. In this paper, we presented accelerated versions of the regularized Newton methods for solving convex composite minimization problems, where the second part of the objective is a simple closed convex function. We assume that the Hessian of the smooth part of the objective is Hölder-continuous. For the case in which the the Hölder parameter $\nu \in [0,1]$ is known, we propose methods with worst-case complexity of $\mathcal{O}\left(\frac{1}{\epsilon^{1/(2+\nu)}}\right)$ iterations, generalizing the results in [12]. For the general case, in which the ν is not known, we propose a universal method which ensures the same precision in at most $\mathcal{O}\left(\frac{1}{\epsilon^{2/[3(1+\nu)]}}\right)$ iterations.

Our problem setting includes, for example, piece-wise linear norms used in regularization techniques and also the indicator function of a closed convex set, making our schemes suitable for several applications (see, for example, [10, 13]).

Acknowledgments. The authors are very grateful to Clóvis C. Gonzaga and to Elizabeth W. Karas for their support and warm hospitality.

REFERENCES

[1] Birgin, E.G., Martínez, J.M.: The use of quadratic regularization with cubic descent condition for unconstrained optimization. SIAM Journal on Optimization **27**, 1049-1074 (2017)

[2] Cartis, C., Gould, N.I.M., and Toint, Ph.L.: Adaptive cubic regularization methods for unconstrained optimization. Part II: worst-case function - and derivative - evaluation complexity. Mathematical Programming **130**, 295-319 (2011)

[3] Cartis, C., Gould, N.I.M., and Toint, Ph.L.: Universal regularized methods - varying the power, the smoothness, and the accuracy. Optimization Online (2016).

[4] Curtis, F.E., Robinson, D.P., Samadi, M.: A trust-region algorithm with a worst-case complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization. Mathematical Programming **162**, 1-32 (2017)

[5] Dussault, J.-P.: ARCq: a new adaptive regularization by cubics variant. Optim.Methods Softw. doi:10. 1080/10556788.2017.1322080 (2017)

[6] Grapiglia, G.N., Nesterov, Yu.: Regularized Newton Methods for minimizing functions with Hölder continuous Hessians. SIAM Journal on Optimization **27**, 478-506 (2017)

[7] Bergou, E., Diouane, Y., Gratton, S.: A line-search algorithm inspired by the adaptive cubic regularization framework with a worst-case complexity $O(\epsilon^{-3/2})$. Optimization Online (2017)

[8] Martínez, J.M., Raydan, M.: Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization. Journal of Global Optimization **68**, 367-385 (2017)

[9] Martínez, J.M.: On high-order model regularization for constrained optimization. Optimization Online (2017).

[10] Nesterov, Yu.: Smooth minimization of non-smooth functions. Mathematical Programming **103**, 127-152 (2005)

[11] Nesterov, Yu, Polyak, B.T.: Cubic regularization of Newton method and its global performance. Mathematical Programming **108**, 177-205 (2006)

[12] Nesterov, Yu.: Accelerating the cubic regularization of Newton's method on convex problems. Mathematical Programming **112**, 159-181 (2008)

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

[14] Nesterov, Yu.: Universal gradient methods for convex optimization problems. Mathematical Programming **152**, 381-404 (2015)