SIAM J. OPTIM.
Vol. 29, No. 1, pp. 77–99

© 2019 Society for Industrial and Applied Mathematics

ACCELERATED REGULARIZED NEWTON METHODS FOR MINIMIZING COMPOSITE CONVEX FUNCTIONS[*]

GEOVANI N. GRAPIGLIA[†] AND YURII 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 $\mathcal{O}\big(\frac{1}{\epsilon^{1/(2+\nu)}}\big)$ 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 $\mathcal{O}\big(\frac{1}{\epsilon^{2/[3(1+\nu)]}}\big)$ iterations without using ν explicitly in the scheme.

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

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

DOI. 10.1137/17M1142077

1. Introduction.

1.1. Motivation.

Following the worst-case complexity analysis presented in [11] for a cubic regularization of the 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 (RNMs) were proposed for unconstrained minimization of a twice-differentiable function with Hölder-continuous Hessians. Some of these methods are "universal" in the sense that they do not require 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 RNMs 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 twice-differentiable convex function with a 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 in [12] for the accelerated cubic regularization

[*] Received by the editors August 4, 2017; accepted for publication (in revised form) October 29, 2018; published electronically January 8, 2019.
http://www.siam.org/journals/siopt/29-1/M114207.html

Funding: The first author was supported by the National Council for Scientific and Technological Development (CNPq), Brazil, under grants 401288/2014-5 and 406269/2016-5. The second author was partially supported by the National Council for Scientific and Technological Development (CNPq), Brazil, grants 401288/2014-5 and 314908/2014-5, and by RSF grant 17-11-01027.

[†] Departamento de Matemática, Universidade Federal do Paraná, Centro Politécnico, 81531-980, Curitiba, Paraná, Brazil (grapiglia@ufpr.br).

[‡] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 1348 Louvain-la-Neuve, Belgium, and National Research University Higher School of Economics, 101000 Moscow, Russia (Yurii.Nesterov@uclouvain.be).

of Newton's method with $\nu = 1$. For the general case, in which parameter 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. In [6], the complexity bounds obtained when ν is known and when ν is not known are the same in terms of the dependence on ϵ. Interestingly, our acceleration technique does not provide the same matching: the complexity bound obtained when ν is known is slightly better than the complexity bound obtained when ν is not known. This justifies the presentation of schemes that deal separately with these two scenarios.

1.2. Contents.

The paper is organized as follows. In section 2, we define our problem. 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 a small residual in the function value.[1] At the end of the paper we also include an appendix with auxiliary results and the main inequalities related to the Hölder continuity of the Hessians of the first term in our objective.

1.3. Notation and generalities.

In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space, and by $\mathbb{E}^*$ its *dual* space, composed of 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 $f : \mathbb{E} \to \mathbb{R}$:

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

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

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

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 $f : \mathbb{E} \to \mathbb{R}$:

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

Operator $B : \mathbb{E} \to \mathbb{E}^*$ is *positive definite* if $\langle Bx, x\rangle > 0$ for $x \in \mathbb{E} \setminus \{0\}$ (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:

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

2. Problem statement.

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).

[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.

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

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

It follows from (2.2) and from an integral form of the mean-value theorem that

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

and

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

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

$$Q(x;y) = f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle,$$

$$M_{\nu,H}(x;y) = Q(x;y) + \frac{H\|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(y), \;\; y \in \operatorname{dom}\varphi,$$

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

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

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:

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

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

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

then by the above inequality we have

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

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

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

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).

3. Numerical schemes for ν known. In this section we consider minimization schemes to solve problem (2.1) when the Hölder parameter ν is known. For example, the solution of a system of linear inequalities can be formulated as an optimization problem whose objective function is convex and has Hölder-continuous Hessian with known parameter $\nu \neq 1$ (see Example 1 in [6]). We also assume that the 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] Let us start with a generic framework. At the beginning of the tth iteration one has an estimate x_t for the solution of (2.1), an auxiliary vector v_t, constant $A_t > 0$, and an approximation $M_t > 0$ for $H_f(\nu)$. Then a new vector y_t is computed as a suitable convex combination of x_t and v_t:

$$y_t = (1-\alpha_t)x_t + \alpha_t v_t. \tag{3.1}$$

Constant α_t plays a crucial role in the acceleration process and is defined as

$$\alpha_t = a_t/(A_t + a_t), \tag{3.2}$$

where the coefficient $a_t > 0$ is computed from the equation

$$a_t^{2+\nu} = \frac{(1+2^{\nu}\sigma_p A_t)}{2M_t}(A_t + a_t)^{1+\nu}. \tag{3.3}$$

We compute a trial point x_t^+ by minimizing the regularized model around y_t, that is,

$$x_t^+ = T_{\nu,M_t}(y_t) \equiv \arg\min_{x \in \operatorname{dom}\varphi} M_{\nu,M_t}(y_t;x). \tag{3.4}$$

If the descent condition

$$\langle \nabla \tilde{f}(x_t^+), y_t - x_t^+ \rangle \geq \left(\frac{1}{2M_t}\right)^{\frac{1}{1+\nu}} \|\nabla \tilde{f}(x_t^+)\|_*^{\frac{2+\nu}{1+\nu}} \tag{3.5}$$

is satisfied, then x_t^+ is accepted and we define $x_{t+1} = x_t^+$. Otherwise, constant M_t is increased until the corresponding trial point x_t^+ is accepted. We assume that there exists $M > 0$ such that $M_t \leq M$ for all t. After obtaining x_{t+1}, we set $A_{t+1} = A_t + a_t$ and compute

$$v_{t+1} = \arg\min_{x \in \operatorname{dom}\varphi} \psi_{t+1}(x), \tag{3.6}$$

where

$$\psi_{t+1}(x) = \psi_t(x) + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right]. \tag{3.7}$$

As we will see, function $\psi_t(\cdot)$ is a proper lower approximation of a multiple of our objective function $A_t\tilde{f}(\cdot)$ augmented by the proximal term $\psi_0(\cdot)$. It aggregates all the information on the objective accumulated after t calls of the oracle. To initialize, we set $A_0 = 0$, $v_0 = x_0$, and $\psi_0(x) = \frac{1}{2+\nu}\|x - x_0\|^{2+\nu}$ for $x_0 \in \operatorname{dom}\varphi$.

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

Algorithm 1 Accelerated RNM for known parameter ν.

Step 0. Choose $x_0 \in \operatorname{dom} \varphi$. Set $v_0 = x_0$, $A_0 = 0$, and $t := 0$.
Step 1. Find $0 < M_t \leq M$, such that

$$\langle \nabla \tilde{f}(x_t^+), y_t - x_t^+ \rangle \geq \left(\frac{1}{2M_t} \right)^{\frac{1}{1+\nu}} \| \nabla \tilde{f}(x_t^+) \|_*^{\frac{2+\nu}{1+\nu}}, \tag{3.8}$$

where $x_t^+ = T_{\nu, M_t}(y_t)$ with y_t being defined by (3.1)–(3.3).
Step 2. Set $x_{t+1} = x_t^+$ and $A_{t+1} = A_t + a_t$, with $a_t > 0$ obtained from (3.3).
Step 3. Define $\psi_{t+1}(\,.\,)$ by (3.7) and compute v_{t+1} by (3.6).
Step 4. Set $t := t + 1$ and go back to Step 1.

Remark 3.1. *At first, the particular definition of the elements in Algorithm* 1 *may seem very mysterious. In this regard, it is worth mentioning that Algorithm* 1 *is developed in the spirit of the estimating sequences technique* [10]. *This means that our accelerated scheme aims to generate sequences* $\{x_t\}_{t=0}^{\infty}$, $\{A_t\}_{t=0}^{\infty}$, *and* $\{\psi_t(\,.\,)\}_{t=0}^{\infty}$ *in such a way that they satisfy the relations*

$$A_t \tilde{f}(x_t) \leq \min_{x \in \mathbb{E}} \psi_t(x) \ \ \forall t \geq 0 \tag{3.9}$$

and

$$\psi_t(x) \leq A_t \tilde{f}(x) + \frac{1}{2+\nu} \|x - x_0\|^{2+\nu} \ \ \forall x \in \mathbb{E}. \tag{3.10}$$

Combining the above inequalities, it follows that

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

and so

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

Thus, the rate of growth of coefficients $\{A_t\}_{t=0}^{\infty}$ *defines the rate of convergence of the method.*

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

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

$$\psi_t(x) \leq A_t \tilde{f}(x) + \frac{1}{(2+\nu)} \|x - x_0\|^{2+\nu} \ \forall x \in \mathbb{E}. \tag{3.11}$$

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

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

Thus, (3.11) is true for $t = 0$. Suppose that (3.11) is true for some $t \geq 0$. Then (3.7) and convexity of f imply that, for all $x \in \mathbb{E}$,

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

Thus, (3.11) is also true for $t + 1$. ◻

Now we are in a position to prove that the sequences in Algorithm 1 satisfy (3.9). By combining (3.9) with (3.11) we obtain global complexity rates for Algorithm 1.

THEOREM 3.3. *Assume that $H_f(\nu) < +\infty$ for some $\nu \in [0, 1]$. If sequence $\{x_t\}_{t=0}^{\infty}$ is generated by Algorithm 1 with $0 < M_t \leq M$, then for all $t \geq 0$ we have*

$$A_t\tilde{f}(x_t) \leq \psi_t^* \equiv \min_{x\in\mathbb{E}} \psi_t(x). \tag{3.12}$$

Moreover,

$$A_t \geq \begin{cases} \frac{1}{2M}\left[\frac{1}{(2+\nu)}\left(\frac{1}{2}\right)^{\frac{1+\nu}{2+\nu}}\right]^{2+\nu}(t-1)^{2+\nu} \quad \forall t \geq 2 \quad \textit{if } \sigma_p = 0,\\ \left(\frac{1}{2M}\right)\left[1 + \left(\frac{\sigma_p}{8M}\right)^{\frac{1}{2+\nu}}\right]^{2(t-1)} \quad \forall t \geq 0 \quad \textit{if } \sigma_p > 0. \end{cases} \tag{3.13}$$

Consequently, we have

$$\tilde{f}(x_t) - \tilde{f}^* \leq \begin{cases} \frac{2M(4+2\nu)^{1+\nu}\|x^* - x_0\|^{2+\nu}}{(t-1)^{2+\nu}} \;\forall t \geq 2 \quad \textit{if } \sigma_p = 0,\\ \frac{2M\|x^* - x_0\|^{2+\nu}}{(2+\nu)}\left[1 + \left(\frac{\sigma_p}{8M}\right)^{\frac{1}{2+\nu}}\right]^{-2(t-1)} \;\forall t \geq 0 \quad \textit{if } \sigma_p > 0, \end{cases} \tag{3.14}$$

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

Proof. Let us prove relation (3.12) by induction over t. Since $A_0 = 0$, for $t = 0$ it is evident that

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

Assume that (3.12) is true for some $t \geq 0$. Note that, for any $x \in \mathbb{E}$,

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

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 A.2 in Appendix A and the induction assumption imply that

$$\begin{aligned} \psi_t(x) &\geq \psi_t^* + \frac{(2^{-\nu} + \sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu} \\ &\geq A_t\tilde{f}(x_t) + \frac{(2^{-\nu} + \sigma_p A_t)}{(2+\nu)}\|x - v_t\|^{2+\nu}. \end{aligned} \tag{3.15}$$

Therefore,

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

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

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

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

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

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

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

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

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

where the last inequality is due to (3.5). Thus, to prove that (3.12) is true for $t+1$, it is enough to show that

$$\begin{aligned}&A_{t+1}\left(\frac{1}{2M_t}\right)^{\frac{1}{1+\nu}}\|\nabla\tilde{f}(x_{t+1})\|_*^{\frac{2+\nu}{1+\nu}} + a_t\langle\nabla\tilde{f}(x_{t+1}), x - v_t\rangle \\ &\qquad + \frac{(2^{-\nu}+\sigma_p A_t)}{(2+\nu)}\|x-v_t\|^{2+\nu} \geq 0\end{aligned} \tag{3.16}$$

for all $x \in \mathbb{E}$. Using Lemma A.3 in Appendix A 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 (3.16) is

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

that is,

$$A_{t+1}\left(\frac{1}{2M_t}\right)^{\frac{1}{1+\nu}} \geq \frac{(1+\nu)}{(2+\nu)}\left(\frac{1}{2^{-\nu}+\sigma_p A_t}\right)^{\frac{1}{1+\nu}} a_t^{\frac{2+\nu}{1+\nu}}, \tag{3.17}$$

which is equivalent to

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

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 (3.3) we have

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

Thus (3.12) is true for $t+1$, completing the induction argument.

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

$$0 < M_t \leq M \quad \forall t \geq 0.$$

Thus, if $\sigma_p = 0$, it follows from (3.3) that $a_t^{2+\nu} \geq \frac{1}{2M}(A_t+a_t)^{1+\nu}$. Hence,

$$A_{t+1} - A_t = a_t \geq \left(\frac{1}{2M}\right)^{\frac{1}{2+\nu}} A_{t+1}^{\frac{1+\nu}{2+\nu}}. \tag{3.18}$$

Now, denoting $B_t = 2MA_t$ for all $t \geq 0$, it follows from (3.18) that

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

Then, by Lemma A.4 in Appendix A, with $\alpha = \frac{1+\nu}{2+\nu}$, we have

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

Note that $A_1 \geq \frac{1}{2M}$. Thus, $B_1 \geq 1$ and consequently

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

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

On the other hand, if $\sigma_p > 0$, it follows from (3.3) that

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

Thus,

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

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

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

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

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

and so (3.13) holds.

Finally, by (3.12) and Lemma 3.2, 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 (3.14) follows immediately from inequality (3.13). □

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]

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

$$0 < M_t \leq M \tag{3.19}$$

for some constant $M > 0$ independent of t, and for which condition (3.5) is satisfied. Let us look now at possible strategies for finding such values.

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

$$M_t = M \equiv (1+\nu)H_f(\nu) \quad \forall t \geq 0.$$

Therefore, in view of the estimate (3.14), the corresponding scheme can find a 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 (3.3) can be replaced by condition

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

Denoting $B_t = 2M_tA_t$, we can see that the latter inequality is equivalent to

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

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

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

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]; see Algorithm 2.

[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 likely 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 long.

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

Step 0. Choose $x_0 \in \operatorname{dom}\varphi$. Define $\psi_0(x) = \frac{1}{2+\nu}\|x - x_0\|^{2+\nu}$. Set $v_0 = x_0$ and $M = (1+\nu)H_f(\nu)$. Define $A_k = \frac{1}{2M}\left(\frac{k}{2+\nu}\right)^{2+\nu}$ for all $k \geq 0$, and set $t := 0$.

Step 1. Compute $y_t = v_t + \frac{A_t}{A_{t+1}}(x_t - v_t)$.

Step 2. Compute

$$\begin{aligned} x_{t+1} = T_{\nu,M}(y_t) \equiv \arg\min_{x \in \operatorname{dom}\varphi} \Big\{ & f(y_t) + \langle \nabla f(y_t), x - y_t\rangle \\ & + \frac{1}{2}\langle \nabla^2 f(y_t)(x - y_t), x - y_t\rangle + \frac{M\|x - y_t\|^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(x)\Big\}. \end{aligned} \tag{3.20}$$

Step 3. 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 compute

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

Step 4. Set $t := t + 1$ and go back to Step 1.

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 adaptive strategy shown in Algorithm 3 for estimating the unknown constant $H_f(\nu)$.

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 the condition*

$$0 < 2^{i_t} H_t \leq 2(1+\nu)\max\{H_f(\nu), H_0\}, \quad t \geq 0. \tag{3.21}$$

Consequently, (3.14) *holds for $M = 2(1+\nu)\max\{H_f(\nu), H_0\}$. Furthermore, the total number N_t of calls to the oracle*[4] *after t iterations of Algorithm* 3 *is bounded as follows:*

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

Proof. Let us prove by induction that the scaling coefficients H_t in Algorithm 3 satisfy $H_t \leq (1+\nu)\max\{H_f(\nu), H_0\}$ for all $t \geq 0$. Indeed, this is true for H_0. Assume that this inequality holds for some $t \geq 0$. In view of Lemma A.6 in Appendix A, the final value $2^{i_t}H_t$ cannot be bigger than $2(1+\nu)\max\{H_f(\nu), H_0\}$, since otherwise we should stop the line-search process earlier. Therefore,

$$H_{t+1} = \frac{1}{2}2^{i_t}H_t \leq (1+\nu)\max\{H_f(\nu), H_0\},$$

completing the induction argument. Thus,

$$0 < 2^{i_t}H_t = 2H_{t+1} \leq 2(1+\nu)\max\{H_f(\nu), H_0\},$$

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

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

Step 0. Choose $x_0 \in \operatorname{dom}\varphi$ and $H_0 > 0$. Define $\psi_0(x) = \frac{1}{2+\nu}\|x - x_0\|^{2+\nu}$. Set $v_0 = x_0$, $A_0 = 0$, and $t := 0$.

Step 1. Set $i := 0$.
Step 1.1. Compute the coefficient $a_{t,i} > 0$ by solving the equation

$$a_{t,i}^{2+\nu} = \frac{(1 + 2^{\nu}\sigma_p A_t)}{2(2^i H_t)}(A_t + a_{t,i})^{1+\nu}. \tag{3.23}$$

Step 1.2. Set $\alpha_{t,i} = \frac{a_{t,i}}{A_t + a_{t,i}}$ and compute

$$y_{t,i} = (1 - \alpha_{t,i})x_t + \alpha_{t,i}v_t.$$

Step 1.3. Compute

$$\begin{aligned} x_{t,i}^{+} = T_{\nu,2^i H_t}(y_{t,i}) \;\equiv\; \arg\min_{x\in\operatorname{dom}\varphi} \Big\{ & f(y_{t,i}) + \langle \nabla f(y_{t,i}), x - y_{t,i}\rangle \\ & + \frac{1}{2}\langle \nabla^2 f(y_{t,i})(x - y_{t,i}), x - y_{t,i}\rangle + \frac{2^i H_t \|x - y_{t,i}\|^{2+\nu}}{(1+\nu)(2+\nu)} + \varphi(x) \Big\}. \end{aligned} \tag{3.24}$$

Step 1.4. If

$$\langle \nabla \tilde{f}(x_{t,i}^{+}), y_{t,i} - x_{t,i}^{+}\rangle \geq \left(\frac{1}{2(2^i H_t)}\right)^{\frac{1}{1+\nu}} \|\nabla \tilde{f}(x_{t,i}^{+})\|_{*}^{\frac{2+\nu}{1+\nu}}, \tag{3.25}$$

set $i_t := i$ and go to Step 2. Otherwise, set $i := i + 1$ and go back to Step 1.1.
Step 2. Set $x_{t+1} = x_{t,i_t}^{+}$, $y_t = y_{t,i_t}$, $a_t = a_{t,i_t}$, and $\alpha_t = \alpha_{t,i_t}$. Define $A_{t+1} = A_t + a_t$ and $H_{t+1} = 2^{i_t - 1}H_t$.
Step 3. Set

$$\psi_{t+1}(x) = \psi_t(x) + a_t\left[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle + \varphi(x)\right] \tag{3.26}$$

and compute

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

Step 4. Set $t := t + 1$ and go back to Step 1.

and, by Theorem 3.3, (3.14) holds with

$$M = 2(1+\nu)\max\left\{H_f(\nu), H_0\right\}.$$

Finally, note that at each iteration the oracle is called $i_t + 1$ times. Since $i_t - 1 = \log_2 \frac{H_{t+1}}{H_t}$, we get upper bound (3.22) for the total number of calls of the 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 A.8 in Appendix A, which guarantees that even when $\nu \neq 1$ we still can obtain a descent condition by using a cubic regularized model for $\tilde{f}$. 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 A.5 in Appendix A we shall use the function

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

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

Algorithm 4 Accelerated universal cubic RNM.

Step 0. Choose $x_0 \in \operatorname{dom} \varphi$ and $H_0 > 0$. Define $\psi_0(x) = \frac{1}{3}\|x - x_0\|^3$. Set $v_0 = x_0$, $A_0 = 0$, and $t := 0$.

Step 1. Set $i := 0$.
Step 1.1. Compute the coefficient $a_{t,i} > 0$ by solving the equation

$$a_{t,i}^3 = \frac{3}{4(2^i H_t)}(A_t + a_{t,i})^2. \tag{4.2}$$

Step 1.2. Set $\alpha_{t,i} = \frac{a_{t,i}}{A_t + a_{t,i}}$ and compute

$$y_{t,i} = (1 - \alpha_{t,i}) x_t + \alpha_{t,i} v_t.$$

Step 1.3. Compute

$$\begin{aligned} x_{t,i}^+ = T_{1,2^i H_t}(y_{t,i}) \ \equiv \ \arg \min_{x \in \operatorname{dom} \varphi} \Big\{ & f(y_{t,i}) + \langle \nabla f(y_{t,i}), x - y_{t,i} \rangle \\ & + \frac{1}{2}\langle \nabla^2 f(y_{t,i})(x - y_{t,i}), x - y_{t,i} \rangle + \frac{2^i H_t \|x - y_{t,i}\|^3}{6} + \varphi(x) \Big\}. \end{aligned} \tag{4.3}$$

Step 1.4. If

$$\langle \nabla \tilde{f}(x_{t,i}^+), y_{t,i} - x_{t,i}^+ \rangle \geq \left(\frac{4}{3(2^i H_t)}\right)^{\frac{1}{2}} \|\nabla \tilde{f}(x_{t,i}^+)\|_*^{\frac{3}{2}}, \tag{4.4}$$

set $i_t := i$ and go to Step 2. Otherwise, set $i := i + 1$ and go back to Step 1.1.
Step 2. Set $x_{t+1} = x_{t,i_t}^+$, $y_t = y_{t,i_t}$, $a_t = a_{t,i_t}$, and $\alpha_t = \alpha_{t,i_t}$. Define $A_{t+1} = A_t + a_t$ and $H_{t+1} = 2^{i_t - 1} H_t$.
Step 3. Set

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

and compute

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

Step 4. Set $t := t + 1$ and go back to Step 1.

To obtain convergence rates for Algorithm 4, we need the following corollary of Lemma 3.2.

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

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

Proof. It can be accomplished as in the proof of Lemma 3.2 with $\nu = 1$. □

Now we can obtain global complexity rates for Algorithm 4 by adapting the proof of Theorem 3.3.

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*

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

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

$$\tilde{f}(x_t) - \tilde{f}(x^*) \leq \frac{96 \max\{\gamma_\nu(\epsilon), H_0\}\, \|x_0 - x^*\|^3}{(t-1)^3}. \tag{4.8}$$

Therefore,

$$T \leq 1 + \frac{14}{3}\|x_0 - x^*\| \max\left\{\inf_{\nu \in [0,1]} \left[\frac{12 H_f(\nu) R(\epsilon)^{\frac{1-\nu}{2}}}{(1+\nu)(2+\nu)\epsilon}\right]^{\frac{2}{3(1+\nu)}}, \left(\frac{H_0}{\epsilon}\right)^{\frac{1}{3}}\right\}. \tag{4.9}$$

Proof. First, let us prove that the sequence $\{x_t\}_{t=0}^{T}$ is well defined. In view of Lemma A.5, at any test point x of the algorithm the norm of the gradient is big enough:

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

Thus, by Lemma A.8, the search procedure at each iteration of Algorithm 4 is finite. In particular, we can guarantee that $2^{i_t} H_t \leq 2 \max\{\gamma_\nu(\epsilon), H_0\}$. Consequently, inequality $H_t \leq \max\{\gamma_\nu(\epsilon), H_0\}$ can be justified by induction.

Now, let us prove by induction that

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

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

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

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 A.2 and the induction assumption imply that

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

Therefore,

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

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

$$f(x_t) \geq f(x_{t+1}) + \langle \nabla f(x_{t+1}), x_t - x_{t+1}\rangle,$$

$$\varphi(x_t) \geq \varphi(x_{t+1}) + \langle g_\varphi(x_{t+1}), x_t - x_{t+1}\rangle,$$
$$\varphi(x) \geq \varphi(x_{t+1}) + \langle g_\varphi(x_{t+1}), x - x_{t+1}\rangle.$$

Substituting these inequalities into the above relation, we get

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

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

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

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

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

where the last inequality is due to (4.4). Thus, to prove that (4.10) is true for $t+1$, it is enough to show that for all $x \in \mathbb{E}$ we have

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

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

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

That is, $A_{t+1}\big(\frac{2}{3(2^{i_t}H_t)}\big)^{\frac{1}{2}} \geq \frac{2\sqrt{2}}{3}a_t^{\frac{3}{2}}$, 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 (4.2), completing our proof by induction.

Let us now estimate the growth of the coefficients A_t. By (4.2) and the bound $2^{i_t}H_t \leq 2\max\{\gamma_\nu(\epsilon), H_0\}$, we have

$$a_t^3 = \frac{3}{4(2^{i_t}H_t)}A_{t+1}^2 \geq \frac{3}{8\max\{\gamma_\nu(\epsilon), H_0\}}A_{t+1}^2.$$

Consequently,

$$A_{t+1} - A_t \geq \left(\frac{3}{8\max\{\gamma_\nu(\epsilon), H_0\}}\right)^{\frac{1}{3}}A_{t+1}^{\frac{2}{3}}. \tag{4.13}$$

Now, denoting $B_t = \frac{8}{3}\max\{\gamma_\nu(\epsilon), H_0\}A_t$, 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 A.4, with $\alpha = 2/3$, we have

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

Therefore, for all $t \geq 2$, we have $A_t \geq \frac{3}{8\max\{\gamma_\nu(\epsilon),H_0\}}\frac{(t-1)^3}{108} = \frac{(t-1)^3}{288\max\{\gamma_\nu(\epsilon),H_0\}}$. Recall that from Lemma 4.1 and (4.10) it follows that

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

Therefore, for $t \geq 2$ we have

$$\tilde{f}(x_t) - \tilde{f}(x^*) \leq \frac{96\max\{\gamma_\nu(\epsilon), H_0\}\|x_0 - x^*\|^3}{(t-1)^3}. \tag{4.14}$$

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

$$\epsilon \leq \tilde{f}(x_t) - \tilde{f}(x^*) \; \leq \; \frac{96\max\{\gamma_\nu(\epsilon), H_0\}\|x_0 - x^*\|^3}{(t-1)^3}, \quad t = 2, \ldots, T.$$

Therefore,

$$\begin{aligned}(T-1)^3 &\leq \frac{96}{\epsilon}\max\left\{\left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}}\left(\frac{R(\epsilon)}{\epsilon}\right)^{\frac{1-\nu}{1+\nu}}, H_0\right\}\|x_0 - x^*\|^3 \\ &= 96\max\left\{\left[\frac{12H_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}}\left(\frac{1}{\epsilon}\right)^{\frac{2}{1+\nu}}R(\epsilon)^{\frac{1-\nu}{1+\nu}}, \frac{H_0}{\epsilon}\right\}\|x_0 - x^*\|^3,\end{aligned}$$

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* 4.2 *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 $\mathcal{O}\left(\frac{1}{\epsilon^{1/(2+\nu)}}\right)$ iterations obtained for Algorithms* 1–3. *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:*

(4.15)

$$T \leq 1 + \frac{14}{3}\|x_0 - x^*\| \max\left\{ \inf_{\nu\in[0,1]} \left[\frac{12 H_f(\nu)}{(1+\nu)(2+\nu)}\left(\frac{p}{\sigma_p}\right)^{\frac{1-\nu}{2p}}\right]^{\frac{2}{3(1+\nu)}} \left(\frac{1}{\epsilon}\right)^{\frac{2p+\nu-1}{3p(1+\nu)}}, \left(\frac{H_0}{\epsilon}\right)^{\frac{1}{3}} \right\}.$$

Proof. Indeed, in view of Lemma A.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). ∎

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

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

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

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

Then, as in section 2, we can see that

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

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 present 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, piecewise 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]).

Appendix A. Auxiliary results.

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

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

$$f((1-\theta)x+\theta y)\le(1-\theta)f(x)+\theta f(y)-\frac{\sigma_p\theta(1-\theta)}{p}\|y-x\|^p.$$

The 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.

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

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

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

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

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

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

and so

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

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

LEMMA A.3. *For any $h\in\mathbb{E}$, $s\in\mathbb{E}^*$, $p\ge 2$, and $\omega>0$, we have*

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

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 is crucial for establishing the complexity results for our accelerated schemes.

LEMMA A.4. *Let $\alpha\in[0,1)$, and suppose that $\{B_t\}_{t\ge0}$ is a sequence of nonnegative numbers with $B_t>0$, $t\ge1$, and*

$$B_{t+1}-B_t\ge B_{t+1}^{\alpha}\quad\forall t\ge0.$$

Then $B_t\ge\left[(1-\alpha)\left(\frac{B_1^{1-\alpha}}{B_1^{1-\alpha}+1}\right)^{\alpha}\right]^{1/(1-\alpha)}(t-1)^{\frac{1}{1-\alpha}}$ for all $t\ge2$.

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

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

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

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

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

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

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,

$$\text{(A.2)} \qquad (B_t + B_{t+1}^{\alpha})^{1-\alpha} - B_t^{1-\alpha} \geq (1-\alpha)(B_t + B_{t+1}^{\alpha})^{-\alpha} B_{t+1}^{\alpha}.$$

Combining (A.1) and (A.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

$$\text{(A.3)} \qquad \begin{aligned} (B_{t+1}^{1-\alpha} - B_t^{1-\alpha})^{\frac{1}{\alpha}} &\geq (1-\alpha)^{\frac{1}{\alpha}} \frac{B_{t+1}}{(B_t + B_{t+1}^{\alpha})} \geq (1-\alpha)^{\frac{1}{\alpha}} \frac{B_{t+1}}{(B_{t+1} + B_{t+1}^{\alpha})} \\ &= (1-\alpha)^{\frac{1}{\alpha}} \frac{1}{1 + B_{t+1}^{\alpha-1}} \geq (1-\alpha)^{\frac{1}{\alpha}} \frac{1}{1 + B_1^{\alpha-1}}, \end{aligned}$$

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

$$\text{(A.4)} \qquad B_{t+1}^{1-\alpha} - B_t^{1-\alpha} \geq (1-\alpha)\left(\frac{B_1^{1-\alpha}}{B_1^{1-\alpha}+1}\right)^{\alpha} \quad \forall t \geq 1.$$

Finally, it follows from (A.4) that, for all $t \geq 2$,

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

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

The next lemma gives us a lower bound on the size of subgradients of convex functions.

Lemma A.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–Schwarz inequality, and the above inequalities that

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

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 A.6. *Let $x_+ = T_{\nu,H}(\bar{x})$ for some $\bar{x} \in \operatorname{dom} \varphi$. If $H \geq (1+\nu)H_f(\nu)$, then*

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

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

$$\|\nabla f(x_+) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x_+ - \bar{x})\|_*^2 \leq \frac{H_f(\nu)^2 r^{2(1+\nu)}}{(1+\nu)^2}. \tag{A.5}$$

On the other hand, by (2.9)

$$\nabla f(\bar{x}) + \nabla^2 f(\bar{x})(x_+ - x) + \frac{H}{1+\nu} r^\nu B(x_+ - x) + g_\varphi(x_+) = 0. \tag{A.6}$$

Thus, combining (A.5) and (A.6), we get

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

Hence,

$$\langle \nabla \tilde{f}(x_+), \bar{x} - x_+ \rangle \geq \frac{(1+\nu)}{2Hr^{\nu}} \|\nabla \tilde{f}(x_+)\|_*^2 + \frac{1}{2(1+\nu)H}(H^2 - H_f(\nu)^2) r^{2+\nu}. \tag{A.7}$$

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 (A.7) as a function of r:

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

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

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

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

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

□

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, i.e., methods that do not require the value of the Hölder parameter to be implemented.

Lemma A.7. *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*

$$\|\nabla \tilde{f}(x_+)\|_* \geq \delta \quad \textit{and} \quad H \geq \left[\frac{CH_f(\nu)}{(1+\nu)(2+\nu)}\right]^{\frac{2}{1+\nu}} \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}}, \tag{A.8}$$

with constant $C \geq 6$, then

$$\|x_+ - \bar{x}\|^{1-\nu} \geq \frac{CH_f(\nu)}{(1+\nu)(2+\nu)H} \tag{A.9}$$

and, consequently,

$$\tilde{f}(x_+) \leq M_{1,H}(\bar{x}, x_+). \tag{A.10}$$

Proof. For $\nu = 1$ the statement is trivial. Assume that $\nu \in [0, 1)$. Denote $r = \|x_+ - \bar{x}\|$. Then the first inequality in (A.8) and inequalities (2.4) and (2.9) imply that

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

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

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

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 (A.8). Therefore, (A.9) holds. Note that if H satisfies the second inequality in (A.8), then $H \geq H_f(\nu)$. Thus, combining (2.5) and (A.9), we obtain (A.10):

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

Using Lemma A.7, we can modify Lemma A.6 in the following way.

LEMMA A.8. *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*

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

then

$$\langle \nabla \tilde{f}(x_+), \bar{x} - x_+\rangle \geq \sqrt{\frac{4}{3H}} \|\nabla \tilde{f}(x_+)\|_*^{\frac{3}{2}}. \tag{A.11}$$

Proof. Denote $r = \|x_+ - \bar{x}\|$. Then, by Lemma A.7 (with $C = 12$),

$$\|\nabla f(x_+) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x_+ - x)\|_* \leq \frac{H_f(\nu) r^{1+\nu}}{1+\nu} \leq \frac{H}{4} r^2. \tag{A.12}$$

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

$$\nabla f(\bar{x}) + \nabla^2 f(\bar{x})(x_+ - x) + \frac{H}{2} r B(x_+ - x) + g_\varphi(x_+) = 0. \tag{A.13}$$

Thus, combining (A.12) and (A.13), we get

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

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,

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

Acknowledgments. The authors are very thankful to two anonymous referees, whose comments significantly improved the readability of the text. The authors also are grateful to Clóvis C. Gonzaga and to Elizabeth W. Karas for their support and warm hospitality.

REFERENCES

[1] E.G. Birgin and J.M. Martínez, *The use of quadratic regularization with cubic descent condition for unconstrained optimization*, SIAM J. Optim., 27 (2017), pp. 1049–1074, https://doi.org/10.1137/16M110280X.

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

[3] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *Universal regularized methods: Varying the power, the smoothness, and the accuracy*, Optim. Methods and Softw., to appear.

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

[5] J.-P. Dussault, ARC$_\mathrm{q}$*: A new adaptive regularization by cubics variant*, Optim. Methods Softw., 33 (2018), pp. 322–335, https://doi.org/10.1080/10556788.2017.1322080.

[6] G.N. Grapiglia and Yu. Nesterov, *Regularized Newton methods for minimizing functions with Hölder continuous Hessians*, SIAM J. Optim., 27 (2017), pp. 478–506, https://doi.org/10.1137/16M1087801.

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

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

[9] J.M. Martínez, *On high-order model regularization for constrained optimization*, SIAM J. Optim., 27 (2017), pp. 2447–2458, https://doi.org/10.1137/17M1115472.

[10] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[11] Yu. Nesterov and B.T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205.

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

[13] Yu. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.