

# ADAPTIVE THIRD-ORDER METHODS FOR COMPOSITE CONVEX OPTIMIZATION

Geovani N. Grapiglia, Yurii Nesterov

REPRINT | 3313

## CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: [lidam-library@uclouvain.be](mailto:lidam-library@uclouvain.be)

<https://uclouvain.be/en/research-institutes/lidam/core/core-reprints.html>

## ADAPTIVE THIRD-ORDER METHODS FOR COMPOSITE CONVEX OPTIMIZATION\*

G. N. GRAPIGLIA<sup>†</sup> AND YU. NESTEROV<sup>‡</sup>

**Abstract.** In this paper we propose third-order methods for composite convex optimization problems in which the smooth part is a three-times continuously differentiable function with Lipschitz continuous third-order derivatives. The methods are adaptive in the sense that they do not require knowledge of the Lipschitz constant. Trial points are computed by the inexact minimization of models that consist in the nonsmooth part of the objective plus a quartic regularization of third-order Taylor polynomial of the smooth part. Specifically, approximate solutions of the auxiliary problems are obtained by using a Bregman gradient method as inner solver. Different from existing adaptive approaches for high-order methods, in our new schemes the regularization parameters are tuned by checking the progress of the inner solver. With this technique, we show that the basic method finds an  $\epsilon$ -approximate minimizer of the objective function performing at most  $\mathcal{O}(|\log(\epsilon)|\epsilon^{-\frac{1}{3}})$  iterations of the inner solver. An accelerated adaptive third-order method is also presented with total inner iteration complexity of  $\mathcal{O}(|\log(\epsilon)|\epsilon^{-\frac{1}{4}})$ .

**Key words.** composite minimization, third-order methods, worst-case complexity bounds

**MSC codes.** 49M15, 49M37, 58C15, 90C25, 90C30

**DOI.** 10.1137/22M1480872

### 1. Introduction.

**1.1. Motivation.** The worst-case complexity analysis of methods based on regularized Taylor models has been a very active research area in recent years (see, e.g., [1, 3, 4, 5, 7, 8, 10, 15, 18, 23]). In these methods, trial points are obtained by the minimization of  $(p+1)$ -regularizations of the  $p$ th-order Taylor approximations of the objective function ( $p \geq 2$ ). To ensure fast global convergence rates, it is enough to compute suitable approximate stationary points of the models [3]. Moreover, if the objective function is convex, then any  $(p+1)$ -regularization of the corresponding  $p$ th-order Taylor approximation is also convex when the regularization parameter is sufficiently large [21]. Remarkably, for  $p = 3$ , additional relative smoothness properties [2, 17] were proved in [21] for third-order tensor models. These properties allow the use of Bregman gradient methods to approximately minimize third-order tensor models with linear rate of convergence [21].

The interest in third-order methods is motivated by their improved worst-case complexity in comparison with first- and second-order methods. For example, it was shown in [21] that the basic third-order method takes at most  $\mathcal{O}(L_3^{1/3}\epsilon^{-1/3})$  iterations and calls of the oracle to find an  $\epsilon$ -approximate minimizer ( $\epsilon > 0$ ), where  $L_3$  is the

\*Received by the editors February 25, 2022; accepted for publication (in revised form) January 18, 2023; published electronically August 2, 2023.

<https://doi.org/10.1137/22M1480872>

**Funding:** This research was financed by European Research Council Advanced Grant 788368. The first author was also partially supported by the National Council for Scientific and Technological Development (grant 312777/2020-5).

<sup>†</sup>Université catholique de Louvain, ICTEAM/INMA, Avenue Geroges Lemaître 4-6/L40501, B-1348 Louvain-la-Neuve, Belgium (geovani.grapiglia@uclouvain.be).

<sup>‡</sup>Université catholique de Louvain, CORE, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium (Yurii.Nesterov@uclouvain.be).

Lipschitz constant of the third-order derivative of the objective function. In contrast, for the cubic regularization of the Newton's method (case  $p = 2$ ) an upper complexity bound of  $\mathcal{O}(L_2^{1/2}\epsilon^{-1/2})$  was proved in [23], with  $L_2$  being the Lipschitz constant of the Hessian of the objective function. With respect to  $\epsilon$ , the two bounds differ by a factor of  $\mathcal{O}(\epsilon^{-1/6})$ , which can be sufficiently large when  $\epsilon$  is close to zero. This means that third-order methods are particularly relevant to problems that need to be solved to high accuracy, such as decision-making problems. Moreover, even when a moderately accurate solution is sought, third-order methods may also outperform second-order methods when  $L_3^{1/3}$  is smaller than  $L_2^{1/2}$ .

The fast minimization of third-order tensor models is crucial for the development of practical third-order methods. However, the linear convergence of the Bregman gradient method applied to these models can be guaranteed only when the regularization parameter is bigger than a certain multiple of the Lipschitz constant of the third-order derivative of the objective function. The simplest way to satisfy this condition is to use explicitly such Lipschitz constant, when it is known. As pointed out in [21], this requirement is one of the main limitations for practical implementations of third-order tensor methods, since in general Lipschitz constants of third-order derivatives are unknown. In this paper we address this problem. Specifically, we propose adaptive third-order methods for composite convex optimization problems. The methods are adaptive in the sense that they do not require knowledge of the Lipschitz constant. Trial points are computed using a Bregman gradient method as inner solver. Our main contribution is a novel way to update the regularization parameters in the third-order models used to compute trial points. Different from existing adaptive tensor methods [12, 13, 16], in our new methods the regularization parameters are tuned by checking the progress of the inner solver. When the regularization parameter is sufficiently large, the inner solver is guaranteed to have a linear rate of convergence. We design an implementable way to check the violation of linear convergence. Thus, once a slow convergence rate is detected, the execution of the inner solver is stopped and the regularization parameter is increased. In standard adaptive third-order methods for convex optimization (such as the ones in [12, 13]), the computation of each trial point takes at most  $\mathcal{O}(\epsilon^{-4})$  iterations of the Bregman gradient method, where  $\epsilon \in (0, 1)$  is the desired accuracy for the norm of the gradient of the objective function (Theorem 3.8 in [11]). Our new adaptive technique brings this complexity down to  $\mathcal{O}(|\log(\epsilon)|)$ . Using this result, we show that our basic adaptive third-order method finds an  $\epsilon$ -approximate minimizer of the objective function performing at most  $\mathcal{O}(|\log(\epsilon)|\epsilon^{-\frac{1}{3}})$  iterations of the inner solver, while our accelerated adaptive third-order method has total inner iteration complexity of  $\mathcal{O}(|\log(\epsilon)|\epsilon^{-\frac{1}{4}})$ . Preliminary numerical results are also reported and illustrate the feasibility of our new adaptive technique.

**1.2. Contents.** The paper is organized as follows. In section 2 we define our problem and present the main results about the Bregman gradient method applied to the minimization of third-order tensor models. In section 3 we present and analyze the basic adaptive third-order method. Section 4 is dedicated to the accelerated adaptive third-order method. Illustrative numerical results are reported in section 5. Finally, in section 6 we discuss the possibility of second-order implementations of our adaptive schemes. For the reader interested in a rapid overview of the present work, the main results of this paper are Lemma 2.8, Theorem 2.9, Corollary 3.5 (with Remark 3.6), Corollary 4.5 and Lemma 4.6 (with Remark 4.7).

**1.3. Notation.** In what follows,  $\|\cdot\|$  denotes the Euclidean norm in  $\mathbb{R}^n$ , or the matrix and tensor norms induced by it. For a set  $S \subset \mathbb{R}^n$ ,  $\text{conv}(S)$  denotes its

convex hull. Given a matrix  $A \in \mathbb{R}^{n \times n}$ ,  $\text{trace}(A) = \sum_{i=1}^n A_{ii}$ . Moreover, for a smooth function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , we denote by  $D^3 f(x)$  the third-order derivative of  $f(\cdot)$  at point  $x$ , which is a tensor defined by

$$[D^3 f(x)]_{ijk} = \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(x), \quad i, j, k \in \{1, \dots, n\}.$$

## 2. Problem statement and auxiliary results.

**2.1. Basic third-order method.** In this paper we consider methods for solving the following composite minimization problem:

$$(2.1) \quad \min_{x \in \mathbb{R}^n} \tilde{f}(x) \equiv f(x) + \psi(x),$$

where  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is a convex three times differentiable function and  $\psi: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is a simple convex function. More specifically, we will assume that the third-order derivative of  $f(\cdot)$  is  $L_f$ -Lipschitz continuous, i.e.,

$$(2.2) \quad \|D^3 f(x) - D^3 f(y)\| \leq L_f \|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

Given  $x \in \mathbb{R}^n$ , the third-order Taylor polynomial of  $f(\cdot)$  around  $x$  is given by

$$(2.3) \quad \Phi_x(y) = f(x) + \sum_{i=1}^3 \frac{1}{i!} D^i f(x) [y - x]^i, \quad y \in \mathbb{R}^n.$$

It follows from (2.2) and (2.3) that

$$(2.4) \quad |f(y) - \Phi_x(y)| \leq \frac{L_f}{4!} \|y - x\|^4,$$

$$(2.5) \quad \|\nabla f(y) - \nabla \Phi_x(y)\| \leq \frac{L_f}{3!} \|y - x\|^3,$$

and

$$(2.6) \quad \|\nabla^2 f(y) - \nabla^2 \Phi_x(y)\| \leq \frac{L_f}{2!} \|y - x\|^2.$$

Define  $d_4: \mathbb{R}^n \rightarrow \mathbb{R}$  by

$$(2.7) \quad d_4(x) = \frac{1}{4} \|x\|^4,$$

and consider the following model of  $\tilde{f}(\cdot)$  around  $x$ :

$$(2.8) \quad \tilde{\Omega}_{x,M}(y) = \Omega_{x,M}(y) + \psi(y), \quad y \in \text{dom } \psi,$$

where

$$(2.9) \quad \Omega_{x,M}(y) = \Phi_x(y) + \frac{M}{2} d_4(y - x),$$

with  $M > 0$  being an estimate of  $L_f$ . If  $M \geq L_f$ , then it follows from (2.4) that

$$\begin{aligned} \tilde{f}(y) &\leq \Phi_x(y) + \frac{4L_f}{4!} d_4(y - x) + \psi(y) \\ &= \Phi_x(y) + \frac{M}{3!} d_4(y - x) + \psi(y) \\ &\leq \Phi_x(y) + \frac{M}{2!} d_4(y - x) + \psi(y) \\ &= \tilde{\Omega}_{x,M}(y). \end{aligned} \quad (2.10)$$

This remark motivates the following third-order method to solve (2.1):

$$(2.11) \quad x_{t+1} \in \arg \min_{y \in \mathbb{R}^n} \tilde{\Omega}_{x_t, M}(y), \quad t \geq 0.$$

Given,  $x \in \mathbb{R}^n$ , notice that  $\Omega_{x, M}(\cdot)$  in (2.9) is a multivariate polynomial. Therefore, in general,  $\tilde{\Omega}_{x, M}(\cdot)$  in (2.8) is a nonconvex function, which makes its exact minimization in (2.11) out of reach. Remarkably, if  $M \geq L_f$ , it follows from Theorem 1 in [21] that  $\tilde{\Omega}_{x, M}(\cdot)$  is convex, allowing the use of efficient methods from convex optimization to approximately solve the subproblem

$$(2.12) \quad \min_{y \in \mathbb{R}^n} \tilde{\Omega}_{x, M}(y).$$

In this work, we are interested in the computation of an inexact solution  $x^+$  of (2.12) such that

$$(2.13) \quad \|\nabla \Omega_{x, M}(x^+) + g_\psi(x^+)\| \leq \frac{M}{6} \|x^+ - x\|^3$$

for some  $g_\psi(x^+) \in \partial\psi(x^+)$ . When  $M$  is sufficiently large, this type of point enjoys the following property.

**LEMMA 2.1.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2), and let  $x^+$  be a point for which (2.13) holds. If  $M \geq L_f$ , then*

$$(2.14) \quad \langle \nabla \tilde{f}(x^+), x - x^+ \rangle \geq \frac{1}{6M^{\frac{1}{3}}} \|\nabla \tilde{f}(x^+)\|^{\frac{4}{3}},$$

where  $\nabla \tilde{f}(x^+) = \nabla f(x^+) + g_\psi(x^+)$ .

*Proof.* By (2.5), (2.9), and (2.13) we have

$$\begin{aligned} & \left\| \nabla \tilde{f}(x^+) + \frac{M}{2} \|x^+ - x\|^2 (x^+ - x) \right\| \\ &= \|\nabla f(x^+) - \nabla \Phi_x(x^+) + \nabla \Omega_{x, M}(x^+) + g_\psi(x^+)\| \\ &\leq \|\nabla f(x^+) - \nabla \Phi_x(x^+)\| + \|\nabla \Omega_{x, M}(x^+) + g_\psi(x^+)\| \\ &\leq \frac{L_f + M}{6} \|x^+ - x\|^3. \end{aligned}$$

Since  $M \geq L_f$ , we get

$$\left\| \nabla \tilde{f}(x^+) + \frac{M}{2} \|x^+ - x\|^2 (x^+ - x) \right\| \leq \frac{M}{3} \|x^+ - x\|^3.$$

Consequently,

$$\begin{aligned} \frac{M^2}{9} \|x^+ - x\|^6 &\geq \left\| \nabla \tilde{f}(x^+) + \frac{M}{2} \|x^+ - x\|^2 (x^+ - x) \right\|^2 \\ &= \|\nabla \tilde{f}(x^+)\|^2 + M \|x^+ - x\|^2 \langle \nabla \tilde{f}(x^+), x^+ - x \rangle + \frac{M^2}{4} \|x^+ - x\|^6, \end{aligned}$$

which implies that

$$(2.15) \quad \langle \nabla \tilde{f}(x^+), x - x^+ \rangle \geq \frac{\|\nabla \tilde{f}(x^+)\|^2}{M\|x^+ - x\|^2} + M \left( \frac{1}{4} - \frac{1}{9} \right) \|x^+ - x\|^4 \geq \frac{5M}{36} \|x^+ - x\|^4.$$

On the other hand, by (2.5), (2.9), and (2.13) we also have

$$\begin{aligned} \|\nabla \tilde{f}(x^+)\| &\leq \|\nabla f(x^+) - \Phi_x(x^+)\| \\ &\quad + \|\nabla \Phi_x(x^+) - \nabla \Omega_{x,M}(x^+)\| + \|\nabla \Omega_{x,M}(x^+) + g_\psi(x^+)\| \\ &\leq \frac{L_f}{3!} \|x^+ - x\|^3 + \frac{M}{2} \|x^+ - x\|^3 + \frac{M}{6} \|x^+ - x\|^3 \\ &= \frac{(L_f + 3M + M)}{6} \|x^+ - x\|^3 \leq \frac{L_f + 4M}{6} \|x^+ - x\|^3. \end{aligned}$$

Since  $M \geq L_f$ , we get

$$\|\nabla \tilde{f}(x^+)\| \leq \frac{5M}{6} \|x^+ - x\|^3,$$

which implies that

$$(2.16) \quad \|x^+ - x\|^3 \geq \frac{6}{5M} \|\nabla \tilde{f}(x^+)\|.$$

Finally, combining (2.15) and (2.16) we obtain (2.14).  $\square$

Combining Lemma 2.1 and the convexity of  $\tilde{f}(\cdot)$ , it follows that

$$\tilde{f}(x) - \tilde{f}(x^+) \geq \langle \nabla \tilde{f}(x^+), x - x^+ \rangle \geq \frac{1}{6M^{\frac{1}{3}}} \|\nabla \tilde{f}(x^+)\|^{\frac{4}{3}}$$

whenever  $x^+$  satisfies (2.13) for  $M \geq L_f$ . This sufficient decrease in the objective function motivates the use of this type of inexact solution of (2.12) in the iterative minimization of  $\tilde{f}(\cdot)$ , as we will see in sections 3 and 4. In what follows we have a sufficient condition for (2.13).

**LEMMA 2.2.** *Suppose that  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2), and let  $x \in \mathbb{R}^n$  and  $M, \epsilon > 0$ . If a point  $x^+$  satisfies*

$$(2.17) \quad \|\nabla f(x^+) + g_\psi(x^+)\| \geq \epsilon$$

and

$$(2.18) \quad \|\nabla \Omega_{x,M}(x^+) + g_\psi(x^+)\| \leq \frac{1}{2} \min \left\{ 1, \frac{M}{L_f + 3M} \right\} \epsilon,$$

then  $x^+$  satisfies (2.13).

*Proof.* By (2.17), (2.5), and (2.18) we get

$$\begin{aligned} \epsilon &\leq \|\nabla f(x^+) + g_\psi(x^+)\| \\ &\leq \|\nabla f(x^+) - \nabla \Phi_x(x^+)\| + \|\nabla \Phi_x(x^+) - \nabla \Omega_{x,M}(x^+)\| \\ &\quad + \|\nabla \Omega_{x,M}(x^+) + g_\psi(x^+)\| \\ &\leq \frac{L_f}{6} \|x^+ - x\|^3 + \frac{M}{2} \|x^+ - x\|^3 + \frac{\epsilon}{2}. \end{aligned}$$

Thus,

$$\frac{\epsilon}{2} \leq \left( \frac{2L_f + 6M}{12} \right) \|x^+ - x\|^3 = \left( \frac{L_f + 3M}{6} \right) \|x^+ - x\|^3,$$

which gives

$$(2.19) \quad \frac{1}{2} \left( \frac{M}{L_f + 3M} \right) \epsilon \leq \frac{M}{6} \|x^+ - x\|^3.$$

Finally, combining (2.18) and (2.19) we conclude that (2.13) holds.  $\square$

*Remark 2.3.* If  $M \geq 4L_f$  and

$$(2.20) \quad \|\nabla \Omega_{x,M}(x^+) + g_\psi(x^+)\| \leq \frac{\epsilon}{7},$$

then

$$\|\nabla \Omega_{x,M}(x^+) + g_\psi(x^+)\| \leq \frac{1}{2} \min \left\{ 1, \frac{M}{L_f + 3M} \right\} \epsilon.$$

Thus, if

$$\|\nabla f(x^+) + g_\psi(x^+)\| > \epsilon,$$

then it follows from Lemma 2.2 that  $x^+$  satisfies (2.13).

In view of Lemma 2.2 and Remark 2.3, the computation of  $x^+$  satisfying (2.13) reduces to the computation of an approximate stationary point of the model  $\tilde{\Omega}_{x,M}(\cdot)$  in the sense of (2.20). As shown in [11], this can be done using a Bregman gradient method (BGM) [2, 17]. When  $M < \frac{3}{2}L_f$ , the BGM in [11] takes at most  $\mathcal{O}(\epsilon^{-4})$  iterations to find  $x^+$  satisfying (2.18). On the other hand, when  $M > \frac{3}{2}L_f$ , the BGM takes at most  $\mathcal{O}(|\log(\epsilon)|)$  iterations to achieve the same goal. The adjustment of  $M$  in existing adaptive third-order methods (such as the ones in [12, 13]) do not take into account the progress of the inner solver. This means that they accept models with  $M < \frac{3}{2}L_f$ , for which the computation of the inexact solutions may be computationally costly. In the next subsection we will see a modified BGM scheme that, within  $\mathcal{O}(|\log(\epsilon)|)$  iterations, either generates an approximate stationary point of  $\tilde{\Omega}_{x,M}(\cdot)$  or gives a certificate that  $M < 4L_f$ .

**2.2. Bregman gradient method.** Exploring the special structure of  $\Omega_{x,M}(\cdot)$ , the next lemma gives bounds on  $\nabla^2 \Omega_{x,M}(\cdot)$  when  $M \geq 4L_f$ .

**LEMMA 2.4.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2), and let  $M \geq 4L_f$ . Then  $\Omega_{x,M}(\cdot)$  satisfies the strong relative smoothness condition*

$$(2.21) \quad \frac{1}{2} \nabla^2 \rho_x(y) \preceq \nabla^2 \Omega_{x,M}(y) \preceq \frac{3}{2} \nabla^2 \rho_x(y) \quad \forall x, y \in \mathbb{R}^n$$

for

$$(2.22) \quad \rho_x(y) = \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{2} d_4(y - x).$$

*Proof.* From the proof of Lemma 4 in [21], for  $\tau^2 = M/L_f$  we have

$$\begin{aligned}\nabla^2 \Omega_{x,M}(y) &\succeq \left(1 - \frac{1}{\tau}\right) \nabla^2 f(x) + \frac{M - \tau L_f}{2} \nabla^2 d_4(y - x) \\ &= \left(1 - \frac{1}{\tau}\right) \left[ \nabla^2 f(x) + \frac{\tau M - \tau^2 L_f}{(\tau - 1)2} \nabla^2 d_4(y - x) \right] \\ &= \left(1 - \frac{1}{\tau}\right) \left[ \nabla^2 f(x) + \frac{\tau M - M}{(\tau - 1)2} \nabla^2 d_4(y - x) \right] \\ &= \left(1 - \frac{1}{\tau}\right) \nabla^2 \rho_x(y).\end{aligned}$$

Since  $M \geq 4L_f$ , it follows that  $\tau \geq 2$  and so

$$(2.23) \quad \nabla^2 \Omega_{x,M}(y) \succeq \frac{1}{2} \nabla^2 \rho_x(y).$$

On the other hand, from the proof of Lemma 4 in [21] we also have

$$\begin{aligned}\nabla^2 \Omega_{x,M}(y) &\preceq \left(\frac{\tau + 1}{\tau - 1}\right) \left[ \left(1 - \frac{1}{\tau}\right) \nabla^2 f(x) + \frac{\tau(\tau - 1)L_f}{2} \nabla^2 d_4(y - x) \right] \\ &= \frac{\tau + 1}{\tau} \nabla^2 f(x) + \frac{(\tau + 1)\tau}{2} L_f \nabla^2 d_4(y - x) \\ &= \left(\frac{\tau + 1}{\tau}\right) \left[ \nabla^2 f(x) + \frac{\tau^2 L_f}{2} \nabla^2 d_4(y - x) \right] \\ &= \left(1 + \frac{1}{\tau}\right) \nabla^2 \rho_x(y).\end{aligned}$$

Since  $\tau \geq 2$ , it follows that

$$(2.24) \quad \nabla^2 \Omega_{x,M}(y) \preceq \frac{3}{2} \nabla^2 \rho_x(y).$$

Thus, combining (2.23) and (2.24) we get (2.21).  $\square$

Lemma 2.4 is a refinement of Lemma 4 in [21]. By Proposition 1.1 in [17], it follows from (2.21) that, for any  $x, z \in \mathbb{R}^n$ ,

$$(2.25) \quad \Omega_{x,M}(y) \leq \Omega_{x,M}(z) + \langle \nabla \Omega_{x,M}(z), y - z \rangle + \frac{3}{2} \beta_{\rho_x}(z, y)$$

and

$$(2.26) \quad \Omega_{x,M}(y) \geq \Omega_{x,M}(z) + \langle \nabla \Omega_{x,M}(z), y - z \rangle + \frac{1}{2} \beta_{\rho_x}(z, y),$$

where

$$(2.27) \quad \beta_{\rho_x}(z, y) \equiv \rho_x(y) - \rho_x(z) - \langle \nabla \rho_x(z), y - z \rangle$$

is the Bregman divergence associated to the function  $\rho_x(\cdot)$ . In view of these inequalities we can approximately solve (2.12) using the following gradient method with Bregman distance (Algorithm B in [11]):

$$(2.28) \quad y_{k+1} = \arg \min_{y \in \mathbb{R}^n} \{ \langle \nabla \Omega_{x,M}(y_k), y - y_k \rangle + 3\beta_{\rho_x}(y_k, y) + \psi(y) \}, \quad k \geq 0.$$

Notice that (2.25) and (2.26) resemble the standard inequalities satisfied by strongly convex functions with Lipschitz continuous gradients. Because of this similarity and the properties of the Bregman divergence it is possible to establish a linear rate of convergence for the BGM defined in (2.28). Usually this type of result is proved for the sequence of functional residuals, i.e.,  $\{\tilde{\Omega}_{x,M}(y_k) - \tilde{\Omega}_{x,M}(y_M^*(x))\}_{k \geq 0}$ , where  $y_M^*$  is a solution of (2.12). However, we are interested in the convergence of the sequence  $\{\|\nabla \Omega_{x,M}(y_k) + g_\psi(y_k)\|\}_{k \geq 0}$ . The next three auxiliary results will allow us to relate these two sequences and then establish a linear rate of convergence for the latter (in Lemma 2.8). In particular, the first auxiliary result gives a sufficient condition for  $\Omega_{x,M}(\cdot)$  to be uniformly convex of degree 4.

**LEMMA 2.5.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2), and let  $M \geq 4L_f$ . Then,  $\Omega_{x,M}(\cdot)$  is uniformly convex of degree 4 with parameter  $M/24$ .*

*Proof.* By Theorem 6 in [19], function  $d_4(\cdot)$  is uniformly convex of degree 4 with parameter  $1/3$ , i.e.,

$$d_4(y) \geq d_4(x) + \langle \nabla d_4(x), y - x \rangle + \frac{1}{4} \left( \frac{1}{3} \right) \|y - x\|^4 \quad \forall x, y \in \mathbb{R}^n.$$

Consequently, function  $\rho_x(\cdot)$  defined in (2.22) is also uniformly convex of degree 4 with parameter  $M/6$ . Combining this fact with the assumption  $M \geq 4L_f$  and the first inequality in (2.21), it follows that, for any  $z, w \in \mathbb{R}^n$ ,

$$\begin{aligned} \frac{M}{12} \|z - w\|^4 &= \frac{2 \left( \frac{M}{6} \right)}{4} \|z - w\|^4 \\ &\leq \langle \nabla \rho_x(z) - \nabla \rho_x(w), z - w \rangle \\ &= \int_0^1 \langle \nabla^2 \rho_x(w + t(z - w))(z - w), z - w \rangle dt \\ &\leq 2 \left\langle \int_0^1 \nabla^2 \Omega_{x,M}(w + t(z - w))(z - w) dt, z - w \right\rangle \\ &= 2 \langle \nabla \Omega_{x,M}(z) - \nabla \Omega_{x,M}(w), z - w \rangle. \end{aligned}$$

Thus, by Lemma 1 in [22] we conclude that  $\Omega_{x,M}(\cdot)$  is uniformly convex of degree 4 with parameter  $M/24$ .  $\square$

Now, given  $x \in \text{dom } \tilde{f}$  and  $M > 0$ , consider the sublevel set

$$\mathcal{L}_M(x) = \left\{ y \in \mathbb{R}^n : \tilde{\Omega}_{x,M}(y) \leq \tilde{f}(x) \right\}.$$

The next lemma gives an upper bound for  $\|y - x\|$  when  $y \in \mathcal{L}_M(x)$ .

**LEMMA 2.6.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2), and let  $M \geq 4L_f$ . Then, given  $x \in \mathbb{R}^n$ ,*

$$(2.29) \quad \|y - x\|^3 \leq \frac{96 \|\nabla \tilde{f}(x)\|}{M} \quad \forall y \in \mathcal{L}_M(x)$$

where  $\nabla \tilde{f}(x)$  is an arbitrary subgradient of  $\tilde{f}(\cdot)$ .

*Proof.* Let  $y \in \mathcal{L}_M(x)$ . For any  $\nabla \tilde{f}(x) \in \partial \tilde{f}(x)$  we have

$$\nabla \tilde{f}(x) = \nabla f(x) + g_\psi(x),$$

where  $g_\psi(x) \in \partial \psi(x)$ . In particular,

$$(2.30) \quad \psi(y) \geq \psi(x) + \langle g_\psi(x), y - x \rangle.$$

Moreover, by Lemma 2.5,  $\Omega_{x,M}(\cdot)$  is uniformly convex of degree 4 with parameter  $M/24$ . Therefore,

$$(2.31) \quad \Omega_{x,M}(y) \geq \Omega_{x,M}(x) + \langle \nabla \Omega_{x,M}(x), y - x \rangle + \frac{M}{96} \|y - x\|^4.$$

Summing up inequalities (2.30) and (2.31) and using the equality  $\nabla \Omega_{x,M}(x) = \nabla f(x)$ , we get

$$(2.32) \quad \tilde{\Omega}_{x,M}(y) \geq \tilde{\Omega}_{x,M}(x) + \langle \nabla \tilde{f}(x), y - x \rangle + \frac{M}{96} \|y - x\|^4.$$

By assumption,  $\tilde{\Omega}_{x,M}(y) \leq \tilde{f}(x) = \tilde{\Omega}_{x,M}(x)$ . Thus, subtracting  $\tilde{\Omega}_{x,M}(x)$  in both sides of (2.32) it follows that

$$0 \geq \langle \nabla \tilde{f}(x), y - x \rangle + \frac{M}{96} \|y - x\|^4.$$

Then

$$\frac{M}{96} \|y - x\|^4 \leq \langle \nabla \tilde{f}(x), x - y \rangle \leq \|\nabla \tilde{f}(x)\| \|x - y\|,$$

which implies (2.29).  $\square$

From Lemma 2.6 we can obtain an upper bound for  $\|\nabla^2 \rho_x(y)\|$  for  $y \in \mathcal{L}_M(x)$ .

LEMMA 2.7. *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2), and let  $M \geq 4L_f$ . Then, given  $x \in \mathbb{R}^n$ , we have*

$$(2.33) \quad \|\nabla^2 \rho_x(y)\| \leq L_{x,M} \quad \forall y \in \text{conv}(\mathcal{L}_M(x)),$$

with

$$(2.34) \quad L_{x,M} \equiv \text{trace}(\nabla^2 f(x)) + \frac{3M}{2} \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{2}{3}},$$

where  $\nabla \tilde{f}(x)$  is an arbitrary subgradient of  $\tilde{f}(\cdot)$ .

*Proof.* Indeed, by (2.22) we have

$$\|\nabla^2 \rho_x(y)\| = \left\| \nabla^2 f(x) + \frac{M}{2} \nabla^2 d_4(y - x) \right\| \leq \|\nabla^2 f(x)\| + \frac{3M}{2} \|y - x\|^2.$$

Thus, by Lemma 2.6 and (2.34) we have

$$\|\nabla^2 \rho_x(y)\| \leq L_{x,M} \quad \forall y \in \mathcal{L}_M(x).$$

Finally, since  $\tilde{\Omega}_{x,M}(\cdot)$  is convex (by Lemma 2.4), it follows that  $\mathcal{L}_M(x) = \text{conv}(\mathcal{L}_M(x))$ . Combining this equality with the previous inequality we see that (2.33) is true.  $\square$

For  $M$  sufficiently large, the next lemma gives a linear rate of convergence for  $\{\|\nabla \Omega_{x,M}(y_k) + g_\psi(y_k)\|\}_{k \geq 0}$  with  $\{y_k\}_{k \geq 0}$  generated by the BGM (2.28). The proof combines the previous results with Lemmas A.8 and A.9 from [11].

LEMMA 2.8. Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2). Given  $x \in \mathbb{R}^n$  and  $M > 0$ , let  $\{y_k\}_{k \geq 0}$  be the sequence generated by method (2.28) with  $y_0 = x$ . If  $M \geq 4L_f$ , then, for all  $k \geq 0$ ,

$$(2.35) \quad g_\psi(y_{k+1}) \equiv -\nabla\Omega_{x,M}(y_k) + 3[\nabla\rho_x(y_k) - \nabla\rho_x(y_{k+1})] \in \partial\psi(y_{k+1})$$

and

$$(2.36) \quad \|\nabla\Omega_{x,M}(y_{k+1}) + g_\psi(y_{k+1})\|^4 \leq \frac{3^8 L_{x,M}^4 \beta_{x,M}}{2M \left(\frac{6}{5}\right)^k},$$

where  $L_{x,M}$  is given in (2.34) and

$$(2.37) \quad \beta_{x,M} \equiv \frac{1}{2} \text{trace}(\nabla^2 f(x)) \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{2}{3}} + \frac{M}{8} \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{4}{3}},$$

with  $\nabla \tilde{f}(x)$  being an arbitrary subgradient of  $\tilde{f}(\cdot)$ .

*Proof.* By Lemma A.8 in [11], (2.35) holds and

$$(2.38) \quad \tilde{\Omega}_{x,M}(y_k) - \tilde{\Omega}_{x,M}(y_{k+1}) \geq \frac{\left(\frac{M}{6}\right)}{4 \left(\frac{3}{2}\right)^3 (3L_{x,M})^4} \|\nabla\Omega_{x,M}(y_{k+1}) + g_\psi(y_{k+1})\|^4.$$

On the other hand, by Lemma A.9 in [11] we also have

$$(2.39) \quad \tilde{\Omega}_{x,M}(y_k) - \tilde{\Omega}_{x,M}(y_M^*(x)) \leq \frac{\left(\frac{1}{2}\right) \beta_{\rho_x}(x, y_M^*(x))}{\left(\frac{3}{3-\frac{1}{2}}\right)^k},$$

where  $y_M^*(x) = \arg \min_{y \in \mathbb{R}^n} \tilde{\Omega}_{x,M}(y)$ . Then, combining (2.38) and (2.39) we obtain

$$(2.40) \quad \|\nabla\Omega_{x,M}(y_{k+1}) + g_\psi(y_{k+1})\|^4 \leq \frac{3^8 L_{x,M}^4 \beta_{\rho_x}(x, y_M^*(x))}{2M \left(\frac{6}{5}\right)^k}.$$

Notice that  $y_M^*(x) \in \mathcal{L}_M(x)$ . Thus, it follows from (2.27), (2.22), and Lemma 2.6 that

$$\begin{aligned} \beta_{\rho_x}(x, y_M^*(x)) &= \rho_x(y_M^*(x)) - \rho_x(x) - \langle \nabla \rho_x(x), y_M^*(x) - x \rangle = \rho_x(y_M^*(x)) \\ &= \frac{1}{2} \langle \nabla^2 f(x)(y_M^*(x) - x), y_M^*(x) - x \rangle + \frac{M}{8} \|y_M^*(x) - x\|^4 \\ &\leq \frac{1}{2} \|\nabla^2 f(x)\| \|y_M^*(x) - x\|^2 + \frac{M}{8} \|y_M^*(x) - x\|^4 \\ &\leq \frac{1}{2} \|\nabla^2 f(x)\| \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{2}{3}} + \frac{M}{8} \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{4}{3}} \\ (2.41) \quad &\leq \beta_{x,M}. \end{aligned}$$

Finally, combining (2.40) and (2.41) we get that (2.36) holds.  $\square$

The main novelty of Lemma 2.8 with respect to similar results in [11] comes from the fact that the right-hand side of (2.36) is computable. This gives us an implementable way to check the eventual violation of linear convergence when  $M < 4L_f$ . This property is the key element in the modified BGM presented in the next subsection.

**2.3. Modified Bregman gradient method.** For any  $x \in \mathbb{R}^n$  and  $M > 0$  we can apply method (2.28) with  $y_0 = x$  to search for an approximate stationary point of  $\tilde{\Omega}_{x,M}(\cdot)$  in the sense of (2.13). By Lemma 2.8, if for some iteration index  $k$  inequality (2.36) is not satisfied, then  $M < 4L_f$ . This fact motivates the following algorithm.

Notice that the output of Algorithm 1 is a pair  $(x^+, \alpha)$ . When  $\alpha = 1$ , this means that Algorithm 1 stopped at Step 3. In this case, by Lemma 2.8 we have a certificate that  $M < 4L_f$ . On the other hand, the output  $\alpha = 0$  means that Algorithm 1 stopped at Step 2. In this case, the corresponding point  $x^+$  is either an  $\epsilon$ -stationary point of  $\tilde{f}(\cdot)$  or it satisfies conditions (2.13). The next theorem establishes a complexity bound for Algorithm 1.

**THEOREM 2.9.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2). Then, Algorithm 1 performs at most*

---

**Algorithm 1**

---

**Step 0.** Given  $x \in \mathbb{R}^n$ ,  $\nabla \tilde{f}(x) \in \partial \tilde{f}(x)$ , and  $M, \epsilon > 0$ , compute

$$L_{x,M} = \text{trace}(\nabla^2 f(x)) + \frac{3M}{2} \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{2}{3}}$$

and

$$\beta_{x,M} = \frac{1}{2} \text{trace}(\nabla^2 f(x)) \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{2}{3}} + \frac{M}{8} \left[ \frac{96 \|\nabla \tilde{f}(x)\|}{M} \right]^{\frac{4}{3}}.$$

Set  $y_0 = x$ ,  $\alpha := 0$  and  $k := 0$ .

**Step 1.** Compute

$$(2.42) \quad y_{k+1} = \arg \min_{y \in \mathbb{R}^n} \{ \langle \nabla \Omega_{x,M}(y_k), y - y_k \rangle + 3\beta_{\rho_x}(y_k, y) + \psi(y) \}$$

and

$$g_\psi(y_{k+1}) = -\nabla \Omega_{x,M}(y_k) + 3[\nabla \rho_x(y_k) - \nabla \rho_x(y_{k+1})].$$

**Step 2.** If

$$(2.43) \quad \|\nabla \Omega_{x,M}(y_{k+1}) + g_\psi(y_{k+1})\| \leq \max \left\{ \frac{\epsilon}{7}, \frac{M}{6} \|y_{k+1} - x\|^3 \right\},$$

set  $x^+ = y_{k+1}$  and stop.

**Step 3.** If

$$(2.44) \quad \|\nabla \Omega_{x,M}(y_{k+1}) + g_\psi(y_{k+1})\|^4 > \frac{3^8 L_{x,M}^4 \beta_{x,M}}{2M \left(\frac{6}{5}\right)^k},$$

set  $x^+ = y_{k+1}$ ,  $\alpha := 1$ , and stop.

**Step 4.** Set  $k := k + 1$ , and go to Step 1.

---

$$1 + \log_{1.2} \left( \frac{3^8 (7L_{x,M})^4 \beta_{x,M}}{2M} \epsilon^{-4} \right)$$

iterations. Moreover, if  $M \geq 4L_f$ , then Algorithm 1 returns  $\alpha = 0$  and  $x^+$  satisfying  $\|\nabla f(x^+) + g_\psi(x^+)\| \leq \epsilon$  or (2.13).

*Proof.* Let  $\{y_k\}_{k=0}^{\ell+1}$  be generated by Algorithm 1. Since  $y_{\ell+1}$  was generated, this means that Algorithm 1 has not stopped at the  $\ell$ th iteration. Consequently,

$$(2.45) \quad \|\nabla \Omega_{x,M}(y_\ell) + g_\psi(y_\ell)\| > \frac{\epsilon}{7},$$

$$(2.46) \quad \|\nabla \Omega_{x,M}(y_\ell) + g_\psi(y_\ell)\| > \frac{M}{6} \|y_\ell - x\|^3,$$

and

$$(2.47) \quad \|\nabla \Omega_{x,M}(y_\ell) + g_\psi(y_\ell)\|^4 \leq \frac{3^8 L_{x,M}^4 \beta_{x,M}}{2M \left(\frac{6}{5}\right)^{\ell-1}}.$$

Assume that

$$\ell \geq 1 + \log_{1.2} \left( \frac{3^8 (7L_{x,M})^4 \beta_{x,M}}{2M} \epsilon^{-4} \right).$$

In this case, by (2.47) we have

$$\|\nabla \Omega_{x,M}(y_\ell) + g_\psi(y_\ell)\|^4 \leq \left(\frac{\epsilon}{7}\right)^4$$

and so

$$(2.48) \quad \|\nabla \Omega_{x,M}(y_\ell) + g_\psi(y_\ell)\| \leq \frac{\epsilon}{7},$$

which contradicts (2.45). Thus, we must have

$$\ell < 1 + \log_{1.2} \left( \frac{3^8 (7L_{x,M})^4 \beta_{x,M}}{2M} \epsilon^{-4} \right).$$

Now, let  $(x^+, \alpha)$  be the output of Algorithm 1, and assume that  $M \geq 4L_f$ . In this case, by Lemma 2.8, Algorithm 1 must stop at Step 2, which gives a pair  $(x^+, \alpha)$  with  $\alpha = 0$  and  $x^+$  satisfying (2.13) or  $\|\nabla f(x^+) + g_\psi(x^+)\| \leq \epsilon$ .  $\square$

**3. Basic adaptive third-order method.** Now that we have a suitable solver for auxiliary problems of the form (2.12), we can describe our basic adaptive third-order method to solve (2.1).

In Algorithm 2,  $t$  is the iteration counter, while  $i_t + 1$  is the number of trial points computed at the  $t$ th iteration. Regarding the trial points, notice that if  $i$  is sufficiently large, we have

$$2^i M_t \geq 4L_f.$$

Thus, in view of Lemma 2.1 and Theorem 2.9, either  $\|\nabla \tilde{f}(x_{t,i})\| \leq \epsilon$  or (3.1) holds. The next lemma gives an upper bound for  $\{M_t\}_{t=0}^T$  and also a bound for the total number of executions of Algorithm 1 up to the  $T$ th iteration of Algorithm 2.

**Algorithm 2**

**Step 0.** Choose  $x_0 \in \mathbb{R}^n$ ,  $M_0, \epsilon > 0$ , and set  $t := 0$ .

**Step 1.** Set  $i := 0$ .

**Step 1.1.** Apply Algorithm 1 to approximately minimize  $\tilde{\Omega}_{x_t, 2^i M_t}(\cdot)$ .

Let  $(x_{t,i}^+, \alpha_{t,i})$  be its output, and let  $\nabla \tilde{f}(x_{t,i}^+) = \nabla f(x_{t,i}^+) + g_\psi(x_{t,i}^+)$ .

**Step 1.2.** If  $\alpha_{t,i} = 1$ , set  $i := i + 1$  and go back to Step 1.1.

**Step 1.3.** If  $\|\nabla \tilde{f}(x_{t,i}^+)\| \leq \epsilon$ , set  $x_{t+1} = x_{t,i}^+$  and stop.

**Step 1.4.** If

$$(3.1) \quad \tilde{f}(x_t) - \tilde{f}(x_{t,i}^+) \geq \frac{1}{6(2^i M_t)^{\frac{1}{3}}} \|\nabla \tilde{f}(x_{t,i}^+)\|^{\frac{4}{3}}$$

holds, set  $i_t = i$  and go to Step 2. Otherwise, set  $i := i + 1$ , and go to Step 1.1.

**Step 2.** Set  $x_{t+1} = x_{t,i_t}^+$ ,  $M_{t+1} = \max\{2^{i_t-1} M_t, M_0\}$ ,  $t := t + 1$ , and go to Step 1.

LEMMA 3.1. Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2). Assume that  $\{x_t\}_{t=0}^T$  is generated by Algorithm 2 with

$$(3.2) \quad \|\nabla \tilde{f}(x_{t,i}^+)\| > \epsilon, \quad i = 0, \dots, i_t \quad \text{and} \quad t = 0, \dots, T.$$

Then

$$(3.3) \quad M_0 \leq M_t \leq \max\{M_0, 4L_f\}$$

for  $t = 0, \dots, T + 1$ . Moreover, the total number  $E_T$  of executions of Algorithm 1 up to the  $T$ th iteration of Algorithm 2 is bounded as follows:

$$(3.4) \quad E_T \leq 2(T + 1) + \log_2(\max\{M_0, 4L_f\}) - \log_2(M_0).$$

*Proof.* First let us prove by induction that (3.3) is true for  $t = 0, \dots, T$ . Clearly, it is true for  $t = 0$ . Assume that (3.3) is true for some  $t$  with  $0 \leq t \leq T$ . If  $i_t = 0$ , then by the definition of  $M_{t+1}$  and the induction assumption we have

$$M_0 \leq M_{t+1} = \max\left\{M_0, \frac{1}{2}M_t\right\} \leq \max\{M_0, 4L_f\},$$

that is, (3.3) holds for  $t + 1$ . Now, let us consider the case  $i_t \geq 1$ . Suppose that

$$(3.5) \quad 2^{i_t} M_t > 2(4L_f).$$

Then,

$$(3.6) \quad 2^{i_t-1} M_t > 4L_f,$$

and, by Theorem 2.9 and (3.2), we would have

$$(3.7) \quad \|\nabla \Omega_{x_t, 2^{i_t-1} M_t}(x_{t,i_t-1}^+) + g_\psi(x_{t,i_t-1}^+)\| \leq \frac{2^{i_t-1} M_t}{6} \|x_{t,i_t-1}^+ - x_t\|^3.$$

In this case, using the convexity of  $\tilde{f}(\cdot)$  it follows from (3.7) and Lemma 2.1 that

$$\tilde{f}(x_t) - \tilde{f}(x_{t,i_t-1}^+) \geq \frac{1}{6(2^{i_t-1} M_t)^{\frac{1}{3}}} \|\nabla \tilde{f}(x_{t,i_t-1}^+)\|^{\frac{4}{3}}.$$

This means that (3.1) would have been satisfied for  $i < i_t$ , contradicting the definition of  $i_t$ . Thus, (3.5) cannot be true; i.e.,

$$2^{i_t} M_t \leq 2(4L_f),$$

and so

$$M_0 \leq M_{t+1} = \max \{M_0, 2^{i_t-1} M_t\} \leq \max \{M_0, 4L_f\},$$

that is, (3.3) is also true for  $t+1$ . Therefore, (3.3) is true for  $t=0, \dots, T$ .

Now, notice that at the  $t$ th iteration of Algorithm 2, Algorithm 1 is executed  $i_t + 1$  times. From  $M_{t+1} = \max \{M_0, 2^{i_t-1} M_t\}$  we have

$$(3.8) \quad i_t - 1 \leq \log_2(M_{t+1}) - \log_2(M_t).$$

Then, by (3.8) and (3.3) we have

$$\begin{aligned} E_T &= \sum_{t=0}^T (i_t + 1) \leq \sum_{t=0}^T 2 + \log_2(M_{t+1}) - \log_2(M_t) \\ &= 2T + \log_2(M_{T+1}) - \log_2(M_0) \\ &\leq 2(T+1) + \log_2(\max \{M_0, 4L_f\}) - \log_2(M_0). \end{aligned}$$

Thus, (3.4) is also true.  $\square$

In what follows we will assume that

$$(3.9) \quad \max_{x \in \mathcal{L}(x_0)} \|x - x^*\| \leq R_0 \in (0, +\infty),$$

where  $x^*$  is a solution of (2.1) and

$$\mathcal{L}(x_0) = \left\{ x \in \mathbb{R}^n : \tilde{f}(x) \leq \tilde{f}(x_0) \right\}.$$

Using (3.9) we can also bound the number of inner iterations at each execution of Algorithm 1 in Algorithm 2.

LEMMA 3.2. *Suppose that  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2) and that (3.9) holds. Assume that  $\{x_t\}_{t=0}^T$  is generated by Algorithm 2 with*

$$(3.10) \quad \|\nabla \tilde{f}(x_{t,i}^+)\| > \epsilon, \quad i=0, \dots, i_t \quad \text{and} \quad t=0, \dots, T.$$

*Then, for each  $t \in \{0, \dots, T\}$  and each  $i \in \{0, \dots, i_t\}$ , the computation of  $x_{t,i}^+$  requires at most*

$$\mathcal{O} \left( \log_{1.2} \left( \frac{3^8 (7L_0)^4 \beta_0}{2M_0} \epsilon^{-4} \right) \right)$$

*iterations of Algorithm 1, where*

$$(3.11) \quad L_0 = 4nL_f R_0^2 + 2\text{trace}(\nabla^2 f(x_0)) + 3 \max \{M_0, 4L_f\} \left( \frac{96}{M_0} \right)^{\frac{2}{3}} G_0^{\frac{2}{3}},$$

$$(3.12) \quad \beta_0 = (2nL_f R_0^2 + \text{trace}(\nabla^2 f(x_0))) \left( \frac{96}{M_0} \right)^{\frac{2}{3}} G_0^{\frac{2}{3}} + \frac{\max \{M_0, 4L_f\}}{4} \left( \frac{96}{M_0} \right)^{\frac{4}{3}} G_0^{\frac{4}{3}},$$

and

$$(3.13) \quad G_0 = \max \left\{ \|\nabla \tilde{f}(x_0)\|, 6^{\frac{3}{4}} \max \{2M_0, 8L_f\}^{\frac{1}{4}} (\tilde{f}(x_0) - \tilde{f}(x^*))^{\frac{3}{4}} \right\}.$$

*Proof.* By Theorem 2.9, the computation of  $x_{t,i}^+$  requires at most

$$\mathcal{O}\left(\log_{1.2}\left(\frac{3^8 (7L_{x_t, 2^i M_t})^4 \beta_{x_t, 2^i M_t}}{2(2^i M_t)} \epsilon^{-4}\right)\right),$$

where

$$(3.14) \quad L_{x_t, 2^i M_t} = \text{trace}(\nabla^2 f(x_t)) + \frac{3(2^i M_t)}{2} \left(\frac{96}{2^i M_t}\right)^{\frac{2}{3}} \|\nabla \tilde{f}(x_t)\|^{\frac{2}{3}}$$

and

$$(3.15) \quad \beta_{x_t, 2^i M_t} = \frac{1}{2} \text{trace}(\nabla^2 f(x_t)) \left(\frac{96}{2^i M_t}\right)^{\frac{2}{3}} \|\nabla \tilde{f}(x_t)\|^{\frac{2}{3}} + \frac{2^i M_t}{8} \left(\frac{96}{2^i M_t}\right)^{\frac{4}{3}} \|\nabla \tilde{f}(x_t)\|^{\frac{4}{3}}.$$

Regarding  $2^i M_t$ , it follows from Step 1 and Lemma 3.1 that

$$(3.16) \quad M_0 \leq 2^i M_t \leq 2^{i_t} M_t \leq 2M_{t+1} \leq 2 \max\{M_0, 4L_f\}.$$

Thus, the computation of  $x_{t,i}^+$  requires at most

$$\mathcal{O}\left(\log_{1.2}\left(\frac{3^8 (7L_{x_t, 2^i M_t})^4 \beta_{x_t, 2^i M_t}}{2M_0} \epsilon^{-4}\right)\right).$$

It remains to show that  $L_{x_t, 2^i M_t}$  and  $\beta_{x_t, 2^i M_t}$  are bounded from above by  $L_0$  and  $\beta_0$ , respectively. In view of (2.6) we have

$$\begin{aligned} \nabla^2 f(x_t) &= \nabla^2 f(x_t) - \nabla^2 \Phi_{x_0}(x_t) + \nabla^2 \Phi_{x_0}(x_t) \\ &\preceq \frac{L_f}{2} \|x_t - x_0\|^2 I + \nabla^2 f(x_0) + D^3 f(x_0)[x_t - x_0], \end{aligned}$$

and so

$$(3.17) \quad \text{trace}(\nabla^2 f(x_t)) \leq \frac{nL_f}{2} \|x_t - x_0\|^2 + \text{trace}(\nabla^2 f(x_0)) + \text{trace}(D^3 f(x_0)[x_t - x_0]).$$

On the other hand, by Lemma 3 in [21] we also have

$$D^3 f(x_0)[x_t - x_0] \preceq \nabla^2 f(x_0) + \frac{L_f}{2} \|x_t - x_0\|^2 I,$$

which gives

$$(3.18) \quad \text{trace}(D^3 f(x_0)[x_t - x_0]) \leq \text{trace}(\nabla^2 f(x_0)) + \frac{nL_f}{2} \|x_t - x_0\|^2.$$

Thus, combining (3.17) and (3.18) we obtain

$$\begin{aligned} \text{trace}(\nabla^2 f(x_t)) &\leq nL_f \|x_t - x_0\|^2 + 2\text{trace}(\nabla^2 f(x_0)) \\ &\leq nL_f (\|x_t - x^*\| + \|x^* - x_0\|)^2 + 2\text{trace}(\nabla^2 f(x_0)) \\ &\leq 2nL_f (\|x_t - x^*\|^2 + \|x_0 - x^*\|^2) + 2\text{trace}(\nabla^2 f(x_0)). \end{aligned}$$

Since  $x_t \in \mathcal{L}(x_0)$ , it follows from (3.9) that

$$(3.19) \quad \text{trace}(\nabla^2 f(x_t)) \leq 4nL_f R_0^2 + 2\text{trace}(\nabla^2 f(x_0)).$$

Now, suppose that  $t \geq 1$ . By (3.1) we have

$$\tilde{f}(x_0) - \tilde{f}(x^*) \geq \tilde{f}(x_{t-1}) - \tilde{f}(x_t) \geq \frac{1}{6(2M_{t+1})^{\frac{1}{3}}} \|\nabla \tilde{f}(x_t)\|^{\frac{4}{3}}.$$

Thus, by Lemma 3.1, we get

$$(3.20) \quad \|\nabla \tilde{f}(x_t)\|^{\frac{4}{3}} \leq 6 \max\{2M_0, 8L_f\}^{\frac{1}{3}} (\tilde{f}(x_0) - \tilde{f}(x^*)),$$

and so

$$(3.21) \quad \|\nabla \tilde{f}(x_t)\| \leq 6^{\frac{3}{4}} \max\{2M_0, 8L_f\}^{\frac{1}{4}} (\tilde{f}(x_0) - \tilde{f}(x^*))^{\frac{3}{4}} \leq G_0,$$

where  $G_0$  is given in (3.13). Finally, it follows from (3.14), (3.16), (3.19), and (3.21) that  $L_{x_t, 2^i M_t} \leq L_0$ , while (3.15), (3.16), (3.19), and (3.20) imply that  $\beta_{x_t, 2^i M_t} \leq \beta_0$ .  $\square$

*Remark 3.3.* From (3.11) and (3.12), it follows that  $L_0$  and  $\beta_0$  depend linearly on  $n$ , the problem dimension. As we can see in (3.17), this dependence is due to the use of  $\text{trace}(\nabla^2 f(x))$  in the definitions of  $L_{x,M}$  and  $\beta_{x,M}$  at Step 0 of Algorithm 1. However, looking at the proof of Lemma 2.7, it is clear that we could use  $\|\nabla^2 f(x)\|$  instead of  $\text{trace}(\nabla^2 f(x))$ , which would give us  $L_0$  and  $\beta_0$  independent of  $n$ . In Algorithm 1, we choose to use  $\text{trace}(\nabla^2 f(x))$  because it can be easily computed. Since  $L_0$  and  $\beta_0$  appear inside the logarithm in our iteration complexity bound for Algorithm 1, the dependence on  $n$  is almost negligible.

**THEOREM 3.4.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2) and that (3.9) holds. Given  $\epsilon > 0$ , assume that  $\{x_t\}_{t=0}^T$  is generated by Algorithm 2 with*

$$(3.22) \quad \min \left\{ \|\nabla \tilde{f}(x_{t,i}^+)\|, \tilde{f}(x_{t,i}^+) - \tilde{f}(x^*) \right\} \geq \epsilon, \quad i = 0, \dots, i_t \quad \text{and} \quad t = 0, \dots, T.$$

*Denote by  $m$  the first iteration number such that*

$$(3.23) \quad \tilde{f}(x_m) - \tilde{f}(x^*) \leq 6^3 (4 \max\{M_0, 4L_f\} R_0^4),$$

*and assume that  $T > m$ . Then,*

$$(3.24) \quad m \leq 1 + \frac{1}{\log_2(\frac{4}{3})} \log_2 \log_2 \left( \frac{\tilde{f}(x_0) - \tilde{f}(x^*)}{6^3 (\max\{2M_0, 8L_f\} R_0^4)} \right)$$

*and, for all  $k$ ,  $m \leq k \leq T$ , we have*

$$(3.25) \quad \tilde{f}(x_k) - \tilde{f}(x^*) \leq \frac{(42)^3 \max\{2M_0, 8L_f\} R_0^4}{(k-m)^3}.$$

*Consequently,*

$$(3.26) \quad T \leq m + 42 \max\{2M_0, 8L_f\}^{\frac{1}{3}} R_0^{\frac{4}{3}} \epsilon^{-\frac{1}{3}}.$$

*Proof.* By (3.22) and Steps 1 and 2 of Algorithm 2, we have

$$(3.27) \quad \tilde{f}(x_k) - \tilde{f}(x_{k+1}) \geq \frac{1}{6(2M_{k+1})^{\frac{1}{3}}} \|\nabla f(x_{k+1})\|^{\frac{4}{3}} \quad \text{for } k = 0, \dots, T-1.$$

Since  $x_{k+1} \in \mathcal{L}(x_0)$  for all  $k \in \{1, \dots, T-1\}$ , using the convexity of  $\tilde{f}(\cdot)$ , (3.9), and (3.3) in (3.27), we get

(3.28)

$$\begin{aligned}
\tilde{f}(x_k) - \tilde{f}(x_{k+1}) &\geq \frac{1}{6(2M_{k+1}R_0^4)^{\frac{1}{3}}} \left( \|\nabla \tilde{f}(x_{k+1})\| \|x_{k+1} - x^*\| \right)^{\frac{4}{3}} \\
&\leq \frac{1}{6(2M_{k+1}R_0^4)^{\frac{1}{3}}} \left( \tilde{f}(x_k) - \tilde{f}(x^*) \right)^{\frac{4}{3}} \\
&\geq \frac{1}{6(2\max\{M_0, 4L_f\}R_0^4)^{\frac{1}{3}}} \left( \tilde{f}(x_k) - \tilde{f}(x^*) \right)^{\frac{4}{3}}, \quad k = 0, \dots, T-1.
\end{aligned}$$

Denote

$$(3.29) \quad \delta_k = \frac{\tilde{f}(x_k) - \tilde{f}(x^*)}{6^3(2\max\{M_0, 4L_f\}R_0^4)}.$$

Then, it follows from (3.28) that

$$(3.30) \quad \delta_k - \delta_{k+1} \geq \delta_{k+1}^{\frac{4}{3}}, \quad k = 0, \dots, T-1.$$

Thus,  $\{\delta_k\}_{k=0}^{T-1}$  satisfies condition (1.1) of Lemma 1.1 in [15] with power equal to  $4/3$ . By definition,  $m$  is the first iteration number such that (3.11) holds. This means that  $m$  is the first iteration number for which  $\delta_m \leq 2$ . If  $m > 0$ , then  $\delta_0 > 2$ . Then, by inequality (1.2) of Lemma 1.1 in [11] we obtain

$$\ln(2) \leq \ln(\delta_{m-1}) \leq \left(\frac{3}{4}\right)^{m-1} \ln(\delta_0).$$

Rearranging the terms in this inequality and then taking the logarithm in both sides, we conclude that (3.24) holds. Using  $\delta_m \leq 2$ , it follows from the inequality (1.3) of Lemma 1.1 in [15] that

$$\delta_k \leq \left[ \frac{1 + \delta_m^{\frac{1}{3}}}{\frac{1}{3}(k-m)} \right]^3 \leq \left( \frac{7}{k-m} \right)^3, \quad m \leq k \leq T.$$

Now, replacing (3.29) in (11) we get

$$\frac{\tilde{f}(x_k) - \tilde{f}(x^*)}{6^3(2\max\{M_0, 4L_f\}R_0^4)} \leq \frac{7^3}{(k-m)^3}, \quad m \leq k \leq T,$$

which implies (3.25).

Finally, by (3.22) and (3.25) we have the inequality

$$\epsilon < \tilde{f}(x_k) - \tilde{f}(x^*) \leq \frac{(42)^3 \max\{2M_0, 8L_f\}R_0^4}{(k-m)^3}, \quad m \leq k \leq T.$$

from which we get the bound (3.26).  $\square$

Combining (3.4) and (3.26) we have the following worst-case complexity result.

**COROLLARY 3.5.** *Under the same assumptions of Theorem 3.4, let  $x_{T+1} = x_{T,i_T}^+$  be the first trial point generated by Algorithm 2 such that*

$$(3.31) \quad \min \left\{ \|\nabla \tilde{f}(x_{T+1})\|, \tilde{f}(x_{T+1}) - \tilde{f}(x^*) \right\} \leq \epsilon.$$

*Then, the total number of executions of Algorithm 1 is bounded as follows:*

$$(3.32) \quad E_T \leq 2 + 2m + 84 \max\{2M_0, 8L_f\}^{\frac{1}{3}} R_0^{\frac{4}{3}} \epsilon^{-\frac{1}{3}} + \log_2(\max\{M_0, 4L_f\}) - \log_2(M_0),$$

*where  $m$  is bounded as in (3.24).*

*Remark 3.6.* Let us refer to the iterations of Algorithm 1 by inner iterations. In view of Corollary 3.5 and Lemma 3.2, Algorithm 2 needs at most  $\mathcal{O}(|\log_{1.2}(\epsilon)|\epsilon^{-\frac{1}{3}})$  inner iterations to generate  $\bar{x}$  such that

$$\min \left\{ \|\nabla \tilde{f}(\bar{x})\|, \tilde{f}(\bar{x}) - \tilde{f}(x^*) \right\} \leq \epsilon.$$

In fact, with a direct adaptation of the proof of Theorem 3.3 in [12], it can be shown that an inner iteration complexity bound of the same order holds for the generation of  $\bar{x}$  such that  $\|\nabla \tilde{f}(\bar{x})\| \leq \epsilon$ .

**4. Accelerated adaptive third-order method.** The adaptive procedure used in Algorithm 2 can also be incorporated in accelerated third-order methods (see, e.g., [13]). The next algorithm is an adaptive accelerated method based on Algorithm 1.

---

### Algorithm 3

---

**Step 0.** Choose  $x_0 \in \text{dom } \psi$ ,  $M_0, \epsilon > 0$ , and define function  $\varphi_0(x) = \frac{1}{4}\|x - x_0\|^4$  for all  $x \in \mathbb{R}^n$ . 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

$$(4.1) \quad a_{t,i}^4 = \frac{4^2 (A_t + a_{t,i})^3}{18^3 (2^i M_t)}.$$

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

$$(4.2) \quad z_{t,i} = (1 - \gamma_{t,i})x_t + \gamma_{t,i}v_t.$$

**Step 1.3.** Apply Algorithm 1 to approximately minimize  $\tilde{\Omega}_{z_{t,i}, 2^i M_t}(\cdot)$ .

Let  $(x_{t,i}^+, \alpha_{t,i})$  be its output, and let  $\nabla \tilde{f}(x_{t,i}^+) = \nabla f(x_{t,i}^+) + g_\psi(x_{t,i}^+)$ .

**Step 1.4.** If  $\alpha_{t,i} = 1$ , set  $i := i + 1$ , and go back to Step 1.1.

**Step 1.5.** If  $\|\nabla \tilde{f}(x_{t,i}^+)\| \leq \epsilon$ , set  $x_{t+1} = x_{t,i}^+$ , and stop.

**Step 1.6.** If

$$(4.3) \quad \langle \nabla \tilde{f}(x_{t,i}^+), z_{t,i} - x_{t,i}^+ \rangle \geq \frac{1}{6(2^i M_t)^{\frac{1}{3}}} \|\nabla \tilde{f}(x_{t,i}^+)\|^{\frac{4}{3}}$$

holds, set  $i_t = i$ , and go to Step 2. Otherwise, set  $i := i + 1$ , and go to Step 1.1.

**Step 2.** Set  $x_{t+1} = x_{t,i_t}^+$ ,  $z_t = z_{t,i_t}$ ,  $a_t = a_{t,i_t}$  and  $\gamma_t = \gamma_{t,i_t}$ . Define  $A_{t+1} = A_t + a_t$  and  $M_{t+1} = \max\{M_0, 2^{i_t-1} M_t\}$ .

**Step 3.** Define

$$(4.4) \quad \varphi_{t+1}(x) = \varphi_t(x) + a_t [f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \psi(x)] \quad \forall x \in \mathbb{R}^n,$$

and compute

$$(4.5) \quad v_{t+1} = \arg \min_{x \in \text{dom } \psi} \varphi_{t+1}(x).$$

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

---

LEMMA 4.1. Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2). Assume that  $\{x_t\}_{t=0}^T$  is generated by Algorithm 3 with

$$\|\nabla \tilde{f}(x_{t,i}^+)\| > \epsilon, \quad i = 0, \dots, i_t \quad \text{and} \quad t = 0, \dots, T.$$

Then

$$(4.6) \quad M_0 \leq M_t \leq \max\{M_0, 4L_f\}$$

for  $t = 0, \dots, T + 1$ . Moreover, the total number  $E_T$  of executions of Algorithm 1 up to the  $T$ th iteration of Algorithm 3 is bounded as follows:

$$(4.7) \quad E_T \leq 2(T + 1) + \log_2(\max\{M_0, 4L_f\}) - \log_2(M_0).$$

*Proof.* By Step 1 of Algorithm 3 and Theorem 2.9, the statements follow as in the proof of Lemma 3.1.  $\square$

LEMMA 4.2. Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2), and assume that  $\{x_t\}_{t=0}^T$  is generated by Algorithm 2. Then,

$$(4.8) \quad A_t \tilde{f}(x_t) \leq \varphi_t^* \equiv \min_{x \in \text{dom } \psi} \varphi_t(x)$$

for all  $t \in \{0, \dots, T\}$ .

*Proof.* By the equality  $A_0 = 0$  and the definition of  $\psi_0(\cdot)$ , we have

$$A_0 \tilde{f}(x_0) = 0 = \min_{x \in \text{dom } \psi} \varphi_0(x),$$

that is, (4.8) is true for  $t = 0$ . Assume that (4.8) is true for some  $t \in \{0, \dots, T - 1\}$ . Since  $\varphi_t(\cdot)$  is uniformly convex of degree 4 with parameter  $1/4$ , by the induction assumption we have

$$(4.9) \quad \varphi_t(x) \geq \varphi_t^* + \frac{1}{16} \|x - v_t\|^4 \geq A_t \tilde{f}(x_t) + \frac{1}{16} \|x - v_t\|^4.$$

Then, combining (4.4) and (4.9) we get

$$(4.10) \quad \begin{aligned} \varphi_{t+1}^* &= \min_{x \in \text{dom } \psi} \{ \varphi_t(x) + a_t [f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \psi(x)] \} \\ &\geq \min_{x \in \text{dom } \psi} \left\{ A_t \tilde{f}(x_t) + \frac{1}{16} \|x - v_t\|^4 + a_t \left[ f(x_{t+1}) \right. \right. \\ &\quad \left. \left. + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \psi(x) \right] \right\}. \end{aligned}$$

Using the convexity of  $\tilde{f}(\cdot)$ , we have

$$(4.11) \quad A_t \tilde{f}(x_t) \geq A_t \tilde{f}(x_{t+1}) + \langle \nabla \tilde{f}(x_{t+1}), A_t x_t - A_{t+1} x_{t+1} \rangle.$$

Moreover, since  $g_\psi(x_{t+1}) \in \partial \psi(x_{t+1})$  we also have

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

which gives

$$(4.12) \quad a_t [f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle + \psi(x)] \geq a_t \left[ \tilde{f}(x_{t+1}) + \langle \nabla \tilde{f}(x_{t+1}), x - x_{t+1} \rangle \right].$$

Thus, using (4.11), (4.12), and  $A_{t+1} = A_t + a_t$  in (4.10), it follows that

$$(4.13) \quad \varphi_{t+1}^* \geq \min_{x \in \text{dom } \psi} \left\{ A_{t+1} \tilde{f}(x_{t+1}) + \langle \nabla \tilde{f}(x_{t+1}), A_t x_t - A_t x_{t+1} + a_t x - a_t x_{t+1} \rangle + \frac{1}{16} \|x - v_t\|^4 \right\}.$$

By (4.2) and the definitions of  $\gamma_t$  and  $A_{t+1}$ , we have

$$z_t = (1 - \gamma_t)x_t + \gamma_t v_t = \left(1 - \frac{a_t}{A_t + a_t}\right)x_t + \frac{a_t}{A_t + a_t}v_t = \frac{A_t}{A_{t+1}}x_t + \frac{a_t}{A_{t+1}}v_t.$$

Consequently  $A_{t+1}z_t = A_t x_t + a_t v_t$ , which gives  $A_t x_t = A_{t+1}z_t - a_t v_t$ . Thus,

$$(4.14) \quad \begin{aligned} A_t x_t - A_t x_{t+1} + a_t x - a_t x_{t+1} &= A_t x_t - A_{t+1}x_{t+1} + a_t x \\ &= A_{t+1}z_t - a_t v_t - A_{t+1}x_{t+1} + a_t x \\ &= A_{t+1}(z_t - x_{t+1}) + a_t(x - v_t). \end{aligned}$$

Now, combining (4.13), (4.14) and (4.3), it follows that

$$(4.15) \quad \begin{aligned} \varphi_{t+1}^* &\geq A_{t+1}\tilde{f}(x_{t+1}) \\ &\quad + \min_{x \in \text{dom } \psi} \left\{ A_{t+1} \langle \nabla \tilde{f}(x_{t+1}), z_t - x_{t+1} \rangle + a_t \langle \nabla \tilde{f}(x_{t+1}), x - v_t \rangle + \frac{1}{16} \|x - v_t\|^4 \right\} \\ &\geq A_{t+1}\tilde{f}(x_{t+1}) \\ &\quad + \min_{x \in \text{dom } \psi} \left\{ \frac{A_{t+1}}{6(2^{i_t} M_t)^{\frac{1}{3}}} \|\nabla \tilde{f}(x_{t+1})\|^{\frac{4}{3}} + a_t \langle \nabla \tilde{f}(x_{t+1}), x - v_t \rangle + \frac{1}{16} \|x - v_t\|^4 \right\}. \end{aligned}$$

By Lemma 2 in [22] with  $p = 4$ ,  $s = a_t \nabla \tilde{f}(x_{t+1})$ , and  $\sigma = 1/4$ , we have

$$\begin{aligned} a_t \langle \nabla \tilde{f}(x_{t+1}), x - v_t \rangle + \frac{1}{16} \|x - v_t\|^4 &\geq - \left( \frac{3}{4} \right) 4^{\frac{1}{3}} a_t^{\frac{4}{3}} \|\nabla \tilde{f}(x_{t+1})\|^{\frac{4}{3}} \\ &= - \left( \frac{3}{4^{\frac{2}{3}}} \right) a_t^{\frac{4}{3}} \|\nabla \tilde{f}(x_{t+1})\|^{\frac{4}{3}}. \end{aligned}$$

Thus,

$$(4.16) \quad \begin{aligned} &\min_{x \in \text{dom } \psi} \left\{ \frac{A_{t+1}}{6(2^{i_t} M_t)^{\frac{1}{3}}} \|\nabla \tilde{f}(x_{t+1})\|^{\frac{4}{3}} + a_t \langle \nabla \tilde{f}(x_{t+1}), x - v_t \rangle + \frac{1}{16} \|x - v_t\|^4 \right\} \\ &\geq \left[ \frac{A_{t+1}}{6(2^{i_t} M_t)^{\frac{1}{3}}} - 4^{-\frac{2}{3}} 3a_t^{\frac{4}{3}} \right] \|\nabla \tilde{f}(x_{t+1})\|^{\frac{4}{3}} \\ &\geq \left[ \frac{(A_t + a_t)}{6(2^{i_t} M_t)^{\frac{1}{3}}} - 4^{-\frac{2}{3}} 3a_t^{\frac{4}{3}} \right] \|\nabla \tilde{f}(x_{t+1})\|^{\frac{4}{3}} \\ &= 0, \end{aligned}$$

where the last equality follows from (4.1). Finally, combining (4.15) and (4.16), we conclude that

$$\varphi_{t+1}^* \geq A_{t+1}\tilde{f}(x_{t+1}),$$

that is, (4.8) is true for  $t + 1$ .  $\square$

LEMMA 4.3. Let  $\{x_t\}_{t=0}^T$  be generated by Algorithm 3. Then, for all  $t \in \{0, \dots, T\}$ ,

$$\varphi_t(x) \leq A_t \tilde{f}(x) + \frac{1}{4} \|x - x_0\|^4 \quad \forall x \in \mathbb{R}^n.$$

*Proof.* The proof follows by induction as in the proof of Lemma 3.2 in [14].  $\square$

Now, combining Lemmas 4.1, 4.2, and 4.3, we can establish the rate of convergence of Algorithm 3.

THEOREM 4.4. Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2). Assume that  $\{x_t\}_{t=0}^T$ , with  $T \geq 2$ , is generated by Algorithm 3 with

$$(4.17) \quad \min \left\{ \|\nabla \tilde{f}(x_{t,i}^+)\|, \tilde{f}(x_t) - \tilde{f}(x^*) \right\} > \epsilon, \quad i = 0, \dots, i_t \quad \text{and} \quad t = 0, \dots, T.$$

Then

$$(4.18) \quad \tilde{f}(x_t) - \tilde{f}(x^*) \leq \frac{18^3 \max\{2M_0, 8L_f\} \|x_0 - x^*\|^4}{2^{\frac{13}{4}} (t-1)^4} \quad \forall t \in \{2, \dots, T\}.$$

Consequently,

$$(4.19) \quad T \leq 1 + \left(\frac{1}{2}\right)^{\frac{13}{16}} \left[ 18^{\frac{3}{4}} \max\{2M_0, 8L_f\}^{\frac{1}{4}} \|x_0 - x^*\| \right] \epsilon^{-\frac{1}{4}}.$$

*Proof.* By Lemmas 4.2 and 4.3, for  $t \in \{0, \dots, T\}$  we have

$$A_t \tilde{f}(x_t) \leq \varphi_t^* \leq A_t \tilde{f}(x^*) + \frac{1}{4} \|x^* - x_0\|^4.$$

Consequently, for  $t \in \{1, \dots, T\}$ ,

$$(4.20) \quad \tilde{f}(x_t) - \tilde{f}(x^*) \leq \frac{\|x^* - x_0\|^4}{4A_t}.$$

By (4.1) and (4.6) we also have

$$a_t^4 = \frac{4^2 (A_t + a_t)^3}{18^3 (2^{i_t} M_t)} = \frac{4^2 A_{t+1}^3}{18^3 (2M_{t+1})} \geq \frac{4^2}{18^3 \max\{2M_0, 8L_f\}} A_{t+1}^3.$$

Since  $a_t = A_{t+1} - A_t$ , it follows that

$$(4.21) \quad A_{t+1} - A_t \geq \left( \frac{4^2}{18^3 \max\{2M_0, 8L_f\}} \right)^{\frac{1}{4}} A_{t+1}^{\frac{3}{4}}.$$

Let us denote

$$(4.22) \quad B_t = \frac{18^3 \max\{2M_0, 8L_f\}}{4^2} A_t.$$

Then, by (4.21) we have

$$(4.23) \quad B_{t+1} - B_t \geq B_{t+1}^{\frac{3}{4}}.$$

Since  $B_0 = 0$ , it follows from (4.23) that  $B_1 \geq 1$ . Thus, by Lemma A.4 in [14] (with  $\alpha = 3/4$ ), we obtain

$$B_t \geq \left[ (1 - \alpha) \left( \frac{B_1^{\frac{1}{1-\alpha}}}{B_1^{\frac{1}{1-\alpha}} + 1} \right)^\alpha \right]^{\frac{1}{1-\alpha}} (t-1)^{\frac{1}{1-\alpha}} = \left[ \left( \frac{1}{4} \right) \left( \frac{1}{2} \right)^{\frac{3}{4}} \right] (t-1)^4 \quad \forall t \in \{2, \dots, T\}.$$

Now, in view of (4.22), it follows that

$$(4.24) \quad A_t \geq \frac{4^2}{18^3 \max\{2M_0, 8L_f\} 2^{\frac{11}{4}}} (t-1)^4 = \frac{2^{\frac{5}{4}}}{18^3 \max\{2M_0, 8L_f\}} (t-1)^4 \quad \forall t \in \{2, \dots, T\}.$$

Then, combining (4.20) and (4.24), we get (4.18). Finally, it follows from (4.17) and (4.18) that

$$\epsilon < \tilde{f}(x_T) - \tilde{f}(x^*) \leq \frac{18^3 \max\{2M_0, 8L_f\} \|x_0 - x^*\|^4}{2^{\frac{13}{4}} (T-1)^4},$$

which implies (4.19).  $\square$

Combining (4.7) and (4.19), we have the following worst-case complexity result.

**COROLLARY 4.5.** *Under the same assumptions of Theorem 4.4, let  $x_{T+1} = x_{T,i_T}^+$  be the first trial point generated by Algorithm 3 such that*

$$(4.25) \quad \min \left\{ \|\nabla \tilde{f}(x_{T+1})\|, \tilde{f}(x_{T+1}) - \tilde{f}(x^*) \right\} \leq \epsilon.$$

*Then, the total number of executions of Algorithm 1 is bounded as follows:*

$$(4.26) \quad E_T \leq 4 + 2 \left( \frac{1}{2} \right)^{\frac{13}{16}} \left[ 18^{\frac{3}{4}} \max\{2M_0, 8L_f\}^{\frac{1}{4}} \|x_0 - x^*\| \right] \epsilon^{-\frac{1}{4}} \\ + \log_2 (\max\{M_0, 4L_f\}) - \log_2(M_0).$$

In what follows we will assume that for all  $z \in \text{dom } \psi$ ,

$$(4.27) \quad \|g_\psi(z)\| \leq L_\psi \quad \forall g_\psi(z) \in \partial\psi(z).$$

The following lemma gives an upper bound for the number of iterations required at each execution of Algorithm 1 as inner solver in Algorithm 3.

**LEMMA 4.6.** *Suppose that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies (2.2) and that (3.9) and (4.27) hold. Assume that  $\{x_t\}_{t=0}^T$  is generated by Algorithm 3 with*

$$(4.28) \quad \|\nabla \tilde{f}(x_{t,i}^+)\| > \epsilon, \quad i = 0, \dots, i_t \quad \text{and} \quad t = 0, \dots, T.$$

*Then, for each  $t \in \{0, \dots, T\}$  and each  $i \in \{0, \dots, i_t\}$ , the computation of  $x_{t,i}^+$  requires at most*

$$\mathcal{O} \left( \log_{1.2} \left( \frac{3^8 (7\hat{L}_0)^4 \hat{\beta}_0}{2M_0} \epsilon^{-4} \right) \right)$$

*iterations of Algorithm 1, where*

$$(4.29) \quad \hat{L}_0 = 20nL_f R_0^2 + 2\text{trace}(\nabla^2 f(x_0)) + 3 \max\{M_0, 4L_f\} \left( \frac{96}{M_0} \right)^{\frac{2}{3}} \hat{G}_0^{\frac{2}{3}},$$

$$(4.30) \quad \hat{\beta}_0 = (10nL_f R_0^2 + \text{trace}(\nabla^2 f(x_0))) \left( \frac{96}{M_0} \right)^{\frac{2}{3}} \hat{G}_0^{\frac{2}{3}} + \frac{\max\{M_0, 4L_f\}}{4} \left( \frac{96}{M_0} \right)^{\frac{4}{3}} \hat{G}_0^{\frac{4}{3}},$$

*and*

$$(4.31) \quad \hat{G}_0 = \frac{4^3 L_f}{6} R_0^3 + \|\nabla f(x_0)\| + 4\|\nabla^2 f(x_0)\| R_0 + \frac{4^2}{2} R_0^2 + L_\psi,$$

*with  $R_0$  given in (3.9) and  $L_\psi$  given in (4.27).*

*Proof.* By Theorem 2.9, the computation of  $x_{t,i}^+$  requires at most

$$\mathcal{O}\left(\log_{1.2}\left(\frac{3^8 (7L_{z_{t,i}, 2^i M_t})^4 \beta_{z_{t,i}, 2^i M_t}}{2(2^i M_t)} \epsilon^{-4}\right)\right),$$

where

$$(4.32) \quad L_{z_{t,i}, 2^i M_t} = \text{trace}(\nabla^2 f(z_{t,i})) + \frac{3(2^i M_t)}{2} \left(\frac{96}{2^i M_t}\right)^{\frac{2}{3}} \|\nabla \tilde{f}(z_{t,i})\|^{\frac{2}{3}}$$

and

$$(4.33) \quad \beta_{z_{t,i}, 2^i M_t} = \frac{1}{2} \text{trace}(\nabla^2 f(z_{t,i})) \left(\frac{96}{2^i M_t}\right)^{\frac{2}{3}} \|\nabla \tilde{f}(z_{t,i})\|^{\frac{2}{3}} + \frac{2^i M_t}{8} \left(\frac{96}{2^i M_t}\right)^{\frac{4}{3}} \|\nabla \tilde{f}(z_{t,i})\|^{\frac{4}{3}}.$$

It follows from Step 1 of Algorithm 3 and Lemma 4.1 that

$$(4.34) \quad M_0 \leq 2^i M_t \leq 2^{it} M_t \leq 2M_{t+1} \leq 2 \max\{M_0, 4L_f\}.$$

Thus, the computation of  $x_{t,i}^+$  requires at most

$$\mathcal{O}\left(\log_{1.2}\left(\frac{3^8 (7L_{z_{t,i}, 2^i M_t})^4 \beta_{z_{t,i}, 2^i M_t}}{2M_0} \epsilon^{-4}\right)\right).$$

Now, it is enough to show that  $L_{z_{t,i}, 2^i M_t}$  and  $\beta_{z_{t,i}, 2^i M_t}$  are bounded from above by  $\hat{L}_0$  and  $\hat{\beta}_0$ , respectively. In view of (2.6) we have

$$\begin{aligned} \nabla^2 f(z_{t,i}) &= \nabla^2 f(z_{t,i}) - \nabla^2 \Phi_{x_0}(z_{t,i}) + \nabla^2 \Phi_{x_0}(z_{t,i}) \\ &\preceq \frac{L_f}{2} \|z_{t,i} - x_0\|^2 I + \nabla^2 f(x_0) + D^3 f(x_0)[z_{t,i} - x_0], \end{aligned}$$

and so

$$(4.35) \quad \text{trace}(\nabla^2 f(z_{t,i})) \leq \frac{nL_f}{2} \|z_{t,i} - x_0\|^2 + \text{trace}(\nabla^2 f(x_0)) + \text{trace}(D^3 f(x_0)[z_{t,i} - x_0]).$$

On the other hand, by Lemma 3 in [21] we also have

$$D^3 f(x_0)[z_{t,i} - x_0] \preceq \nabla^2 f(x_0) + \frac{L_f}{2} \|z_{t,i} - x_0\|^2 I,$$

which gives

$$(4.36) \quad \text{trace}(D^3 f(x_0)[z_{t,i} - x_0]) \leq \text{trace}(\nabla^2 f(x_0)) + \frac{nL_f}{2} \|z_{t,i} - x_0\|^2.$$

Thus, combining (4.35) and (4.36) we obtain

$$(4.37) \quad \begin{aligned} \text{trace}(\nabla^2 f(z_{t,i})) &\leq nL_f \|z_{t,i} - x_0\|^2 + 2\text{trace}(\nabla^2 f(x_0)) \\ &\leq nL_f (\|z_{t,i} - x^*\| + \|x^* - x_0\|)^2 + 2\text{trace}(\nabla^2 f(x_0)) \\ &\leq 2nL_f (\|z_{t,i} - x^*\|^2 + \|x_0 - x^*\|^2) + 2\text{trace}(\nabla^2 f(x_0)). \end{aligned}$$

By (4.1) we have

$$\begin{aligned}
 \|z_{t,i} - x^*\| &= \|(1 - \gamma_{t,i})x_t + \gamma_{t,i}v_t - (1 - \gamma_{t,i})x^* - \gamma_{t,i}x^*\| \\
 &\leq (1 - \gamma_{t,i})\|x_t - x^*\| + \gamma_{t,i}\|v_t - x^*\| \\
 (4.38) \quad &\leq \|x_t - x^*\| + \|v_t - x^*\|,
 \end{aligned}$$

where the last inequality is due to  $\gamma_{t,i} \in (0, 1]$ . Since  $\varphi_t(\cdot)$  is uniformly convex of degree 4 with parameter  $1/4$ , by Lemmas 4.2 and 4.3 we also have

$$\frac{1}{16}\|v_t - x\|^4 + A_t\tilde{f}(x_t) \leq \frac{1}{16}\|v_t - x\|^4 + \varphi_t^* \leq \varphi_t(x) \leq A_t\tilde{f}(x) + \frac{1}{4}\|x - x_0\|^4.$$

In particular, for  $x = x^*$  it follows that

$$\frac{1}{16}\|v_t - x^*\|^4 + A_t(\tilde{f}(x_0) - \tilde{f}(x^*)) \leq \frac{1}{4}\|x^* - x_0\|^4,$$

and so

$$(4.39) \quad \|v_t - x^*\| \leq 4^{\frac{1}{4}}\|x_0 - x^*\| \leq 2\|x_0 - x^*\|.$$

Moreover, by Lemmas 4.2 and 4.3 we also have

$$A_t\tilde{f}(x_t) \leq \varphi_t^* \leq \varphi_t(x_0) \leq A_0\tilde{f}(x_0).$$

Consequently,  $x_t \in \mathcal{L}(x_0)$ , and by (3.9) we get

$$(4.40) \quad \|x_t - x^*\| \leq R_0.$$

Thus, combining (4.38)–(4.40) and (3.9) it follows that

$$(4.41) \quad \|z_{t,i} - x^*\| \leq 3R_0.$$

Then, by (4.37), (4.41), and (3.9) we obtain

$$(4.42) \quad \text{trace}(\nabla^2 f(z_{t,i})) \leq 20nL_f R_0^2 + 2\text{trace}(\nabla^2 f(x_0)).$$

Moreover, by (2.5), (4.41), and (3.9) we have

$$\begin{aligned}
 \|\nabla f(z_{t,i})\| &= \|\nabla f(z_{t,i}) - \nabla \Phi_{x_0}(z_{t,i})\| + \|\nabla \Phi_{x_0}(z_{t,i})\| \\
 &\leq \frac{L_f}{6}\|z_{t,i} - x_0\|^3 + \|\nabla f(x_0)\| + \|\nabla^2 f(x_0)\|\|z_{t,i} - x_0\| \\
 &\quad + \frac{1}{2}\|D^3 f(x_0)\|\|z_{t,i} - x_0\|^2 \\
 &\leq \frac{4^3 L_f}{6}R_0^3 + \|\nabla f(x_0)\| + 4\|\nabla^2 f(x_0)\|R_0 + \frac{4^2}{2}\|D^3 f(x_0)\|R_0^2.
 \end{aligned}$$

Thus, by (4.27) and (4.31) we get

$$(4.43) \quad \|\nabla \tilde{f}(z_{t,i})\| \leq \hat{G}_0.$$

Finally, it follows from (4.32), (4.42), (4.43), and (4.29) that  $L_{z_{t,i}, 2^i M_t} \leq \hat{L}_0$ , while (4.33), (4.42), (4.43), and (4.30) imply that  $\beta_{z_{t,i}, 2^i M_t} \leq \hat{\beta}_0$ .  $\square$

*Remark 4.7.* In view of Corollary 4.5 and Lemma 4.6, to generate an  $\epsilon$ -approximate minimizer of  $\tilde{f}(\cdot)$ , Algorithm 3 needs at most  $\mathcal{O}(|\log_{1.2}(\epsilon)|\epsilon^{-\frac{1}{4}})$  iterations of Algorithm 1.

**5. Preliminary numerical results.** To investigate the feasibility of our new adaptive strategy, we tested MATLAB implementations of Algorithm 2 and its inner solver (Algorithm 1). We applied the corresponding code to smooth logistic regression problems of the form

$$\min_{x \in \mathbb{R}^{n+1}} f(x) = - \sum_{i=1}^m \left[ b^{(i)} \log(m_x(a^{(i)})) + (1 - b^{(i)}) \log(1 - m_x(a^{(i)})) \right],$$

where  $\{(a^{(i)}, b^{(i)})\}_{i=1}^m \subset \mathbb{R}^{n+1} \times \{0, 1\}$  is the dataset (with  $a_1^{(i)} = 1$  for  $i = 1, \dots, m$ )  $m_x(a) \equiv 1/(1 + e^{-\langle a, x \rangle})$ , and  $\log(\cdot)$  denotes the natural logarithm. Three datasets from [9] were considered. For each of them we used the starting point  $x_0 = [1 \ 1 \ \dots \ 1]^T$ ,  $M_0 = 1$ , and the stopping criterion  $\|\nabla f(x_k)\| \leq \epsilon$ . The results are presented in Tables 5.1, 5.2, and 5.3, where “IT” represents the number of outer iterations (i.e., iterations of Algorithm 2), “CO” represents the number of calls of the oracle,<sup>1</sup> “BGM-E” represents the number of executions of Algorithm 1, “BGM-IT” represents the total number of inner iterations (i.e., iterations of Algorithm 1), and “BGM-A” represents the average number of inner iterations per execution of Algorithm 1.

As we can see, the new adaptive scheme is able to find approximate stationary points with efficiency as predicted by the theory. Moreover, the average number of inner iterations required to compute a suitable trial point is usually small.

We also applied the adaptive Cubic Newton Method (CNM)<sup>2</sup> to the same problems. Table 5.4 shows the number of iterations and the number of calls of the oracle

TABLE 5.1  
Results for the dataset *Diabetes* ( $n = 8$  and  $m = 768$ ).

| $\epsilon$ | IT | CO  | BGM-E | BGM-IT | BGM-A |
|------------|----|-----|-------|--------|-------|
| $10^{-2}$  | 42 | 249 | 82    | 459    | 5.60  |
| $10^{-4}$  | 42 | 249 | 82    | 470    | 5.70  |
| $10^{-6}$  | 43 | 253 | 83    | 486    | 5.85  |
| $10^{-8}$  | 43 | 253 | 83    | 498    | 6.00  |

TABLE 5.2  
Results for the dataset *Phishing* ( $n = 9$  and  $m = 1353$ ).

| $\epsilon$ | IT | CO | BGM-E | BGM-IT | BGM-A |
|------------|----|----|-------|--------|-------|
| $10^{-2}$  | 4  | 17 | 4     | 51     | 12.75 |
| $10^{-4}$  | 4  | 17 | 4     | 62     | 15.50 |
| $10^{-6}$  | 5  | 17 | 4     | 74     | 18.50 |
| $10^{-8}$  | 5  | 21 | 5     | 93     | 18.60 |

TABLE 5.3  
Results for the dataset *Ionosphere* ( $n = 34$  and  $m = 351$ ).

| $\epsilon$ | IT   | CO   | BGM-E | BGM-IT | BGM-A |
|------------|------|------|-------|--------|-------|
| $10^{-2}$  | 59   | 242  | 60    | 258    | 4.30  |
| $10^{-4}$  | 125  | 506  | 126   | 522    | 4.14  |
| $10^{-6}$  | 411  | 1650 | 412   | 1666   | 4.04  |
| $10^{-8}$  | 1731 | 6930 | 1732  | 6946   | 4.01  |

<sup>1</sup>Each call of the oracle corresponds to either one function evaluation or one (high-order) derivative evaluation.

<sup>2</sup>Specifically, our implementation corresponds to the universal version of Algorithm 2 in [13] with  $p = 2$ ,  $\theta = 10^{-2}$ , and  $H_0 = 1$ .

TABLE 5.4  
Comparison between Algorithm 2 and CNM.

| Dataset    | Algorithm 2 |      | CNM  |       |
|------------|-------------|------|------|-------|
|            | IT          | CO   | IT   | CO    |
| Diabetes   | 43          | 253  | 58   | 289   |
| Phishing   | 5           | 21   | 10   | 43    |
| Ionosphere | 1731        | 6930 | 2198 | 10975 |

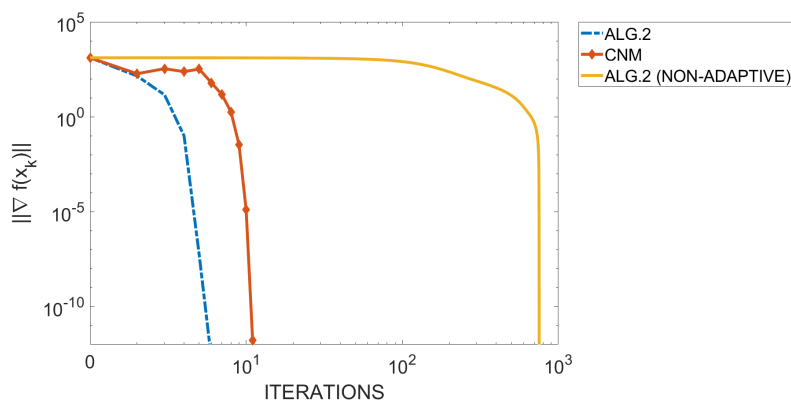


FIG. 5.1. Relative performance of Algorithm 2.

that Algorithm 2 and CNM executed to find  $\epsilon$ -approximate stationary points with  $\epsilon = 10^{-8}$ . Algorithm 2 outperformed CNM in these three problems, which reinforces our interest in third-order methods.

Finally, for the Phishing dataset, Figure 5.1 shows the norm of the gradient of the objective function at iterates generated by Algorithm 2, by a nonadaptive version of it (with  $M_t = 4L_f$  for all  $t$ ) and by CNM. As we can see, Algorithm 2 outperforms CNM (which is an adaptive second-order method). Remarkably, in this particular case, the two adaptive methods (Algorithm 2 and CNM) outperform the nonadaptive third-order method.<sup>3</sup> Moreover, this example suggests that Algorithm 2 enjoys a local superlinear rate of convergence.

**6. Discussion.** In our adaptive third-order methods, third-order information about  $f(\cdot)$  is only required for the computation of

$$(6.1) \quad \nabla \Omega_{x,M}(y) = \nabla f(x) + \nabla^2 f(x)(y-x) + \frac{1}{3} D^3 f(x)[y-x]^2 + \frac{M}{2} \|y-x\|^2 (y-x)$$

inside the inner solver (Algorithm 2). Notice that inequality (2.5) can be rewritten as

$$\left\| \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x) - \frac{1}{2} D^3 f(x)[y-x]^2 \right\| \leq \frac{L_f}{3!} \|y-x\|^3 \quad \forall x, y \in \mathbb{R}^n.$$

In particular, for any  $h \in \mathbb{R}^n$  and  $\tau > 0$ , we have

$$\left\| \nabla f(x + \tau h) - \nabla f(x) - \tau \nabla^2 f(x)h - \frac{\tau^2}{2} D^3 f(x)[h]^2 \right\| \leq \frac{L_f}{3!} \|\tau h\|^3$$

<sup>3</sup>In general, this superiority of adaptive methods depends on proper choices of the initial regularization parameters.

and

$$\left\| \nabla f(x - \tau h) - \nabla f(x) + \tau \nabla^2 f(x) h - \frac{\tau^2}{2} D^3 f(x)[h]^2 \right\| \leq \frac{L_f}{3!} \|h\|^3.$$

Combining these two inequalities, it follows that

$$\| \nabla f(x + \tau h) + \nabla f(x - \tau h) - 2\nabla f(x) - \tau^2 D^3 f(x)[h]^2 \| \leq \frac{L_f}{3} \tau^3 \|h\|^3,$$

and so we can approximate  $D^3 f(x)[h]^2$  by the vector

$$(6.2) \quad T_\tau(h) = \frac{\nabla f(x + \tau h) + \nabla f(x - \tau h) - 2\nabla f(x)}{\tau^2}$$

with an error bounded as follows:

$$\|T_\tau(h) - D^3 f(x)[h]^2\| \leq \frac{L_f}{3} \tau \|h\|^3.$$

Thus, by replacing  $D^3 f(x)[y - x]^2$  in (6.1) by  $T_\tau(y - x)$ , we obtain the vector

$$(6.3) \quad g_{x,M,\tau}(y) = \nabla f(x) + \nabla^2 f(x)(y - x) + \frac{1}{3} T_\tau(y - x) + \frac{M}{2} \|y - x\|^2 (y - x),$$

which approximates  $\nabla \Omega_{x,M}(y)$  with an error bound

$$\|g_{x,M,\tau}(y) - \nabla \Omega_{x,M}(y)\| \leq \frac{L_f}{9} \tau \|y - x\|^3.$$

Since  $g_{x,M,\tau}(\cdot)$  defined in (6.3) depends only on first- and second-order information about  $f(\cdot)$ , this remark opens the possibility for the development of second-order implementations of our third-order methods. As shown in [20], with a careful choice of  $\tau$ , it is indeed possible to obtain this type of second-order scheme preserving the global convergence rate of the original third-order method. However, the choice for  $\tau$  in [20] requires the knowledge of the Lipschitz constant  $L_f$ . Therefore, the development of adaptive strategies in this context remains a very interesting research direction.

**7. Conclusion.** In this paper we presented adaptive third-order methods for composite convex optimization problems in which the smooth part is a three-times continuously differentiable function with Lipschitz continuous third-order derivatives. Trial points are computed using a Bregman gradient method as inner solver. Our adaptive procedure to update the regularization parameters in the third-order tensor models takes into account the complexity of minimizing these models. When the regularization parameter is sufficiently large, the inner solver is guaranteed to find a suitable approximate stationary point of the model within  $\mathcal{O}(|\log(\epsilon)|)$  iterations. Thus, once a slow convergence rate is detected, the execution of the inner solver is stopped and the regularization parameter is increased. Using this strategy we obtained a basic adaptive third-order method that finds an  $\epsilon$ -approximate minimizer of the objective function performing at most  $\mathcal{O}(|\log(\epsilon)|\epsilon^{-\frac{1}{3}})$  iterations of the inner solver. We also obtained an accelerated adaptive third-order method with an improved bound of  $\mathcal{O}(|\log(\epsilon)|\epsilon^{-\frac{1}{4}})$  inner iterations. Preliminary numerical experiments illustrated the feasibility of our new adaptive technique. Topics for future research include the investigation about the local convergence rate of our basic adaptive third-order method, the development of adaptive high-order methods tailored to more structured problems classes [6], and also possible extensions in the context of nonconvex optimization [5] where adaptivity is a necessary feature of any practical method.

**Acknowledgments.** The authors are very grateful to Coralia Cartis and to the two anonymous referees, whose comments helped us to improve the paper.

## REFERENCES

- [1] M. BAES, *Estimate Sequence Methods: Extensions and Approximations*, Optimization Online, 2009.
- [2] H. H. BAUSCHKE, J. BOLTE, AND M. TEBoulLE, *A descent lemma beyond Lipschitz gradient continuity: First order methods revisited and applications*, Math. Oper. Res., 42 (2016), pp. 330–348.
- [3] E. G. BIRGIN, J. L. GARDENGHI, M. MARTÍNEZ, S. A. SANTOS, AND PH. L. TOINT, *Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*, Math. Program., 163 (2017), pp. 359–368.
- [4] C. CARTIS, N. I. GOULD, AND PH. L. TOINT, *Universal regularization methods: Varying the power, the smoothness, and the accuracy*, SIAM J. Optim., 29 (2019), pp. 595–615, <https://doi.org/10.1137/16M1106316>.
- [5] C. CARTIS, N. I. M. GOULD, AND PH. L. TOINT, *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation, and Perspectives*, SIAM, Philadelphia, 2022, <https://doi.org/10.1137/1.9781611976991>.
- [6] X. CHEN, B. JIANG, T. LIN, AND S. ZHANG, *Accelerating adaptive cubic regularization of Newton's method via random sampling*, J. Mach. Learn. Res., 23 (2022), pp. 1–38.
- [7] N. DOIKOV AND YU. NESTEROV, *Inexact tensor methods with dynamic accuracies*, in Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 2020, pp. 2577–2586.
- [8] J.-P. DUSSAULT, *ARCq: A new adaptive regularization by cubics variant*, Optim. Methods Softw., 33 (2018), pp. 322–335.
- [9] D. DUA AND C. GRAFF, *UCI Machine Learning Repository*, School of Information and Computer Science, University of California, Irvine, CA, 2019.
- [10] A. GASNIKOV, P. DVURECHENSKY, E. GORBUNOV, E. VORONTOVA, D. SELIKHANOVYCH, C. A. URIBE, B. JIANG, H. WANG, S. ZHANG, S. BUBECK, Q. JIANG, Y. T. LEE, Y. LI, AND A. SIDFORD, *Near optimal methods for minimizing convex functions with Lipschitz  $p$ -th Derivatives*, in Proceedings of the Thirty-Second Conference on Learning Theory, PMLR 99, 2019, pp. 1392–1393.
- [11] G. N. GRAPIGLIA AND YU. NESTEROV, *On inexact solution of auxiliary problems in tensor methods for convex optimization*, Optim. Methods Softw., 36 (2021), pp. 145–170.
- [12] G. N. GRAPIGLIA AND YU. NESTEROV, *Tensor methods for finding approximate stationary points of convex functions*, Optim. Methods Softw., 37 (2022), pp. 605–638, <https://doi.org/10.1080/10556788.2020.1818082>.
- [13] G. N. GRAPIGLIA AND YU. NESTEROV, *Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives*, SIAM J. Optim., 30 (2020), pp. 2750–2779, <https://doi.org/10.1137/19M1259432>.
- [14] G. N. GRAPIGLIA AND YU. NESTEROV, *Accelerated regularized Newton methods for minimizing composite convex functions*, SIAM J. Optim., 29 (2019), pp. 77–99, <https://doi.org/10.1137/17M1142077>.
- [15] 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>.
- [16] B. JIANG, T. LIN, AND S. ZHANG, *A unified adaptive tensor approximation scheme to accelerate composite convex optimization*, SIAM J. Optim., 30 (2020), pp. 2897–2926, <https://doi.org/10.1137/19M1286025>.
- [17] H. LU, R. M. FREUND, AND YU. NESTEROV, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. Optim., 28 (2018), pp. 333–354, <https://doi.org/10.1137/16M1099546>.
- [18] 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.
- [19] YU. NESTEROV, *Quartic Regularity*, preprint, <https://arxiv.org/abs/2201.04852>, [math.OC], 2022.
- [20] YU. NESTEROV, *Superfast second-order methods for unconstrained convex optimization*, J. Optim. Theory Appl., 191 (2021), pp. 1–30.
- [21] YU. NESTEROV, *Implementable tensor methods in unconstrained convex optimization*, Math. Program., 186 (2021), pp. 157–183.

- [22] YU. NESTEROV, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program., 112 (2008), pp. 159–181.
- [23] YU. NESTEROV AND B. T. POLYAK, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205.