

On inexact solution of auxiliary problems in tensor methods for convex optimization

G.N. Grapiglia & Yu. Nesterov

To cite this article: G.N. Grapiglia & Yu. Nesterov (2020): On inexact solution of auxiliary problems in tensor methods for convex optimization, Optimization Methods and Software, DOI: [10.1080/10556788.2020.1731749](https://doi.org/10.1080/10556788.2020.1731749)

To link to this article: <https://doi.org/10.1080/10556788.2020.1731749>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 23 Apr 2020.



Submit your article to this journal [↗](#)



Article views: 332





View related articles [↗](#)



View Crossmark data [↗](#)

On inexact solution of auxiliary problems in tensor methods for convex optimization

G.N. Grapiglia ^a and Yu. Nesterov ^b

^aDepartamento de Matemática, Universidade Federal do Paraná, Centro Politécnico, Curitiba, Brazil; ^bCenter for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium

ABSTRACT

In this paper, we study the auxiliary problems that appear in p -order tensor methods for unconstrained minimization of convex functions with ν -Hölder continuous p th derivatives. This type of auxiliary problems corresponds to the minimization of a $(p + \nu)$ -order regularization of the p th-order Taylor approximation of the objective. For the case $p = 3$, we consider the use of Gradient Methods with Bregman distance. When the regularization parameter is sufficiently large, we prove that the referred methods take at most $\mathcal{O}(\log(\epsilon^{-1}))$ iterations to find either a suitable approximate stationary point of the tensor model or an ϵ -approximate stationary point of the original objective function.

ARTICLE HISTORY

Received 31 August 2019
Accepted 27 January 2020

KEYWORDS

Unconstrained minimization; high-order methods; tensor methods; Hölder condition; worst-case global complexity bounds



2010 MATHEMATICS SUBJECT CLASSIFICATIONS

49M15; 49M37; 58C15; 90C25; 90C30

1. Introduction

1.1. Motivation

In [19], a cubic regularization of Newton's method (CNM) was proposed for convex and nonconvex minimization of functions with Lipschitz continuous Hessian. At each iteration of CNM, a trial point is computed by minimizing a third-order regularization of the second-order Taylor approximation of the objective function around the current iterate. When the objective f is convex, it was shown that CNM takes at most $\mathcal{O}(\epsilon^{-1/2})$ iterations to generate \bar{x} such that $f(\bar{x}) - f_* \leq \epsilon$, where f_* is the optimal value of f . An accelerated version of CNM was proposed in [16] with an improved complexity bound of $\mathcal{O}(\epsilon^{-1/3})$. In the sequel, accelerated p -order tensor methods with complexity of $\mathcal{O}(\epsilon^{-1/(p+1)})$ were proposed by Baes [1], generalizing the accelerated CNM. However, each iteration of these tensor methods require the exact minimization of a potentially nonconvex model, namely, a $(p + 1)$ -order regularization of the p th-order Taylor approximation of the objective. Since the global minimization of general nonconvex multivariate polynomials is computationally out of reach, the contribution in [1] remained restricted to the theoretical field.

CONTACT G.N. Grapiglia  grapiglia@ufpr.br, geovani_mat@outlook.com  Departamento de Matemática, Universidade Federal do Paraná, Centro Politécnico, Cx. postal 19.081, Curitiba, Paraná 81531-980, Brazil

Recently, two important works have pointed new ways towards practical tensor methods. In the context of nonconvex optimization, Birgin et al. [3] presented a p -order tensor method that can find \bar{x} with $\|\nabla f(\bar{x})\|_* \leq \epsilon$ in at most $\mathcal{O}(\epsilon^{-(p+1)/p})$ iterations, generalizing the bound of $\mathcal{O}(\epsilon^{-3/2})$ proved in [19] for the CNM (case $p = 2$). The method is based on the same regularized models used in [1] but allows the trial points to be approximate stationary points of the tensor models. On the other hand, in the context of convex optimization, Nesterov [17] proved that regularized tensor models are convex if the corresponding regularization parameter is sufficiently large. This makes possible the iterative solution of tensor auxiliary problems by efficient methods from Convex Optimization.

The tensor methods in [17] make explicit use of the Lipschitz constant of the higher-order derivative of the objective and also require the exact solution of the convex auxiliary problems. In [10,11], we proposed adaptive tensor methods for unconstrained minimization of convex functions with ν -Hölder continuous p th derivatives. These methods generalize the regularized Newton methods presented in [8,9] for $p = 2$ and allow inexact solution of the auxiliary problems as in [3].

In this paper, we investigate the use of Gradient Methods with Bregman distance to approximately solve the auxiliary problems in third-order tensor methods. When the regularization parameter is sufficiently large, we prove that these schemes applied to the corresponding tensor model take at most $\mathcal{O}(\log(\epsilon^{-1}))$ iterations to find either an approximate stationary point of the model (in the sense of [3]) or an ϵ -approximate stationary point of the original objective function.

1.2. Contents

The paper is organized as follows. In Section 2, we state the general problem. In Section 3, we establish convexity and smoothness properties of regularized third-order tensor models. In Section 4, we consider a Bregman Gradient Method for the approximate solution of smooth third-order tensor auxiliary problems. In Section 4, we consider possibly nonsmooth auxiliary problem that arise in composite convex optimization. General complexity results for Bregman Gradient Methods are provided in the Appendix.

1.3. Notations and generalities

In what follows, we denote by \mathbb{E} a finite-dimensional real vector space, and by \mathbb{E}^* its *dual* space, composed by linear functionals on \mathbb{E} . The value of function $s \in \mathbb{E}^*$ at point $x \in \mathbb{E}$ is denoted by $\langle s, x \rangle$. Given a self-adjoint positive definite operator $B : \mathbb{E} \rightarrow \mathbb{E}^*$ (notation $B \succ 0$), we can endow these spaces with conjugate Euclidean norms:

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

For a smooth function $f : \text{dom } f \rightarrow \mathbb{R}$ with convex and open domain $\text{dom } f \subset \mathbb{E}$, denote by $\nabla f(x)$ its gradient, and by $\nabla^2 f(x)$ its Hessian evaluated at point $x \in \text{dom } f$. Note that $\nabla f(x) \in \mathbb{E}^*$ and $\nabla^2 f(x)h \in \mathbb{E}^*$ for $x \in \text{dom } f$ and $h \in \mathbb{E}$.

For any integer $p \geq 1$, denote by

$$D^p f(x)[h_1, \dots, h_p]$$

the directional derivative of function f at x along directions $h_i \in \mathbb{E}$, $i = 1, \dots, p$. In particular, for any $x \in \text{dom } f$ and $h_1, h_2 \in \mathbb{E}$ we have

$$Df(x)[h_1] = \langle \nabla f(x), h_1 \rangle \quad \text{and} \quad D^2f(x)[h_1, h_2] = \langle \nabla^2 f(x)h_1, h_2 \rangle.$$

If $h_1 = \dots = h_p = h \in \mathbb{E}$, we denote $D^p f(x)[h_1, \dots, h_p]$ as $D^p f(x)[h]^p$. With this notation, the p th-order Taylor approximation of function f at $x \in \text{dom } f$ can be written as follows:

$$f(x+h) = \Phi_{x,p}(x+h) + o(\|h\|^p), \quad x+h \in \text{dom } f, \quad (1)$$

where

$$\Phi_{x,p}(y) \equiv f(x) + \sum_{i=1}^p \frac{1}{i!} D^i f(x)[y-x]^i, \quad y \in \mathbb{E}. \quad (2)$$

Since $D^p f(x)[\cdot]$ is a symmetric p -linear form, its norm is defined as:

$$\|D^p f(x)\| = \max_{h_1, \dots, h_p} \{ |D^p f(x)[h_1, \dots, h_p]| : \|h_i\| \leq 1, i = 1, \dots, p \}.$$

It can be shown that (see, e.g. Appendix 1 in [18])

$$\|D^p f(x)\| = \max_h \{ |D^p f(x)[h]^p| : \|h\| \leq 1 \}.$$

Similarly, since $D^p f(x)[\cdot, \dots, \cdot] - D^p f(y)[\cdot, \dots, \cdot]$ is also a symmetric p -linear form for fixed $x, y \in \text{dom } f$, it follows that

$$\|D^p f(x) - D^p f(y)\| = \max_h \{ |D^p f(x)[h]^p - D^p f(y)[h]^p| : \|h\| \leq 1 \}.$$

2. Problem statement

Let $f : \mathbb{E} \rightarrow \mathbb{R}$ be a p -times differentiable convex function with ν -Hölder continuous p th derivatives, that is,

$$\|D^p f(x) - D^p f(y)\| \leq H_{f,p}(\nu) \|x - y\|^\nu, \quad \forall x, y \in \mathbb{E}, \quad (3)$$

for some $\nu \in [0, 1]$. Given $x \in \mathbb{E}$, let us consider the following minimization problem:

$$\min_{y \in \mathbb{E}} \Omega_{x,p,H}^{(\nu)}(y) \equiv \Phi_{x,p}(y) + \frac{H}{p!} \|y - x\|^{p+\nu}, \quad (4)$$

where $\Phi_{x,p}(\cdot)$ is defined in (2) and $H > 0$. Problems of the form (4) appear as auxiliary problems in p -order tensor methods for convex and nonconvex unconstrained optimization (see, e.g. [3,5,10,11,15]). In these methods, only approximate stationary points of $\Omega_{x,p,H}^{(\nu)}(\cdot)$ are required [3]. Specifically, it is enough to find x^+ such that

$$\Omega_{x,p,H}^{(\nu)}(x^+) \leq f(x), \quad (5)$$

and

$$\|\nabla \Omega_{x,p,H}^{(\nu)}(x^+)\|_* \leq \theta \|x^+ - x\|^{p+\nu-1}, \quad (6)$$

where $\theta > 0$. The next lemma gives a sufficient condition for (6) to be satisfied.

Lemma 2.1: Let $x \in \mathbb{E}$, $H, \theta > 0$ and $\delta \in (0, 1)$. If

$$\|\nabla f(x^+)\|_* \geq \delta \quad \text{and} \quad \|\nabla \Omega_{x,p,H}^{(v)}(x^+)\|_* \leq \min \left\{ \frac{1}{2}, \frac{\theta(p-1)!}{2[H_{f,p}(v) + H(p+v)]} \right\} \delta, \quad (7)$$

then x^+ satisfies (6).

Proof: From (3), it follows that

$$\|\nabla f(y) - \nabla \Phi_{x,p}(y)\|_* \leq \frac{H_{f,p}(v)}{(p-1)!} \|y - x\|^{p+v-1}, \quad \forall y \in \mathbb{E}. \quad (8)$$

Combining (7) and (8) we obtain

$$\begin{aligned} \delta \leq \|\nabla f(x^+)\|_* &\leq \|\nabla f(x^+) - \nabla \Phi_{x,p}(x^+)\|_* + \|\nabla \Phi_{x,p}(x^+) - \nabla \Omega_{x,p,H}^{(v)}(x^+)\|_* \\ &\quad + \|\nabla \Omega_{x,p,H}^{(v)}(x^+)\|_* \\ &\leq \frac{H_{f,p}(v)}{(p-1)!} \|x^+ - x\|^{p+v-1} + \frac{H(p+v)}{p!} \|x^+ - x\|^{p+v-1} + \frac{\delta}{2} \\ &\leq \frac{H_{f,p}(v) + H(p+v)}{(p-1)!} \|x^+ - x\|^{p+v-1} + \frac{\delta}{2}. \end{aligned}$$

Thus,

$$\frac{\delta}{2} \leq \left(\frac{H_{f,p} + H(p+v)}{(p-1)!} \right) \|x^+ - x\|^{p+v-1},$$

which gives

$$\left[\frac{\theta(p-1)!}{2[H_{f,p}(v) + H(p+v)]} \right] \delta \leq \theta \|x^+ - x\|^{p+v-1}. \quad (9)$$

Finally, (6) follows directly from the second inequality in (7) and (9). ■

In view of Lemma 2.1, x^+ satisfying (5)–(6) can be computed by any monotone optimization scheme that drives the gradient of the objective to zero. It is worth mentioning that the lemma above does not require the convexity of f . Therefore, a slight modification of it also applies to the tensor models in [3,5,15]. Our goal in the next sections is to describe iterative schemes to solve (4) with $p = 3$, and also provide iteration-complexity bounds for reducing the norm of the gradient below the threshold specified in the second inequality in (7).

3. Gradient method for smooth third-order tensor models

3.1. Convexity and relative smoothness properties

The next lemma gives a sufficient condition for function $\Omega_{x,p,H}^{(v)}(\cdot)$ to be convex.

Lemma 3.1: Let $p \geq 2$. Then, for any $x, y \in \mathbb{E}$ we have

$$\nabla^2 f(y) \leq \nabla^2 \Phi_{x,p}(y) + \frac{H_{f,p}(v)}{(p-2)!} \|y-x\|^{p+\nu-2} B. \quad (10)$$

Moreover, if $H \geq (p-1)H_{f,p}(v)$, then function $\Omega_{x,p,H}^{(v)}(\cdot)$ is convex for any $x \in \mathbb{E}$.

Proof: See Lemma 5.1 in [10]. ■

In order to exploit additional properties of $\Omega_{x,p,H}^{(v)}(\cdot)$, let us focus on the case $p = 3$. Note that

$$\Phi_{x,3}(y) = f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y-x), (y-x) \rangle + \frac{1}{6} D^3 f(x)[y-x]^3, \quad (11)$$

and

$$\Omega_{x,3,H}^{(v)}(y) = \Phi_{x,3}(y) + \frac{H}{6} \|y-x\|^{3+\nu}. \quad (12)$$

The next auxiliary result gives bounds on the third-order derivatives of f . Its proof is an adaptation of the proof of Lemma 3 in [17].

Lemma 3.2: For any $x, y \in \mathbb{E}$ and $\tau > 0$ we have

$$\begin{aligned} -\frac{1}{\tau} \nabla^2 f(x) - \tau^\nu H_{f,3}(v) \|y-x\|^{1+\nu} B &\leq D^3 f(x)[y-x] \\ &\leq \frac{1}{\tau} \nabla^2 f(x) + \tau^\nu H_{f,3}(v) \|y-x\|^{1+\nu} B. \end{aligned} \quad (13)$$

Proof: Given $u, y \in \mathbb{E}$, by Lemma 3.1 (for $p = 3$) and the convexity of f , we have:

$$\begin{aligned} 0 &\leq \langle \nabla^2 f(y)u, u \rangle \leq \langle \nabla^2 \Phi_{x,3}(y)u, u \rangle + H_{f,3}(v) \|y-x\|^{1+\nu} \|u\|^2 \\ &= \langle (\nabla^2 f(x) + D^3 f(x)[y-x])u, u \rangle + H_{f,3}(v) \|y-x\|^{1+\nu} \|u\|^2. \end{aligned}$$

Thus, replacing y by $\bar{y} = x + \tau(y-x)$, we obtain

$$\begin{aligned} 0 &\leq \langle \nabla^2 f(\bar{y})u, u \rangle \\ &\leq \langle \nabla^2 f(x)u, u \rangle + \tau \langle D^3 f(x)[y-x]u, u \rangle + \tau^{1+\nu} H_{f,3}(v) \|y-x\|^{1+\nu} \|u\|^2 \\ &\implies -\tau \langle D^3 f(x)[y-x]u, u \rangle \leq \langle \nabla^2 f(x)u, u \rangle + \tau^{1+\nu} H_{f,3}(v) \|y-x\|^{1+\nu} \|u\|^2. \end{aligned}$$

Then, dividing this inequality by $-\tau$, it follows that

$$\langle D^3 f(x)[y-x]u, u \rangle \geq -\frac{1}{\tau} \langle \nabla^2 f(x)u, u \rangle - \tau^\nu H_{f,3}(v) \|y-x\|^{1+\nu} \|u\|^2. \quad (14)$$

Since u is arbitrary, this gives the first inequality in (13). The second inequality in (13) can be obtained by replacing $y-x$ by $-(y-x)$ in (14). ■

Now, using Lemma 3.2, we can prove relative smoothness properties¹ [14] of $\Omega_{x,3,H}^{(v)}(\cdot)$.

Theorem 3.3: Let $\tau_H = [(3 + v)H/6H_{f,3}(v)]^{1/(1+v)}$ and

$$\rho_x(y) \equiv \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{1}{3 + v} \|y - x\|^{3+v}. \quad (15)$$

Then, the following assertions hold:

(a) Function $\Omega_{x,3,H}^{(v)}(\cdot)$ is L_H -smooth relative to $\rho_x(\cdot)$ for

$$L_H = \max \left\{ \frac{\tau_H + 1}{\tau_H}, \tau_H^v (\tau_H + 1) H_{f,3}(v) \right\}. \quad (16)$$

(b) If $\tau_H \geq 1$, then function $\Omega_{x,3,H}^{(v)}(\cdot)$ is μ_H -strongly convex relative to $\rho_x(\cdot)$ for

$$\mu_H = \min \left\{ \frac{\tau_H - 1}{\tau_H}, \tau_H^v (\tau_H - 1) H_{f,3}(v) \right\}. \quad (17)$$

Proof: In view of (12) and (13), we have

$$\begin{aligned} \nabla^2 \Omega_{x,3,H}^{(v)}(y) &= \nabla^2 f(x) + D^3 f(x)[y - x] + \nabla^2 \left(\frac{H}{6} \|y - x\|^{3+v} \right) \\ &\leq \left(\frac{\tau_H + 1}{\tau_H} \right) \nabla^2 f(x) + \tau_H^v H_{f,3}(v) \|y - x\|^{1+v} B + \nabla^2 \left(\frac{H}{6} \|y - x\|^{3+v} \right) \\ &\leq \left(\frac{\tau_H + 1}{\tau_H} \right) \nabla^2 f(x) + \tau_H^v H_{f,3}(v) \nabla^2 \left(\frac{1}{3 + v} \|y - x\|^{3+v} \right) \\ &\quad + \frac{(3 + v)H}{6} \nabla^2 \left(\frac{1}{3 + v} \|y - x\|^{3+v} \right) \\ &= \left(\frac{\tau_H + 1}{\tau_H} \right) \nabla^2 f(x) + \tau_H^v (\tau_H + 1) H_{f,3}(v) \nabla^2 \left(\frac{1}{3 + v} \|y - x\|^{3+v} \right) \\ &\leq \max \left\{ \frac{\tau_H + 1}{\tau_H}, \tau_H^v (\tau_H + 1) H_{f,3}(v) \right\} \left[\nabla^2 f(x) + \nabla^2 \left(\frac{1}{3 + v} \|y - x\|^{3+v} \right) \right] \\ &= L_H \nabla^2 \rho_x(y). \end{aligned}$$

Since $\rho_x(\cdot)$ is convex, by Proposition 1.1 in [14] we conclude that $\Omega_{x,3,H}^{(v)}(\cdot)$ is L_H -smooth relative to $\rho_x(\cdot)$. This proves (a).

Now, suppose that $\tau_H \geq 1$. In this case, by (12) and (13) we have

$$\begin{aligned} \nabla^2 \Omega_{x,3,H}^{(v)}(y) &= \nabla^2 f(x) + D^3 f(x)[y - x] + \nabla^2 \left(\frac{H}{6} \|y - x\|^{3+v} \right) \\ &\geq \left(\frac{\tau_H - 1}{\tau_H} \right) \nabla^2 f(x) - \tau_H^v H_{f,3}(v) \|y - x\|^{1+v} B \end{aligned}$$

$$\begin{aligned}
& + \frac{(3+\nu)H}{6} \nabla^2 \left(\frac{1}{3+\nu} \|y-x\|^{3+\nu} \right) \\
\geq & \left(\frac{\tau_H-1}{\tau_H} \right) \nabla^2 f(x) - \tau_H^\nu H_{f,3}(\nu) \nabla^2 \left(\frac{1}{3+\nu} \|y-x\|^{3+\nu} \right) \\
& + \frac{(3+\nu)H}{6} \left(\frac{1}{3+\nu} \|y-x\|^{3+\nu} \right) \\
= & \left(\frac{\tau_H-1}{\tau_H} \right) \nabla^2 f(x) + \tau_H^\nu (\tau_H-1) H_{f,3}(\nu) \nabla^2 \left(\frac{1}{3+\nu} \|y-x\|^{3+\nu} \right) \\
\geq & \min \left\{ \frac{\tau_H-1}{\tau_H}, \tau_H^\nu (\tau_H-1) H_{f,3}(\nu) \right\} \left[\nabla^2 f(x) + \nabla^2 \left(\frac{1}{3+\nu} \|y-x\|^{3+\nu} \right) \right] \\
= & \mu_H \nabla^2 \rho_x(y).
\end{aligned}$$

Thus, by Proposition 1.1 in [14], we conclude that $\Omega_{x,3,H}^{(\nu)}(\cdot)$ is μ_H -strongly convex relative to $\rho_x(\cdot)$, and this proves (b). \blacksquare

Remark 3.1: Note that

$$\nabla^2 \left(\frac{1}{3+\nu} \|y-x\|^{3+\nu} \right) = (1+\nu) \|y-x\|^{\nu-1} B(y-x)(y-x)^T B + \|y-x\|^{1+\nu} B.$$

Consequently, for all $y \in \mathbb{E}$, we have

$$\|\nabla^2 \rho_x(y)\| \leq \|\nabla^2 f(x)\| + (2+\nu) \|y-x\|^{1+\nu}, \quad (18)$$

where $\|A\| = \max_{\|h\|=1} \|Ah\|$, for any matrix A . Moreover, by Lemma 5 in [7], it follows that $\rho_x(\cdot)$ is uniformly convex of degree $3+\nu$ with parameter $2^{-(1+\nu)}$.

The next lemma establishes an upper bound for the Hessians of function $\rho_x(\cdot)$ when $H \geq H_{f,p}(\nu)$.

Lemma 3.4: Given $x \in \mathbb{E}$ and $H \geq H_{f,3}(\nu)$, let

$$\mathcal{L}_H(x) = \left\{ z \in \mathbb{E} : \Omega_{x,3,H}^{(\nu)}(z) \leq f(x) \right\}.$$

Suppose that f has a global minimizer x^* and that

$$x \in \mathcal{F}(x_0) \equiv \left\{ z \in \mathbb{E} : f(z) \leq f(x_0) \right\},$$

with

$$\sup_{y \in \mathcal{F}(x_0)} \|y-x^*\| \leq R_0 < +\infty, \quad (19)$$

and $R_0 \geq 1$. Then,

$$\sup \left\{ \|\nabla^2 \rho_x(y)\| : y \in \text{co}(\mathcal{L}_H(x)) \right\} \leq \|\nabla^2 f(x)\| + 12R_0^2 \equiv N_x, \quad (20)$$

where $\text{co}(X)$ denotes the convex hull of the set X .

Proof: If $y \in \mathcal{L}_H(x)$, then

$$\Omega_{x,3,H}^{(v)}(y) \leq f(x) \leq f(x_0). \quad (21)$$

Since $H \geq H_{f,3}(v)$, it follows from (3) that

$$f(y) \leq \Omega_{x,3,H}^{(v)}(y). \quad (22)$$

Combining (21) and (22), we conclude that $y \in \mathcal{F}(x_0)$ and, by (19), we obtain

$$\|y - x\| \leq \|y - x^*\| + \|x^* - x\| \leq 2R_0. \quad (23)$$

Now, let $y \in \text{co}(\mathcal{L}_H(x))$. Then, there exists $\lambda \in [0, 1]$ and $y_1, y_2 \in \mathcal{L}_H(x)$ such that $y = (1 - \lambda)y_1 + \lambda y_2$. Consequently, using (23), we get

$$\|y - x\| \leq (1 - \lambda)\|y_1 - x\| + \lambda\|y_2 - x\| \leq 2R_0. \quad (24)$$

Finally, by (18) and (24), we conclude that (20) holds. ■

Even when $H < H_{f,p}(v)$ and $x \notin \mathcal{F}(x_0)$, we can bound the Hessians of $\rho_x(\cdot)$ on $\text{co}(\mathcal{L}_H(x))$. For that, we need first to establish the coercivity of $\Omega_{x,3,H}^{(v)}(\cdot)$ when $v \neq 0$.

Lemma 3.5: *Let $x \in \mathbb{E}$, $H > 0$ and $v \neq 0$. Then, the following statements are true:*

(a) *Given $A > 0$, if*

$$\|y - x\| > \max \left\{ \left[\frac{6(A - f(x))}{H} \right]^{1/3}, \left[\frac{6\|\nabla f(x)\|_*}{H} \right]^{1/2}, \frac{3\|\nabla^2 f(x)\|}{H}, \left[3 + \frac{\|D^3 f(x)\|}{H} \right]^{1/v} \right\}, \quad (25)$$

then $\Omega_{x,3,H}^{(v)}(y) > A$.

(b) $\Omega_{x,3,H}^{(v)}(\cdot)$ *is coercive.*

Proof: First, by the definition of $\Omega_{x,3,H}(\cdot)$ and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \Omega_{x,3,H}^{(v)}(y) &\geq f(x) - \|\nabla f(x)\|_* \|y - x\| - \frac{1}{2} \|\nabla^2 f(x)\| \|y - x\|^2 \\ &\quad - \frac{1}{6} \|D^3 f(x)\| \|y - x\|^3 + \frac{H}{6} \|y - x\|^{3+v}. \end{aligned}$$

Thus, to ensure $\Omega_{x,3,H}^{(v)}(y) > A$, it is enough to have

$$\begin{aligned} \frac{H}{6} \|y - x\|^{3+v} &> (A - f(x)) + \|\nabla f(x)\|_* \|y - x\| + \frac{1}{2} \|\nabla^2 f(x)\| \|y - x\|^2 \\ &\quad + \frac{1}{6} \|D^3 f(x)\| \|y - x\|^3, \end{aligned}$$

which is equivalent to

$$\|y - x\|^v > \frac{6(A - f(x))}{H\|y - x\|^3} + \frac{6\|\nabla f(x)\|_*}{H\|y - x\|^2} + \frac{3\|\nabla^2 f(x)\|}{H\|y - x\|} + \frac{\|D^3 f(x)\|}{H}. \quad (26)$$

Note that, if (25) holds, then (26) holds. Therefore,

$$(25) \implies (26) \implies \Omega_{x,3,H}^{(v)}(y) > A.$$

This proves statement (a).

Finally, given $A > 0$, if

$$\|y\| > \|x\| + \max \left\{ \left[\frac{6(A - f(x))}{H} \right]^{1/3}, \left[\frac{6\|\nabla f(x)\|_*}{H} \right]^{1/2}, \frac{3\|\nabla^2 f(x)\|}{H}, \left[3 + \frac{\|D^3 f(x)\|}{H} \right]^{1/v} \right\},$$

then, by (a), we have $\Omega_{x,3,H}^{(v)}(y) > A$. Since $A > 0$ is arbitrary, we conclude that

$$\lim_{\|y\| \rightarrow +\infty} \Omega_{x,3,H}^{(v)}(y) = +\infty.$$

This proves statement (b). ■

As a corollary of Lemma 3.5, we can establish the following upper bound for $\|y - x\|$ whenever y belongs to the convex hull of a suitable sublevel set of $\Omega_{x,3,H}(\cdot)$.

Lemma 3.6: *Given $x \in \mathbb{E}$, $H > 0$ and $v \neq 0$, let*

$$\mathcal{L}_H(x) = \left\{ z \in \mathbb{E} : \Omega_{x,3,H}^{(v)}(z) \leq f(x) \right\}. \quad (27)$$

Then,

$$\begin{aligned} \|y - x\| &\leq \max \left\{ 1, \left[\frac{6\|\nabla f(x)\|_*}{H} \right]^{1/2}, \frac{3\|\nabla^2 f(x)\|}{H}, \left[3 + \frac{\|D^3 f(x)\|}{H} \right]^{1/v} \right\} \\ &\equiv D_{x,H}, \end{aligned} \quad (28)$$

for all $y \in \text{co}(\mathcal{L}_H(x))$. Consequently,

$$\sup \left\{ \|\nabla^2 \rho_x(y)\| : y \in \text{co}(\mathcal{L}_H(x)) \right\} \leq \|\nabla^2 f(x)\| + (2 + v)D_{x,H}^2 \equiv \hat{N}_{x,H}. \quad (29)$$

Proof: By Lemma 3.5(a) with $A = f(x)$, we have the implication

$$\begin{aligned} \|y - x\| &> \max \left\{ \left[\frac{6\|\nabla f(x)\|_*}{H} \right]^{1/2}, \frac{3\|\nabla^2 f(x)\|}{H}, \left[3 + \frac{\|D^3 f(x)\|}{H} \right]^{1/\nu} \right\} \\ &\implies \Omega_{x,3,H}^{(\nu)}(y) > f(x), \end{aligned}$$

whose contrapositive is

$$\begin{aligned} \Omega_{x,3,H}^{(\nu)}(y) \leq f(x) &\implies \|y - x\| \\ &\leq \max \left\{ \left[\frac{6\|\nabla f(x)\|_*}{H} \right]^{1/2}, \frac{3\|\nabla^2 f(x)\|}{H}, \left[3 + \frac{\|D^3 f(x)\|}{H} \right]^{1/\nu} \right\}. \end{aligned}$$

Thus, if $y \in \mathcal{L}_H(x)$, then the bound (28) holds for y . Consequently, as in the proof of Lemma 3.4, we obtain

$$\|y - x\| \leq D_{x,H}, \quad \forall y \in \text{co}(\mathcal{L}_H(x)). \quad (30)$$

Finally, (29) follows by (18), $D_{x,H} \geq 1$ and (30). ■

3.2. Gradient method and its efficiency

Let us consider the problem

$$\min_{y \in \mathbb{E}} \Omega_{x,3,H}^{(\nu)}(y) \quad (31)$$

By Theorem 3.3, Remark 3.1 and Lemma 3.6, it follows that:

- $\Omega_{x,3,H}^{(\nu)}(\cdot)$ is L_H -smooth;
- $\rho_x(\cdot)$ is uniformly convex of degree $3 + \nu$ with parameter $2^{-(1+\nu)}$;
- $\rho_x(\cdot)$ is twice differentiable and $\|\nabla^2 \rho_x(y)\|$ is bounded on $\text{co}(\mathcal{L}_H(x))$.

This means that $\Omega_{x,3,H}^{(\nu)}(\cdot)$ and $\rho_x(\cdot)$ satisfy assumptions H1–H3 in Appendix. Therefore, we can apply Algorithm A (see page 17) to solve (31). The Bregman distance corresponding to $\rho_x(\cdot)$ is

$$\beta_{\rho_x}(u, v) = \rho_x(v) - \rho_x(u) - \langle \nabla \rho_x(u), v - u \rangle. \quad (32)$$

Thus, Algorithm A applied to (31) can be rewritten as follows.

Algorithm 1. Algorithm A applied to (31)**Step 0.** Choose $L_0 > 0$. Set $y_0 = x$ and $k := 0$.**Step 1.** Set $i := 0$.**Step 1.1.** Compute $y_{k,i}^+ = \arg \min_{z \in \mathbb{E}} \{ \langle \nabla \Omega_{x,3,H}^{(v)}(y_k), z - y_k \rangle + 2^i L_k \beta_{\rho_x}(y_k, z) \}$.**Step 1.2.** If

$$\Omega_{x,3,H}^{(v)}(y_{k,i}^+) \leq \Omega_{x,3,H}^{(v)}(y_k) + \langle \nabla \Omega_{x,3,H}^{(v)}(y_k), y_{k,i}^+ - y_k \rangle + 2^i L_k \beta_{\rho_x}(y_k, y_{k,i}^+),$$

set $i_k := i$ and go to Step 2. Otherwise, set $i := i + 1$ and go to Step 1.1.**Step 2.** Set $y_{k+1} = y_{k,i_k}^+$ and $L_{k+1} = 2^{i_k-1} L_k$.**Step 3.** Set $k := k + 1$ and go to Step 1.

When H is sufficiently large, the next theorem establishes that Algorithm 1 takes at most $\mathcal{O}(\log(\epsilon^{-1}))$ iterations to find an ϵ -stationary point of $\Omega_{x,3,H}^{(v)}(\cdot)$.

Theorem 3.7: Suppose that f has a global minimizer x^* and that $x \in \mathcal{F}(x_0)$ with

$$\sup_{y \in \mathcal{F}(x_0)} \|y - x^*\| \leq R_0 < +\infty, \quad R_0 \geq 1. \quad (33)$$

Denote $M_H = \max\{2L_0, 4L_H\}$, with L_H defined in (16) and

$$N_x = \|\nabla^2 f(x)\| + 12R_0^2. \quad (34)$$

Let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm 1. If $H > [6/(3 + \nu)]H_{f,3}(\nu)$ and $\|\nabla \Omega_{x,3,H}^{(v)}(y_{T+1})\|_* > \epsilon$ for a given $\epsilon \in (0, 1)$, then

$$T \leq \left[\log_2 \left(\frac{M_H}{M_H - \mu_H} \right) \right]^{-1} [C_{x,H} + (3 + \nu)] \log_2(\epsilon^{-1}), \quad (35)$$

where

$$C_{x,H} = \log_2 \left(\frac{4(3 + \nu)M_H^{2+\nu}N_x^3\mu_H}{2^{-(1+\nu)}} \right). \quad (36)$$

Proof: Since $H > [6/(3 + \nu)]H_{f,3}(\nu)$, it follows from Theorem 3.3 that $\Omega_{x,3,H}^{(v)}(\cdot)$ is L_H -smooth and μ_H -strongly convex relative to $\rho_x(\cdot)$, with $\mu_H > 1$. Moreover, by Remark 3.1 and Lemma 3.4, function $\rho_x(\cdot)$ is twice differentiable, uniformly convex of degree $3 + \nu$ with parameter $2^{-(1+\nu)}$ and satisfies

$$\sup \{ \|\nabla^2 \rho_x(y)\| : y \in \text{co}(\mathcal{L}_H(x)) \} \leq N_x.$$

Thus, $\Omega_{x,3,H}^{(v)}(\cdot)$ and $\rho_x(\cdot)$ satisfy assumptions H1-H4 in Appendix with $L = L_H$, $q = 3 + \nu$, $\sigma_q = 2^{-(1+\nu)}$, $N = N_x$ and $\mu = \mu_H$. Consequently, by Corollary A.6, we must have

$$T \leq \left[\log_2 \left(\frac{M_H}{M_H - \mu_H} \right) \right]^{-1} [\tilde{C}_{x,H} + (3 + \nu)] \log_2(\epsilon^{-1}), \quad (37)$$

where

$$\tilde{C}_{x,H} = \log_2 \left(\frac{2(3+\nu)M_H^{2+\nu}N_x\mu_H\beta_{\rho_x}(x,S(x))}{2^{-(1+\nu)}} \right). \quad (38)$$

with $S(x) \in \arg \min_{y \in \mathbb{E}} \Omega_{x,3,H}^{(\nu)}(y)$. Clearly, $S(x) \in \mathcal{L}_H(x)$. Thus, it follows from (32), (14), $R_0 \geq 1$ and (34) that

$$\begin{aligned} \beta_{\rho_x}(x,S(x)) &= \rho_x(S(x)) \\ &= \frac{1}{2} \langle \nabla^2 f(x)(S(x) - x), S(x) - x \rangle + \frac{1}{3+\nu} \|S(x) - x\|^{3+\nu} \\ &\leq \frac{1}{2} \|\nabla^2 f(x)\| \|S(x) - x\|^2 + \frac{1}{3+\nu} \|S(x) - x\|^{3+\nu} \\ &\leq \frac{1}{2} [\|\nabla^2 f(x)\| + \|S(x) - x\|^{1+\nu}] \|S(x) - x\|^2 \\ &\leq \frac{1}{2} [\|\nabla^2 f(x)\| + (2R_0)^{1+\nu}] (2R_0)^2 \\ &\leq 2 [\|\nabla^2 f(x)\| + 4R_0^{1+\nu}] R_0^2 \\ &\leq 2N_x^2. \end{aligned} \quad (39)$$

Finally, combining (37)–(39), we obtain (35)–(36). \blacksquare

Remark 3.2: If $x \notin \mathcal{F}(x_0)$, by Lemma 3.6 we also have $T \leq \mathcal{O}(\log_2(\epsilon^{-1}))$ with N_x replaced by $\hat{N}_{x,H}$ in (36), as long as $\nu \neq 0$. In both cases, it is worth mentioning that the potentially ‘bad’ constants N_x and $\hat{N}_{x,H}$ are inside the $\log_2(\cdot)$ in (36).

When $H \leq [6/(3+\nu)]H_{f,3}(\nu)$, problem (31) may be nonconvex. Even in this case, we can establish complexity bounds for Algorithm 1 if $\nu \neq 0$.

Theorem 3.8: Given $\epsilon \in (0, 1)$, let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm 1 such that

$$\|\nabla \Omega_{x,3,H}^{(\nu)}(y_k)\|_* > \epsilon \quad \text{for } k = 0, \dots, T. \quad (40)$$

Then, the following statements are true:

(a) If $H < [6/(3+\nu)]H_{f,3}(\nu)$, then

$$T \leq \left\lceil \frac{\hat{N}_{x,H}^{3+\nu} (3+\nu) M_H^{2+\nu} F_x}{2^{-(1+\nu)}} \right\rceil \epsilon^{-(3+\nu)},$$

where $\hat{N}_{x,H}$ is defined in (29) and

$$F_x = D_{x,H} \left[\|\nabla f(x)\|_* + \frac{1}{2} \|\nabla^2 f(x)\| D_{x,H} + \|D^3 f(x)\| D_{x,H}^2 \right],$$

with $D_{x,H}$ given in (28).

(b) If $H = [6/(3 + \nu)]H_{f,3}(\nu)$, then

$$T \leq 3 \left[M_H \hat{N}_{x,H} \right]^{(3+\nu)/2} \left[\frac{(3 + \nu) \hat{N}_{x,H}^2}{2^{-(2+\nu)}} \right]^{1/2} \epsilon^{-(3+\nu)/2}$$

Proof: Combining Theorem 3.3(a), Remark 3.1, Lemma 3.6 and (40) with Theorem A.2, we obtain

$$T \leq \left[\frac{\hat{N}_{x,H}^{3+\nu} (3 + \nu) M_H^{2+\nu} \left(f(x) - \Omega_{x,3,H}^{(\nu)}(S(x)) \right)}{2^{-(1+\nu)}} \right] \epsilon^{-(3+\nu)}, \quad (41)$$

with $S(x) \in \arg \min_{y \in \mathbb{E}} \Omega_{x,3,H}^{(\nu)}(y)$. Since $S(x) \in \mathcal{L}_H(x)$, it follows from (19) that

$$\begin{aligned} f(x) - \Omega_{x,3,H}^{(\nu)}(S(x)) &= \langle \nabla f(x), x - S(x) \rangle + \frac{1}{2} \langle \nabla^2 f(x)(S(x) - x), S(x) - x \rangle \\ &\quad - \frac{1}{6} D^3 f(x)[S(x) - x]^3 - \frac{H}{6} \|S(x) - x\|^{3+\nu} \\ &\leq \|\nabla f(x)\|_* \|S(x) - x\| + \frac{1}{2} \|\nabla^2 f(x)\| \|S(x) - x\|^2 \\ &\quad + \|D^3 f(x)\| \|S(x) - x\|^3 \\ &\leq D_{x,H} \left[\|\nabla f(x)\|_* + \frac{1}{2} \|\nabla^2 f(x)\| D_{x,H} + \|D^3 f(x)\| D_{x,H}^2 \right] \\ &= F_x. \end{aligned} \quad (42)$$

Thus, from (41) and (42) we see that statement (a) is true.

Now, suppose that $H = [6/(3 + \nu)]H_{f,3}(\nu)$. Then, by Theorem 3.3(b) functions $\Omega_{x,3,H}^{(\nu)}(\cdot)$ and $\rho_x(\cdot)$ satisfy assumption H4 in Appendix with $\mu = 0$. Consequently, by (40) and Corollary A.5 we have

$$T \leq 3 \left[M_H \hat{N}_{x,H} \right]^{(3+\nu)/2} \left[\frac{(3 + \nu) \beta_{\rho_x}(x, S(x))}{2^{-(1+\nu)}} \right]^{1/2} \epsilon^{-(3+\nu)/2}. \quad (43)$$

As in the proof of Theorem 3.7, by (29) we have

$$\beta_{\rho_x}(x, S(x)) \leq \frac{1}{2} \hat{N}_{x,H}^2. \quad (44)$$

Thus, combining (43) and (44), we see that statement (b) is also true. ■

In view of Lemma 2.1 and Theorem 3.7, if $H > [6/(3 + \nu)]H_{f,3}(\nu)$, then Algorithm 1 takes at most $\mathcal{O}(\log_2(\epsilon^{-1}))$ iterations to generate x^+ such that either $\|\nabla f(x^+)\|_* \leq \epsilon$ or (5)–(6) holds for $p = 3$. In contrast, by Theorem 3.8, if $H = [6/(3 + \nu)]H_{f,3}(\nu)$ or $H < [6/(3 + \nu)]H_{f,3}(\nu)$, this iteration complexity bound is increased to $\mathcal{O}(\epsilon^{-(3+\nu)/2})$ and $\mathcal{O}(\epsilon^{-(3+\nu)})$, respectively, in the case $\nu \neq 0$.

4. Auxiliary problems in composite minimization

For third-order tensor methods designed to composite minimization [10,12], the auxiliary problems take the form:

$$\min_{y \in \mathbb{E}} \tilde{\Omega}_{x,3,H}^{(v)}(y) \equiv \Omega_{x,3,H}^{(v)}(y) + \varphi(y), \quad (45)$$

where $\Omega_{x,3,H}^{(v)}(\cdot)$ is defined by (4), $H \geq (p-1)H_{f,3}(v)$ and $\varphi: \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ is a simple closed convex function whose effective domain has nonempty relative interior. In this case, we are interested in finding an approximate solution x^+ for (45) such that²

$$\tilde{\Omega}_{x,3,H}^{(v)}(x^+) \leq f(x) + \varphi(x) \equiv \tilde{f}(x), \quad (46)$$

and

$$\|\nabla \Omega_{x,3,H}^{(v)}(x^+) + g_\varphi(x^+)\|_* \leq \theta \|x^+ - x\|^{2+v}, \quad (47)$$

for some $g_\varphi(x^+) \in \partial\varphi(x^+)$. For general $p \geq 2$, we have the following generalization of Lemma 2.1.

Lemma 4.1: *Let $x \in \mathbb{E}$, $H, \theta > 0$ and $\delta \in (0, 1)$. Given $g_\varphi(x^+) \in \partial\varphi(x^+)$, if*

$$\|\nabla f(x^+) + g_\varphi(x^+)\|_* \geq \delta, \quad (48)$$

and

$$\|\nabla \Omega_{x,p,H}^{(v)}(x^+) + g_\varphi(x^+)\|_* \leq \min \left\{ \frac{1}{2}, \frac{\theta(p-1)!}{2[H_{f,p}(v) + H(p+v)]} \right\} \delta, \quad (49)$$

then x^+ satisfies (47).

Proof: It follows as in the proof of Lemma 2.1. ■

Suppose that $H \geq 2H_{f,3}(v)$. Then, in view of the relative smoothness properties of $\Omega_{x,3,H}^{(v)}(\cdot)$ established in Section 3.1, we can apply Algorithm B (see page 25) to solve (45):

Algorithm 2. Algorithm B applied to (45)

Step 0. Set $y_0 = x$ and $k := 0$.

Step 1. Compute $y_{k+1} = \arg \min_{z \in \mathbb{E}} \{ \langle \nabla \Omega_{x,3,H}(y_k), z - y_k \rangle + 2L_H \beta_{\rho_x}(y_k, z) + \varphi(z) \}$.

Step 2. Set $k := k + 1$ and go to Step 1.

The next theorem establishes that Algorithm 2 takes at most $\mathcal{O}(\log_2(\epsilon^{-1}))$ iterations to generate x^+ such that

$$\|\nabla\Omega_{x,3,H}^{(v)}(x^+) + g_\varphi(x^+)\|_* \leq \epsilon,$$

with $g_\varphi(x^+) \in \partial\varphi(x^+)$.

Theorem 4.2: *Suppose that $\tilde{f} = f + \varphi$ has a global minimizer x^* and that*

$$x \in \tilde{\mathcal{F}}(x_0) \equiv \left\{ z \in \mathbb{E} : \tilde{f}(z) \leq \tilde{f}(x_0) \right\},$$

with

$$\sup_{y \in \tilde{\mathcal{F}}(x_0)} \|y - x^*\| \leq R_0 < +\infty, \quad R_0 \geq 1.$$

Assume that $H \geq 2H_{f,3}(v)$ and let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm 2. Then, for all $k \geq 1$, we have

$$g_\varphi(y_k) \equiv 2L_H [\nabla\rho_x(y_{k-1}) - \nabla\rho_x(y_k)] - \nabla\Omega_{x,3,H}^{(v)}(y_{k-1}) \in \partial\varphi(y_k). \quad (50)$$

Moreover, if

$$\|\nabla\Omega_{x,3,H}^{(v)}(y_{T+1}) + g_\varphi(y_{T+1})\|_* > \epsilon \quad (51)$$

for a given $\epsilon \in (0, 1)$, then

$$T \leq \left[\log_2 \left(\frac{2L_H}{2L_H - \mu_H} \right) \right]^{-1} [K_{x,H} + (3 + v)] \log_2(\epsilon^{-1}), \quad (52)$$

where

$$K_{x,H} = \log_2 \left(\frac{4(3 + v)(2L_H)^{2+v} N_x^3 \mu_H}{2^{-(1+v)}} \right) \quad (53)$$

with N_x given in (34).

Proof: By Lemma A.8 and $\text{ri}(\text{dom } \varphi) \neq \emptyset$, we have

$$u(y_k) \equiv g_\varphi(y_k) + \nabla\Omega_{x,3,H}^{(v)}(y_k) \in \partial\tilde{\Omega}_{x,3,H}^{(v)}(y_k) = \left\{ \nabla\Omega_{x,3,H}^{(v)}(y_k) \right\} + \partial\varphi(y_k).$$

Thus, $g_\varphi(y_k) = u(y_k) - \nabla\Omega_{x,3,H}^{(v)}(y_k) \in \partial\varphi(y_k)$, and so (50) holds. Moreover, by (51), we have

$$\|u(y_{T+1})\|_* > \epsilon.$$

Then, the bound (52) on T follows directly from Corollary A.10. ■

In view of Theorem 4.2, if $H \geq 2H_{f,3}(v)$, Algorithm 2 takes at most $\mathcal{O}(\log_2(\epsilon^{-1}))$ iterations to generate x^+ such that either $\|\nabla f(x^+) + g_\varphi(x^+)\|_* \leq \epsilon$ or (48)–(49) holds, for $g_\varphi(x^+) \in \partial\varphi(x^+)$ defined in (50).

5. Conclusion

In this paper, we studied the auxiliary problems that appear in non-universal adaptive p -order tensor methods for unconstrained minimization of convex functions with Hölder continuous p th derivatives [10,11]. For $p = 3$, we consider the use of Gradient Methods with Bregman Distance. When the regularization parameter is sufficiently large, we prove that Bregman Gradient Methods applied to the corresponding tensor model takes at most $\mathcal{O}(\log(\epsilon^{-1}))$ iterations to find either a suitable approximate stationary point of the tensor model or an ϵ -approximate stationary point of the original objective function. The authors believe this work is a step towards implementable third-order tensor methods for convex optimization. Future research includes the development of methods for the auxiliary problems in universal tensor methods and numerical experimentation.

Notes

1. See also [2] for a version of relative smoothness without strong convexity.
2. See, e.g. Section 5 in [10].

Acknowledgments

The authors are very grateful to two anonymous referees, whose comments helped to improve the first version of this paper. Generous support of Alibaba Group is also acknowledged.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

G.N. Grapiglia was supported by the National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico) – Brazil (grant 406269/2016-5) and by the European Research Council Advanced Grant 788368. Yu. Nesterov was supported by the European Research Council Advanced Grant 788368.

Notes on contributors

G.N. Grapiglia obtained his doctoral degree in Mathematics in 2014 at Universidade Federal do Paraná (UFPR), Brazil. Currently he is an Assistant Professor at UFPR. His research cover the development, analysis and application of optimization methods.

Y. Nesterov is a professor at the Center for Operations Research and Econometrics (CORE) in Catholic University of Louvain (UCL), Belgium. He received his Ph.D. degree (Applied Mathematics) in 1984 at the Institute of Control Sciences, Moscow. His research interests are related to complexity issues and efficient methods for solving various optimization problems. He has received several international prizes, among which are the Dantzig Prize from SIAM and Mathematical Programming society (2000), the John von Neumann Theory Prize from INFORMS (2009), the SIAM

Outstanding Paper Award (2014), and the Euro Gold Medal from the Association of European Operations Research Societies (2016). In 2018 he also won an Advanced Grant from the European Research Council.

ORCID

G.N. Grapiglia  <http://orcid.org/0000-0003-3284-3371>

Yu. Nesterov  <http://orcid.org/0000-0002-0542-8757>

References

- [1] M. Baes, *Estimate sequence methods: Extensions and approximations*. Optimization Online (2009). Available at http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf
- [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, J.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] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM J. Optim. 28 (2018), pp. 2131–2151.
- [5] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *Universal regularized methods—varying the power, the smoothness, and the accuracy*, SIAM J. Optim. 29 (2019), pp. 595–615.
- [6] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim. 3 (1993), pp. 538–543.
- [7] N. Doikov and Yu. Nesterov, *Minimizing uniformly convex functions by cubic regularization of newton method* (2019). Available at arXiv: 1905.02671 [math. OC].
- [8] G.N. Grapiglia and Yu. Nesterov, *Regularized Newton methods for minimizing functions with Hölder continuous Hessians*, SIAM J. Optim. 27 (2017), pp. 478–506.
- [9] G.N. Grapiglia and Yu. Nesterov, *Accelerated regularized Newton methods for minimizing composite convex functions*, SIAM J. Optim. 29 (2019), pp. 77–99.
- [10] G.N. Grapiglia and Yu. Nesterov, *Tensor methods for finding approximate stationary points of convex functions* (2019). Available at arXiv: 1907.07053 [math. OC].
- [11] G.N. Grapiglia and Yu. Nesterov, *Tensor methods for minimizing functions with Hölder continuous higher-order derivatives* (2019). Available at arXiv: 1904.12559 [math. OC].
- [12] B. Jiang, T. Lin, and S. Zhang, *A unified adaptive tensor approximation scheme to accelerated composite convex optimization* (2018). Available at arXiv: 1811.02427 [math O.C.].
- [13] G. Lan, Z. Lu, and R.D.C. Monteiro, *Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming*, Math. Program. 126 (2011), pp. 1–29.
- [14] 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.
- [15] J.M. Martínez, *On high-order model regularization for constrained optimization*, SIAM J. Optim. 27 (2017), pp. 2447–2458.
- [16] Yu. Nesterov, *Accelerating the cubic regularization of Newton’s method on convex problems*, Math. Program. 112 (2008), pp. 159–181.
- [17] Yu. Nesterov, *Implementable tensor methods in unconstrained convex optimization*. CORE Discussion Paper 2018/05.
- [18] Yu. Nesterov and A. Nemirovskii, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*, SIAM, Philadelphia, PA, 1994.
- [19] Yu Nesterov and B.T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program. 108 (2006), pp. 177–205.
- [20] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, Tech. Rep., May 21, 2008.

Appendix. Adaptive Bregman proximal gradient method

A.1 Smooth minimization

Consider the following optimization problem

$$\min_{y \in \mathbb{E}} g(y), \quad (\text{A1})$$

where $g : \mathbb{E} \rightarrow \mathbb{R}$ is L -smooth relative to a convex and smooth function $d(\cdot)$, that is, for all $x, y \in \mathbb{E}$,

$$g(x) \leq g(y) + \langle \nabla g(y), x - y \rangle + L\beta_d(y, x), \quad (\text{A2})$$

with

$$\beta_d(y, x) := d(x) - d(y) - \langle \nabla d(y), x - y \rangle \quad (\text{A3})$$

being the Bregman distance corresponding to $d(\cdot)$. We assume that $g(\cdot)$ has at least one global minimizer $y^* \in \mathbb{E}$. We do not assume the convexity of $g(\cdot)$ yet.

We shall consider the following adaptive version of the Proximal Gradient Scheme proposed in [14]:

Algorithm A. Adaptive Proximal Gradient Method

Step 0. Choose $y_0 \in \mathbb{E}$, $L_0 > 0$ and set $k := 0$.

Step 1. Set $i := 0$.

Step 1.1. Compute

$$y_{k,i}^+ = \arg \min_{x \in \mathbb{E}} \{ \langle \nabla g(y_k), x - y_k \rangle + 2^i L_k \beta_d(y_k, x) \}. \quad (\text{A4})$$

Step 1.2. If

$$g(y_{k,i}^+) \leq g(y_k) + \langle \nabla g(y_k), y_{k,i}^+ - y_k \rangle + 2^i L_k \beta_d(y_k, y_{k,i}^+), \quad (\text{A5})$$

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

Step 2. Set $y_{k+1} = y_{k,i_k}^+$ and $L_{k+1} = 2^{i_k-1} L_k$.

Step 3. Set $k := k + 1$ and go to Step 1.

Let us assume that:

(H1) $g(\cdot)$ is L -smooth relative to $d(\cdot)$.

(H2) $d(\cdot)$ is twice differentiable and uniformly convex of degree q , with parameter $\sigma_q > 0$.

(H3) There exists a constant $N > 0$ such that

$$\sup \{ \|\nabla^2 d(y)\| : y \in \text{co}(\mathcal{L}(y_0)) \} \leq N,$$

where $\mathcal{L}(y_0) = \{y \in \mathbb{E} : g(y) \leq g(y_0)\}$.

The next lemma gives a global upper bound on L_k and a lower bound on the functional decrease in successive iterations.

Lemma A.1: *Suppose that H1 holds and let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm A. Then, for all k ,*

$$L_k \leq \max \{L_0, 2L\}, \quad (\text{A6})$$

and

$$g(y_k) - g(y_{k+1}) \geq 2L_{k+1} \beta_d(y_{k+1}, y_k). \quad (\text{A7})$$

Proof: Let us prove by induction that (A6) is true. It is obvious for $k = 0$. Assume that (A6) is true for some $k \geq 0$. Then, it follows from H1 and (A2) that $2^{ik}L_k$ cannot be bigger than $4L$, since otherwise the line search procedure should have stopped earlier. Thus,

$$L_{k+1} = \max \{L_0, 2^{ik-1}L_k\} \leq \max \{L_0, 2L\},$$

that is, (A6) also holds for $k + 1$, which concludes the induction argument.

Now, let us prove (A7). In view of (A4), we have

$$\nabla g(y_k) + 2^{ik}L_k (\nabla d(y_{k+1}) - \nabla d(y_k)) = 0,$$

which gives

$$\langle \nabla g(y_k), y_{k+1} - y_k \rangle = -2^{ik}L_k \langle \nabla d(y_{k+1}) - \nabla d(y_k), y_{k+1} - y_k \rangle. \quad (\text{A8})$$

Then, combining (A5) and (A8), we get

$$\begin{aligned} g(y_{k+1}) &\leq g(y_k) - 2^{ik}L_k \langle \nabla d(y_{k+1}) - \nabla d(y_k), y_{k+1} - y_k \rangle + 2^{ik}L_k \beta_d(y_k, y_{k+1}) \\ &= g(y_k) - 2^{ik}L_k \langle \nabla d(y_{k+1}) - \nabla d(y_k), y_{k+1} - y_k \rangle \\ &\quad + 2^{ik}L_k [d(y_{k+1}) - d(y_k) - \langle \nabla d(y_k), y_{k+1} - y_k \rangle] \\ &= g(y_k) - 2^{ik}L_k [d(y_k) - d(y_{k+1}) - \langle \nabla d(y_{k+1}), y_k - y_{k+1} \rangle] \\ &= g(y_k) - 2^{ik}L_k \beta_d(y_{k+1}, y_k), \end{aligned}$$

that is

$$g(y_k) - g(y_{k+1}) \geq 2^{ik}L_k \beta_d(y_{k+1}, y_k). \quad (\text{A9})$$

Finally, since $L_{k+1} = 2^{ik-1}L_k$, (A7) follows directly from (A9). \blacksquare

Theorem A.2: Suppose that H1–H3 hold. Then, for all $k \geq 0$ we have

$$g(y_k) - g(y_{k+1}) \geq \frac{\sigma_q}{q[\max \{2L_0, 4L\}]^{q-1}N^q} \|\nabla g(y_k)\|_*^q, \quad (\text{A10})$$

where σ_q and N are specified in H2 and H3, respectively. Moreover, for all $T \geq 1$,

$$\min_{0 \leq k \leq T-1} \|\nabla g(y_k)\|_* \leq N \left[\frac{q[\max \{2L_0, 4L\}]^{q-1} (g(y_0) - g(y^*))}{\sigma_q} \right]^{1/q} \left(\frac{1}{T} \right)^{1/q}. \quad (\text{A11})$$

Consequently, if

$$\|\nabla g(y_k)\|_* > \epsilon, \quad \text{for } k = 0, \dots, T-1, \quad (\text{A12})$$

for a given $\epsilon > 0$, we have

$$T \leq \left[\frac{N^q q [\max \{2L_0, 4L\}]^{q-1} (g(y_0) - g(y^*))}{\sigma_q} \right] \epsilon^{-q}. \quad (\text{A13})$$

Proof: By H2, $d(\cdot)$ is uniformly convex of degree q with parameter $\sigma_q > 0$. Therefore,

$$\beta_d(y_{k+1}, y_k) \geq \frac{\sigma_q}{q} \|y_{k+1} - y_k\|^q.$$

In this case, by (A7) we obtain

$$g(y_k) - g(y_{k+1}) \geq \frac{2L_{k+1}\sigma_q}{q} \|y_{k+1} - y_k\|^q. \quad (\text{A14})$$

By the definition of y_{k+1} , this point satisfies the following first-order optimality condition:

$$\nabla g(y_k) + 2^{ik}L_k (\nabla d(y_{k+1}) - \nabla d(y_k)) = 0. \quad (\text{A15})$$

In view of H3, it follows from the mean value theorem that ∇d is N -Lipschitz continuous on $\text{co}(\mathcal{L}(y_0))$. From (A14), we see that $\{g(y_k)\}_{k \geq 0}$ is nonincreasing, and so $\{y_k\} \subset \mathcal{L}(y_0)$. Combining

these facts, we get

$$\|\nabla d(y_{k+1}) - \nabla d(y_k)\|_* \leq N\|y_{k+1} - y_k\|, \quad \forall k. \quad (\text{A16})$$

Then, it follows from (A15), (A16) and (A6) that

$$\|\nabla g(y_k)\|_* \leq 2^{ik} L_k \|\nabla d(y_{k+1}) - \nabla d(y_k)\|_* \leq (2^{ik} L_k) N \|y_{k+1} - y_k\|.$$

Thus,

$$\|y_{k+1} - y_k\| \geq \frac{1}{2L_{k+1}N} \|\nabla g(y_k)\|_*. \quad (\text{A17})$$

Combining (A14), (A17) and (A6), we obtain

$$\begin{aligned} g(y_k) - g(y_{k+1}) &\geq \frac{2L_{k+1}\sigma_q}{q} \|y_{k+1} - y_k\|^q \\ &\geq \frac{2L_{k+1}\sigma_q}{q} \frac{1}{(2L_{k+1})^q N^q} \|\nabla g(y_k)\|_*^q \\ &= \frac{\sigma_q}{q(2L_{k+1})^{q-1} N^q} \|\nabla g(y_k)\|_*^q \\ &\geq \frac{\sigma_q}{q[2\max\{L_0, 2L\}]^{q-1} N^q} \|\nabla g(y_k)\|_*^q, \end{aligned}$$

which gives (A10). Summing up inequalities (A10) for $k = 0, \dots, T-1$, we get

$$\begin{aligned} g(y_0) - g(y^*) &\geq g(y_0) - g(y_T) \\ &= \sum_{k=0}^{T-1} g(y_k) - g(y_{k+1}) \\ &\geq \sum_{k=0}^{T-1} \frac{\sigma_q}{q[2\max\{L_0, 2L\}]^{q-1} N^q} \|\nabla g(y_k)\|_*^q \\ &\geq T \frac{\sigma_q}{q[2\max\{L_0, 2L\}]^{q-1} N^q} \left(\min_{0 \leq k \leq T-1} \|\nabla g(y_k)\|_* \right)^q, \end{aligned}$$

which gives (A11). Finally, (A13) follows directly from (A11) and (A12). ■

Now, let us consider the following additional assumption:

(H4) $g(\cdot)$ is μ -strongly convex relative to $d(\cdot)$.

Lemma A.3 (Three-Point Property): *Let $\phi(\cdot)$ and $d(\cdot)$ be convex functions and let $\beta_d(\cdot, \cdot)$ be the Bregman distance from $d(\cdot)$. Given $y \in \mathbb{E}$, let*

$$y^+ = \arg \min_{x \in \mathbb{E}} \{ \phi(x) + \beta_d(y, x) \}.$$

Then,

$$\phi(x) + \beta_d(y, x) \geq \phi(y^+) + \beta_d(y, y^+) + \beta_d(y^+, x), \quad \forall x \in \mathbb{E}. \quad (\text{A18})$$

Proof: See [6,13,20]. ■

The next theorem establishes sublinear and linear convergence rates for Algorithm A.

Theorem A.4: Suppose that H1, H2 and H4 hold and let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm A. Then,

$$g(y_k) - g(y^*) \leq \frac{\mu \beta_d(y_0, y^*)}{\left(1 + \frac{\mu}{\max\{2L_0, 4L\} - \mu}\right)^k - 1} \leq \frac{(\max\{2L_0, 4L\} - \mu) \beta_d(y_0, y^*)}{k}, \quad (\text{A19})$$

where, in the case $\mu = 0$, the middle expression is defined in the limit as $\mu \rightarrow 0^+$.

Proof: By H1 and Lemma A.1, it follows that $\{y_k\}_{k \geq 0}$ is well-defined. Let us denote $M_k = 2^{ik} L_k$. Then, for all $k \geq 1$, it follows from (A5) that

$$g(y_k) \leq g(y_{k-1}) + \langle \nabla g(y_{k-1}), y_k - y_{k-1} \rangle + M_k \beta_d(y_{k-1}, y_k). \quad (\text{A20})$$

In order to get an upper bound for the inner product in (A20), let us apply Lemma A.3 with $h = d$ and

$$\phi(x) = \frac{1}{M_k} \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle.$$

In this case, $y^+ = y_k$ and, for $y = y_{k-1}$, we obtain

$$\phi(x) + \beta_d(y_{k-1}, x) \geq \phi(y_k) + \beta_d(y_{k-1}, y_k) + \beta_d(y_k, x), \quad \forall x \in \mathbb{E},$$

that is

$$\langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + M_k \beta_d(y_{k-1}, x) \geq \langle \nabla g(y_{k-1}), y_k - y_{k-1} \rangle + M_k \beta_d(y_{k-1}, y_k) + M_k \beta_d(y_k, x).$$

This gives the upper bound

$$\begin{aligned} \langle \nabla g(y_{k-1}), y_k - y_{k-1} \rangle &\leq \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + M_k \beta_d(y_{k-1}, x) \\ &\quad - M_k \beta_d(y_{k-1}, y_k) - M_k \beta_d(y_k, x). \end{aligned} \quad (\text{A21})$$

Combining (A20) and (A21), we obtain

$$g(y_k) \leq g(y_{k-1}) + \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + M_k \beta_d(y_{k-1}, x) - M_k \beta_d(y_k, x). \quad (\text{A22})$$

By (A4), we have

$$g(x) \geq g(y_{k-1}) + \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + \mu \beta_d(y_{k-1}, x),$$

and so

$$\langle \nabla g(y_{k-1}), x - y_{k-1} \rangle \leq g(x) - g(y_{k-1}) - \mu \beta_d(y_{k-1}, x). \quad (\text{A23})$$

Now, using inequality (A23) in (A22), it follows that

$$g(y_k) \leq g(x) + (M_k - \mu) \beta_d(y_{k-1}, x) - M_k \beta_d(y_k, x).$$

Substituting $x = y^*$, we get

$$g(y_k) \leq g(y^*) + (M_k - \mu) \beta_d(y_{k-1}, y^*) - M_k \beta_d(y_k, y^*). \quad (\text{A24})$$

Since $\beta_d(y_{k-1}, y^*) \geq 0$ and $\mu \geq 0$, it follows that

$$\begin{aligned} 0 \leq g(y_k) - g(y^*) &\leq (M_k - \mu) \beta_d(y_{k-1}, y^*) - M_k \beta_d(y_k, y^*) \\ &\leq M_k [\beta_d(y_{k-1}, y^*) - \beta_d(y_k, y^*)] \end{aligned}$$

and so

$$\beta_d(y_{k-1}, y^*) - \beta_d(y_k, y^*) \geq 0. \quad (\text{A25})$$

Moreover, by Lemma A.1 we have

$$M_k = 2^{ik} L_k = 2(2^{i(k-1)} L_k) \leq 2L_{k+1} \leq \max\{2L_0, 4L\}. \quad (\text{A26})$$

Denote $M = \max\{2L_0, 4L\}$. In view of (A24)–(A26), we obtain

$$\begin{aligned} g(y_k) &\leq g(y^*) + (M_k - \mu)\beta_d(y_{k-1}, y^*) - M_k\beta_d(y_k, y^*) \\ &= g(y^*) + M_k[\beta_d(y_{k-1}, y^*) - \beta_d(y_k, y^*)] - \mu\beta_d(y_{k-1}, y^*) \\ &\leq g(y^*) + M[\beta_d(y_{k-1}, y^*) - \beta_d(y_k, y^*)] - \mu\beta_d(y_{k-1}, y^*) \\ &= g(y^*) + (\tilde{M} - \mu)\beta_d(y_{k-1}, y^*) - M\beta_d(y_k, y^*). \end{aligned} \quad (\text{A27})$$

Now, as in the proof of Theorem 3.1 in [14], we can show by induction that, for all $k \geq 1$,

$$\sum_{i=1}^k \left(\frac{M}{M-\mu}\right)^i g(y_i) \leq \sum_{i=1}^k \left(\frac{M}{M-\mu}\right)^i g(y^*) + M\beta_d(y_0, y^*) - \left(\frac{M}{M-\mu}\right)^k M\beta_d(y_k, y^*). \quad (\text{A28})$$

Since $\{g(y_k)\}$ is nonincreasing and $\beta_d(y_k, y^*)$ is nonnegative, it follows from (A28) that

$$\left[\sum_{i=1}^k \left(\frac{\tilde{M}}{M-\mu}\right)^i \right] (g(y_k) - g(y^*)) \leq M\beta_d(y_0, y^*), \quad \forall k \geq 1.$$

Thus, denoting

$$C_k = \frac{1}{\sum_{i=1}^k \left(\frac{M}{M-\mu}\right)^i}$$

we get

$$g(y_k) - g(y^*) \leq C_k M \beta_d(y_0, y^*), \quad \forall k \geq 1. \quad (\text{A29})$$

If $\mu = 0$, it follows that $C_k = 1/k$ and so (A29) becomes

$$g(y_k) - g(y^*) \leq \frac{M}{k} \beta_d(y_0, y^*). \quad (\text{A30})$$

On the other hand, if $\mu > 0$ we have

$$\sum_{i=1}^k \left(\frac{M}{M-\mu}\right)^i = \frac{\left(\frac{M}{M-\mu}\right) \left[\left(\frac{M}{M-\mu}\right)^k - 1 \right]}{\left(\frac{M}{M-\mu}\right) - 1} = \frac{M \left[\left(1 + \frac{\mu}{M-\mu}\right)^k - 1 \right]}{\mu}$$

which gives

$$C_k = \frac{\mu}{M \left[\left(1 + \frac{\mu}{M-\mu}\right)^k - 1 \right]}. \quad (\text{A31})$$

In this case, combining (A29) and (A31) we obtain

$$g(y_k) - g(y^*) \leq \frac{\mu \beta_d(y_0, y^*)}{\left[\left(1 + \frac{\mu}{M-\mu}\right)^k - 1 \right]}. \quad (\text{A32})$$

Thus, (A19) follows from (A30), (A32) and $M = \max\{2L_0, 4L\}$. ■

Corollary A.5: *Suppose that H1–H3 hold and let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm A. Additionally, assume that H4 holds with $\mu = 0$. If $T = 3s$ for some $s \geq 1$, then*

$$\min_{0 \leq k \leq T-1} \|\nabla g(y_k)\|_* \leq MN \left[\frac{q\beta_d(y_0, y^*)}{\sigma_q} \right]^{1/q} \left(\frac{3}{T}\right)^{2/q}, \quad (\text{A33})$$

where $M = \max\{2L_0, 4L\}$. Consequently, if

$$\|\nabla g(y_k)\|_* > \epsilon \quad \text{for } k = 0, \dots, T-1, \quad (\text{A34})$$

for a given $\epsilon > 0$, then

$$T \leq 3 [\max\{2L_0, 4L\} N]^{q/2} \left[\frac{q\beta_d(y_0, y^*)}{\sigma_q} \right]^{1/2} \epsilon^{-q/2}. \quad (\text{A35})$$

Proof: By Theorem A.4, we have

$$g(y_i) - g(y^*) \leq \frac{M\beta_d(y_0, y^*)}{i}, \quad \forall i \geq 1.$$

Since $T = 3s$, in particular, it follows that

$$\begin{aligned} \frac{M\beta_d(y_0, y^*)}{2s} &\geq g(y_{2s}) - g(y^*) \\ &= g(y_T) - g(y^*) + \sum_{k=2s}^{T-1} [g(y_k) - g(y_{k+1})] \\ &\geq s \frac{\sigma_q}{qM^{q-1}N^q} \left(\min_{0 \leq k \leq T-1} \|\nabla g(y_k)\|_* \right)^q. \end{aligned}$$

Therefore,

$$\left(\min_{0 \leq k \leq T-1} \|\nabla g(y_k)\|_* \right)^q \leq \left[\frac{q(MN)^q \beta_d(y_0, y^*)}{\sigma_q} \right] \frac{1}{s^2}$$

which gives (A33). Finally, (A35) follows directly from (A24) and (A34). ■

Corollary A.6: Suppose that H1–H3 hold and let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm A. Additionally, assume that H4 holds with $\mu > 0$. Then, for all $T \geq \lceil \log_2(1 + \mu/(M - \mu)) \rceil^{-1}$, with $M = \max\{2L_0, 4L\}$, we have

$$\|\nabla g(y_T)\|_* \leq \left[\frac{2qM^{q-1}N\mu\beta_d(y_0, y^*)}{\sigma_q} \right]^{1/q} \left(\frac{1}{1 + \frac{\mu}{M-\mu}} \right)^{T/q} \quad (\text{A36})$$

Consequently, if $\|\nabla g(y_T)\|_* > \epsilon$, for a given $\epsilon \in (0, 1)$, then

$$T \leq \left[\log_2 \left(\frac{\max\{2L_0, 4L\}}{\max\{2L_0, 4L\} - \mu} \right) \right]^{-1} [C + q] \log_2(\epsilon^{-1}), \quad (\text{A37})$$

where

$$C = \log_2 \left(\frac{2q \max\{2L_0, 4L\}^{q-1} N \mu \beta_d(y_0, y^*)}{\sigma_q} \right). \quad (\text{A38})$$

Proof: By Theorems A.2 and A.4, for all $k \geq 1$ we have

$$\begin{aligned} \frac{\sigma_q}{qM^{q-1}N^q} \|\nabla g(y_k)\|_*^q &\leq g(y_k) - g(y^*) \\ &\leq \frac{\mu\beta_d(y_0, y^*)}{\left[\left(1 + \frac{\mu}{M-\mu}\right)^k - 1 \right]} \end{aligned}$$

In particular, it follows that

$$\|\nabla g(y_T)\|_*^q \leq \frac{qM^{q-1}N^q\mu\beta_d(y_0, y^*)}{\sigma_q \left[\left(1 + \frac{\mu}{M-\mu}\right)^T - 1 \right]} \quad (\text{A39})$$

Since $T \geq \lceil \log_2(1 + \mu/(M - \mu)) \rceil^{-1}$ we have

$$\left(1 + \frac{\mu}{M - \mu}\right)^T - 1 \geq \frac{1}{2} \left(1 + \frac{\mu}{M - \mu}\right)^T. \quad (\text{A40})$$

Thus, combining (A39) and (A40), it follows that

$$\|\nabla g(y_T)\|_*^q \leq \frac{2qM^{q-1}N^q\mu\beta_d(y_0, y^*)}{\sigma_q \left(1 + \frac{\mu}{M - \mu}\right)^T},$$

which gives (A36). Finally, (A37) follows directly from (A36), $\|g(y_T)\|_* > \epsilon$ and $\epsilon \in (0, 1)$. \blacksquare

In summary, if $g(\cdot)$ is L -smooth relative to a convex function $d(\cdot)$ which is uniformly convex of degree q , then Algorithm A takes at most $\mathcal{O}(\delta^{-q})$ iterations to generate a point y_k such that $\|\nabla g(y_k)\| \leq \delta$. If $g(\cdot)$ is also μ -strongly convex relative to $d(\cdot)$ with $\mu = 0$, then this complexity bound is reduced to $\mathcal{O}(\delta^{-q/2})$. Moreover, if $\mu > 0$, the complexity bound is further improved to $\mathcal{O}(\log(\delta^{-1}))$.

A.2 Composite minimization

Consider now the composite minimization problem

$$\min_{y \in \mathbb{E}} \tilde{g}(y) \equiv g(y) + \varphi(y), \quad (\text{A41})$$

where $g: \mathbb{E} \rightarrow \mathbb{R}$ is a twice-differentiable function satisfying H1 and H4 (on pages 17 and 20, respectively), and $\varphi: \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ is a simple closed convex function whose effective domain has nonempty relative interior. We assume that there exists at least one optimal solution $y^* \in \mathbb{E}$ for (A41). Moreover, for the sake of brevity, we suppose that the constant L in H1 is known. Thus, to approximately solve (A41), we may use the following adaptation of Algorithm A:

Algorithm B. Proximal Gradient Method

Step 0. Choose $y_0 \in \mathbb{E}$ and set $k := 0$.

Step 1. Compute

$$y_{k+1} = \arg \min_{x \in \mathbb{E}} \{ \langle \nabla g(y_k), x - y_k \rangle + 2L\beta_d(y_k, x) + \varphi(x) \}. \quad (\text{A42})$$

Step 2. Set $k := k + 1$ and go to Step 1.

Algorithm B can be viewed as a particular instance of the NoLips Algorithm in [2]. The next lemma gives a lower bound on the functional decrease in terms of the Bregman distance. It corresponds to Lemma 4.1 in [4]. We give its proof here for completeness.

Lemma A.7: *Suppose that H1 and H4 hold and let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm B. Then, for all $k \geq 0$,*

$$\tilde{g}(y_k) - \tilde{g}(y_{k+1}) \geq L\beta_d(y_k, y_{k+1}). \quad (\text{A43})$$

Proof: In view of (A42), we have

$$\langle \nabla g(y_k), y_{k+1} - y_k \rangle + 2L\beta_d(y_k, y_{k+1}) + \varphi(y_{k+1}) \leq \varphi(y_k).$$

Then,

$$\langle \nabla g(y_k), y_{k+1} - y_k \rangle \leq -2L\beta_d(y_k, y_{k+1}) - \varphi(y_{k+1}) + \varphi(y_k). \quad (\text{A44})$$

Now, combining H1 and (A44), we obtain

$$\begin{aligned} g(y_{k+1}) &\leq g(y_k) + \langle \nabla g(y_k), y_{k+1} - y_k \rangle + L\beta_d(y_k, y_{k+1}) \\ &\leq g(y_k) - 2L\beta_d(y_k, y_{k+1}) - \varphi(y_{k+1}) + \varphi(y_k) + L\beta_d(y_k, y_{k+1}). \end{aligned}$$

Therefore,

$$\tilde{g}(y_{k+1}) \leq \tilde{g}(y_k) - L\beta_d(y_k, y_{k+1}),$$

which gives (A43). ■

The next lemma gives a lower bound on the functional decrease in terms of the norm of a certain subgradient of $\tilde{g}(\cdot)$.

Lemma A.8: *Suppose that H1–H4 hold and let $\{y_k\}_{k \geq 0}$ be generated by Algorithm B. Then, for all $k \geq 0$,*

$$u(y_{k+1}) \equiv \nabla g(y_{k+1}) - \nabla g(y_k) + 2L[\nabla d(y_k) - \nabla d(y_{k+1})] \in \partial \tilde{g}(y_{k+1}), \quad (\text{A45})$$

and

$$\tilde{g}(y_k) - \tilde{g}(y_{k+1}) \geq \frac{\sigma_q}{qL^{q-1}(3N)^q} \|u(y_{k+1})\|_*^q, \quad (\text{A46})$$

where σ_q and N are specified in H2 and H3 (see page 18), respectively.

Proof: By H2 and Lemma A.7, for all k , we have

$$\tilde{g}(y_k) - \tilde{g}(y_{k+1}) \geq L\beta_d(y_k, y_{k+1}) \geq \frac{L\sigma_q}{q} \|y_k - y_{k+1}\|^q. \quad (\text{A47})$$

By the definition of y_{k+1} , this point satisfies the first-order optimality condition:

$$0 \in \{\nabla g(y_k) + 2L[\nabla d(y_{k+1}) - \nabla d(y_k)]\} + \partial \varphi(y_{k+1}).$$

Since $\text{ri}(\text{dom } \varphi) \neq \emptyset$, it follows that (A45) is true.

On the other hand, by H1, H4 and Proposition 1.1 in [14], we have

$$0 \leq \mu \nabla^2 d(y) \leq \nabla^2 g(y) \leq L \nabla^2 d(y), \quad \forall y \in \mathbb{E}.$$

Consequently,

$$\|\nabla^2 g(y)\| \leq L \|\nabla^2 d(y)\|, \quad \forall y \in \mathbb{E}. \quad (\text{A48})$$

Thus, in view of H3 and (A48), it follows from the mean value theorem that ∇d and ∇g are Lipschitz continuous on $\text{co}(\mathcal{L}(y_0))$ with constants N and LN , respectively. From (A47), we see that $\{\tilde{g}(y_k)\}_{k \geq 0}$ is nonincreasing, and so $\{y_k\} \subset \mathcal{L}(y_0)$. Therefore,

$$\begin{aligned} \|u(y_{k+1})\|_* &\leq \|\nabla g(y_{k+1}) - \nabla g(y_k)\|_* + 2L\|\nabla d(y_k) - \nabla d(y_{k+1})\|_* \\ &\leq (LN + 2LN) \|y_k - y_{k+1}\|, \end{aligned}$$

that is,

$$\|y_k - y_{k+1}\| \geq \frac{1}{3LN} \|u(y_{k+1})\|_*. \quad (\text{A49})$$

Combining (A46) and (A48), we obtain

$$\begin{aligned} \tilde{g}(y_k) - \tilde{g}(y_{k+1}) &\geq \frac{L\sigma_q}{q} \frac{1}{(3LN)^q} \|u(y_{k+1})\|_*^q \\ &= \frac{\sigma_q}{qL^{q-1}(3N)^q} \|u(y_{k+1})\|_*^q, \end{aligned}$$

which is (A46). ■

Theorem A.9: *Suppose that H1–H4 hold and let $\{y_k\}_{k \geq 0}$ be generated by Algorithm B. Then, for all $k \geq 1$, we have*

$$\tilde{g}(y_k) - \tilde{g}(y^*) \leq \frac{\mu\beta_d(y_0, y^*)}{\left(1 + \frac{\mu}{2L-\mu}\right)^k - 1} \leq \frac{(2L - \mu)\beta_d(y_0, y^*)}{k}, \quad (\text{A50})$$

where, in case $\mu = 0$, the middle expression is defined by the limit as $\mu \rightarrow 0^+$.

Proof: By H1, for all $k \geq 1$, we have

$$g(y_k) \leq g(y_{k-1}) + \langle \nabla g(y_{k-1}), y_k - y_{k-1} \rangle + L\beta_d(y_{k-1}, y_k). \quad (\text{A51})$$

To obtain an upper bound for the inner product in (A51), let us apply Lemma A.3 with $h = d$ and

$$\phi(x) = \frac{1}{2L} [\langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + \varphi(x)].$$

In this case, $y^+ = y_k$ and, for $y = y_{k-1}$ we have

$$\phi(x) + \beta_d(y_{k-1}, x) \geq \phi(y_k) + \beta_d(y_{k-1}, y_k) + \beta_d(y_k, x), \quad \forall x \in \mathbb{E},$$

that is,

$$\begin{aligned} \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + \varphi(x) + 2L\beta_d(y_{k-1}, x) &\geq \langle \nabla g(y_{k-1}), y_k - y_{k-1} \rangle + \varphi(y_k) \\ &\quad + 2L\beta_d(y_{k-1}, y_k) + 2L\beta_d(y_k, x). \end{aligned}$$

This gives the upper bound

$$\begin{aligned} \langle \nabla g(y_{k-1}), y_k - y_{k-1} \rangle &\leq \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + \varphi(x) + 2L\beta_d(y_{k-1}, x) \\ &\quad - \varphi(y_k) - 2L\beta_d(y_{k-1}, y_k) - 2L\beta_d(y_k, x). \end{aligned} \quad (\text{A52})$$

Combining (A51) and (A52), we obtain

$$\begin{aligned} g(y_k) &\leq g(y_{k-1}) + \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + \varphi(x) + 2L\beta_d(y_{k-1}, x) \\ &\quad - \varphi(y_k) - 2L\beta_d(y_{k-1}, y_k) - 2L\beta_d(y_k, x) + L\beta_d(y_{k-1}, y_k) \\ \tilde{g}(y_k) &\leq g(y_{k-1}) + \langle \nabla g(y_{k-1}), x - y_{k-1} \rangle + \varphi(x) + 2L\beta_d(y_{k-1}, x) - 2L\beta_d(y_k, x). \end{aligned} \quad (\text{A53})$$

Combining (A53) and (A23), we get

$$\begin{aligned} \tilde{g}(y_k) &\leq g(y_{k-1}) + g(x) - g(y_{k-1}) - \mu\beta_d(y_{k-1}, x) + \varphi(x) + 2L\beta_d(y_{k-1}, x) - 2L\beta_d(y_k, x) \\ &= \tilde{g}(x) + (2L - \mu)\beta_d(y_{k-1}, x) - 2L\beta_d(y_k, x). \end{aligned}$$

Substituting $x = y^*$, it follows that

$$\tilde{g}(y_k) \leq \tilde{g}(y^*) + (M - \mu)\beta_d(y_{k-1}, y^*) - M\beta_d(y_k, y^*),$$

where $M = 2L$. Then, the rest of the proof follows exactly as in the proof of Theorem A.4 (from inequality (A27)). ■

Corollary A.10: *Suppose that H1–H3 hold and let $\{y_k\}_{k \geq 0}$ be a sequence generated by Algorithm B. Additionally, assume that H4 holds with $\mu > 0$ and let $u(y_k) \in \partial \tilde{g}(y_k)$ be defined in (A45) for $k \geq 1$. Given $\epsilon \in (0, 1)$, if $\|u(y_{T+1})\|_* > \epsilon$, then*

$$T \leq \left[\log_2 \left(\frac{2L}{2L - \mu} \right) \right]^{-1} [C + q] \log_2(\epsilon^{-1}),$$

where

$$C = \log_2 \left(\frac{2q(2L)^{q-1} N \mu \beta_d(y_0, y^*)}{\sigma_q} \right).$$

Proof: By Lemma A.8 and Theorem A.9, it follows as in the proof of Corollary A.6. ■