2020/07

DP

Yurii Nesterov

Superfast second-order methods for unconstrained convex optimization

CORE
CENTER FOR OPERATIONS RESEARCH AND ECONOMETRICS

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: immaq-library@uclouvain.be

https://uclouvain.be/en/research-institutes/

lidam/core/discussion-papers.html

CORE DISCUSSION PAPER

2020/07

Superfast second-order methods for Unconstrained Convex Optimization

Yurii Nesterov *

February 4, 2020

Abstract

In this paper, we present new second-order methods with converge rate $O\left(k^{-4}\right)$, where k is the iteration counter. This is faster that the existing lower bound for this type of schemes [1, 2], which is $O\left(k^{-7/2}\right)$. Our progress can be explained by a finer specification of the problem class. The main idea of this approach consists in implementation of the third-order scheme from [15] using the second-order oracle. At each iteration of our method, we solve a nontrivial auxiliary problem by a linearly convergent scheme based on the relative non-degeneracy condition [3, 10]. During this process, the Hessian of the objective function is computed once, and the gradient is computed $O\left(\ln \frac{1}{\epsilon}\right)$ times, where ϵ is the desired accuracy of the solution for our problem.

Keywords: Convex Optimization, tensor methods, lower complexity bounds, second-order methods

*Center for Operations Research and Econometrics (CORE) and Department of Mathematical Engineering (INMA) of Catholic University of Louvain (UCL). E-mail: Yurii.Neterov@uclouvain.be.
Research results presented in this paper are obtained in the framework of Advanced Grant 788368 of the European Research Council.

1 Introduction

Motivation. In the last years, the theory of high-order methods in Convex Optimization was developed seemingly up to its natural limits. After discovering the simple fact that the auxiliary problem in tensor methods can be posed as a problem of minimizing a convex multivariate polynomial [15], very soon the performance of these methods was increased up to the maximal limits [5, 7, 9], given by the theoretical lower complexity bounds [1, 2].

It is interesting that the first accelerated tensor methods were analyzed in the unpublished paper [4], where the author did not express any hope for their practical implementations in the future. In [4] and [15], it was shown that the p-th order methods can accelerate up to the level $O\left(k^{-(p+1)}\right)$, where k in the iterations counter. The main advantage of the theory in [15] is that it corresponds to the methods with convex polynomial subproblems.

However, the fastest tensor methods [5, 7, 9] are based on the trick discovered in [11] for the second-order methods. It allows to increase the rate of convergence of tensor methods up to the level $O\left(k^{-(3p+1)/2}\right)$, which matches the lower complexity bounds for functions with Lipschitz-continuous pth derivative. Thus, for example, the best possible rate of convergence of the second-order methods on the corresponding problem class is of the order $O\left(k^{-7/2}\right)$.

Unfortunately, this advanced technique requires finding at each iteration a root of a univariate nonlinear non-monotone equation defined by inverse Hessians of the objective function. Hence, from the practical point of view, the methods proposed in [15] remain the most attractive.

The developments of this paper are based on one simple observation. In [15] it was shown that the accelerated tensor method of degree three with the rate of convergence $O\left(k^{-4}\right)$ can be implemented by using at each iteration a simple gradient method based on the *relative non-degeneracy* condition [3, 10]. This auxiliary method has to minimize an augmented Taylor polynomial of degree three, computed at the current test point $x \in \mathbb{R}^n$:

$$\langle \nabla f(x), h \rangle + \tfrac{1}{2}\langle \nabla^2 f(x)h, h\rangle + \tfrac{1}{6} D^3 f(x)[h]^3 + \tfrac{H}{24}\|h\|_2^4 \quad \to \quad \min_{h\in\mathbb{R}^n}.$$

At each iteration of this linearly convergent scheme, we need to compute the gradient of the auxiliary objective function in h. The only non-trivial part of this gradient comes from the gradient of the third derivative. This is the vector $D^3 f(x)[h]^2 \in \mathbb{R}^n$. It is the *only place* where we need the third-order information. However, it is well known that

$$D^3 f(x)[h]^2 \quad = \quad \lim_{\tau\to 0} \tfrac{1}{\tau^2}[\nabla f(x+\tau h) + \nabla f(x-\tau h) - 2\nabla f(x)].$$

In other words, the vector $D^3 f(x)[h]^2$ can be approximated with any accuracy by the *first-order* information. This means that we have a chance to implement the third-order method with the convergence rate $O\left(k^{-4}\right)$ using only the second-order information.

So, formally our method will be of the order two. However, it will have the rate of convergence, which is higher than the formal lower bound $O\left(k^{-7/2}\right)$ for the second-order schemes. Of course, the reason for this is that it will work with the problem class initially reserved for the third-order methods. However, interestingly enough, our method will demonstrate on this class *the same* rate of convergence as the third-order schemes.

In order to implement our hint into rigorous statements, we need to introduce in the constructions of Section 5 in [15] some modifications related to the inexactness of the available information. This is the subject of the remaining sections of this paper.

Contents. The paper is organized as follows. In Section 2. We introduce a convenient definition of the acceptable neighborhood of the exact tensor step. It differs from the previous ones (e.g. [6, 8, 16]) since for its verification it is necessary to call oracle of the main objective function. However, we will see that it significantly simplifies the overall complexity analysis. We prove that every point from this neighborhood ensures a good decrease of the objective functions, which is sufficient for implementing the Basic Tensor Method and its accelerated version without spoiling their rates of convergence.

In Section 3 we analyze the rate of convergence of the gradient method based on the relative smoothness condition [3, 10], under assumption that the gradient of the objective function is computed with a small absolute error. We need this analysis in order to replace the exact value of the third derivative along two vectors by a finite difference of the gradients. We show that the perturbed method converges linearly to a small neighborhood of the exact solution.

Finally, in Section 4 we put all our results together in order to justify a second-order implementation of the accelerated third-order tensor method. The rate of convergence of the resulting algorithm is of the order $O\left(k^{-4}\right)$, where k is the iteration counter. At each iteration we compute the Hessian once and the gradient is computed $O\left(\ln \frac{1}{\epsilon}\right)$ times, where ϵ is the desired accuracy of the solution of the main problem. Recall that this rate of convergence is impossible for the second-order schemes working with the functions with Lipschitz continuous third derivative (see [1, 2]). However, our problem class is smaller (see Lemma 4).

We conclude the paper with Section 6, where we discuss our results and directions for the future research.

Notation and generalities. In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space, and by $\mathbb{E}^*$ its dual spaced composed by linear functions on $\mathbb{E}$. For such a function $s \in \mathbb{E}^*$, we denote by $\langle s, x\rangle$ its value at $x \in \mathbb{E}$.

If it is not mentioned explicitly, we measure distances in $\mathbb{E}$ and $\mathbb{E}^*$ in a Euclidean norm. For that, using a self-adjoint positive-definite operator $B : \mathbb{E} \to \mathbb{E}^*$ (notation $B = B^* \succ 0$), we define

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

Sometimes, in the formulas involving products of linear operators, it will be convenient to treat $x \in \mathbb{E}$ as a linear operator from $\mathbb{R}$ to $\mathbb{E}$, and x^* as a linear operator from $\mathbb{E}^*$ to $\mathbb{R}$. In this case, xx^* is a linear operator from $\mathbb{E}^*$ to $\mathbb{E}$, acting as follows:

$$(xx^*)g \;=\; \langle g, x\rangle x \;\in\; \mathbb{E}, \quad g \in \mathbb{E}^*.$$

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

$$\nabla f(x) \;\in\; \mathbb{E}^*, \quad \nabla^2 f(x)h \;\in\; \mathbb{E}^*, \quad x \in \mathrm{dom}\, f,\; h \in \mathbb{E}.$$

In what follows, we often work with directional derivatives. For $p \geq 1$, denote by

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

the directional derivative of function f at x along directions $h_i \in \mathbb{E}$, $i = 1, \dots, p$. Note that $D^p f(x)[\cdot]$ is a *symmetric p-linear form*. Its *norm* is defined in a standard way:

$$\| D^p f(x) \| \;=\; \max_{h_1,\dots,h_p} \left\{ \left| D^p f(x)[h_1, \dots, h_p] \right| :\; \| h_i \| \leq 1,\, i = 1, \dots, p \right\}. \tag{1.1}$$

In terms of our previous notation, for any $x \in \mathrm{dom}\, f$ and $h_1, h_2 \in \mathbb{E}$, we have

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

Thus, for the Hessian, our definition corresponds to the *spectral norm* of self-adjoint linear operator (the maximal module of all eigenvalues computed with respect to operator B).

If all directions $h_1, \dots, h_p$ are the same, we apply notation

$$D^p f(x)[h]^p, \quad h \in \mathbb{E}.$$

Then, Taylor approximation of function $f(\cdot)$ at $x \in \mathrm{dom}\, f$ can be written as follows:

$$f(y) \;=\; \Omega_{x,p}(y) + o(\| y - x \|^p), \quad y \in \mathrm{dom}\, f,$$

$$\Omega_{x,p}(y) \;\stackrel{\mathrm{def}}{=}\; f(x) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(x)[y - x]^k, \quad y \in \mathbb{E}.$$

Note that, in general, we have (see, for example, Appendix 1 in [12])

$$\| D^p f(x) \| \;=\; \max_{h} \left\{ \left| D^p f(x)[h]^p \right| :\; \| h \| \leq 1 \right\}. \tag{1.2}$$

Similarly, since for $x, y \in \mathrm{dom}\, f$ being fixed, the form $D^p f(x)[\cdot, \dots, \cdot] - D^p f(y)[\cdot, \dots, \cdot]$ is p-linear and symmetric, we also have

$$\| D^p f(x) - D^p f(y) \| \;=\; \max_{h} \left\{ \left| D^p f(x)[h]^p - D^p f(y)[h]^p \right| :\; \| h \| \leq 1 \right\}. \tag{1.3}$$

In this paper, we consider functions from the problem classes $\mathcal{F}_p$, which are convex and p times differentiable on $\mathbb{E}$. Denote by L_p its uniform bound for the Lipschitz constant of their pth derivative:

$$\| D^p f(x) - D^p f(y) \| \;\leq\; L_p \| x - y \|, \quad x, y \in \mathrm{dom}\, f, \quad p \geq 1. \tag{1.4}$$

If an ambiguity can arise, we use notation $L_p(f)$. Sometimes it is more convenient to work with uniform bounds on the derivatives:

$$M_p(f) \;=\; \sup_{x \in \mathrm{dom}\, f} \| D^p f(x) \|. \tag{1.5}$$

If both values are well defined, we suppose that $L_p(f) = M_{p+1}(f)$, $p \geq 1$.

Let $F(\cdot)$ be a sufficiently smooth vector function, $F : \mathrm{dom}\, F \to \mathbb{E}_2$. Then, by the well-known Taylor formula, we have

$$\begin{array}{rl} F(y) \; - & F(x) - \sum\limits_{k=1}^{p} \frac{1}{k!} D^k F(x)[y-x]^k \\ = & \frac{1}{p!} \int\limits_{1}^{1} (1-\tau)^p D^{p+1} F(x+\tau(y-x))[y-x]^{p+1} d\tau, \quad x, y \in \mathrm{dom}\, F. \end{array} \tag{1.6}$$

Hence, we can bound the residual between the function value and its Taylor approximation as follows:

$$|f(y) - \Omega_{x,p}(y)| \quad \leq \quad \frac{L_p}{(p+1)!} \|y-x\|^{p+1}, \quad x, y \in \mathrm{dom}\, f. \tag{1.7}$$

By the same reason, for functions $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$, we get the following guarantees:

$$\|\nabla f(y) - \nabla \Omega_{x,p}(y)\|_* \quad \leq \quad \frac{L_p}{p!} \|y-x\|^p, \tag{1.8}$$

$$\|\nabla^2 f(y) - \nabla^2 \Omega_{x,p}(y)\| \quad \leq \quad \frac{L_p}{(p-1)!} \|y-x\|^{p-1}, \tag{1.9}$$

which are valid for all $x, y \in \mathrm{dom}\, f$.

Finally, for simplifying the long expressions we often use the trivial inequality

$$\left(a^{1/p} + b^{1/p}\right)^p \quad \leq \quad 2^{p-1}(a+b), \tag{1.10}$$

which is valid for all $a, b \geq 0$ and $p \geq 1$.

2 Tensor methods with inexact iteration

Consider the following unconstrained optimization problem:

$$\min_{x \in \mathbb{E}} \; f(x), \tag{2.1}$$

where $f(\cdot)$ is a smooth convex function with Lipschitz-continuous derivative of order $p \geq 1$:

$$\|D^p f(x) - D^p f(y)\| \quad \leq \quad L_p \|x-y\|, \quad x, y \in \mathbb{E}. \tag{2.2}$$

In this section, we work only with Euclidean norms.

We are going to solve problem (2.1) by tensor methods. Their performance crucially depends on ability to achieve a significant improvement in the objective function at the current test point.

Definition 1 *We say that point $T \in \mathbb{E}$ ensures pth-order improvement of some point $x \in \mathbb{E}$ with factor $c > 0$ if it satisfies the following inequality:*

$$\langle \nabla f(T), x - T \rangle \quad \geq \quad c \|\nabla f(T)\|_*^{\frac{p+1}{p}}. \tag{2.3}$$

This terminology has the following justification. Consider the *augmented Taylor polynomial* of degree $p \geq 1$:

$$\hat{\Omega}_{x,p,H}(y) \stackrel{\rm def}{=} \Omega_{x,p}(y) + \frac{H}{(p+1)!}\|y-x\|^{p+1}, \quad y \in \mathbb{E}.$$

In view of (1.7), for $H \geq L_p$, this function provides us with a uniform upper estimate for the objective. Moreover, for $H \geq pL_p$ this function is convex (see Theorem 1 in [15]).

We are going to generate new test point T as a close approximation to the minimum of function $\hat{\Omega}_{x,p,H}(\cdot)$. Namely, we are interested in points from the following nested neighborhoods:

$$\mathcal{N}^{\gamma}_{p,H}(x) = \{T \in \mathbb{E} : \|\nabla\hat{\Omega}_{x,p,H}(T)\|_* \leq \gamma\|\nabla f(T)\|_*\}, \tag{2.4}$$

where $\gamma \in [0,1)$ is an accuracy parameter. The smallest set $\mathcal{N}^0_{p,H}(x)$ contains only the exact minimizers of the augmented Taylor polynomial. Note that $\hat{\Omega}_{x,p,H}(x) = \nabla f(x)$. Therefore, if $\nabla f(x) \neq 0$, then $x \notin \mathcal{N}^{\gamma}_{p,H}(x)$ for any $\gamma \in [0,1)$.

These neighborhoods are important by the following reason.

Theorem 1 *Let $x \in \mathbb{E}$ and parameters γ and H satisfy the following condition:*

$$\gamma + \frac{L_p}{H} \leq \frac{1}{p}. \tag{2.5}$$

Then any point T from the set $\mathcal{N}^{\gamma}_{p,H}(x)$ ensures a pth-order improvement of x with factor

$$c_{\gamma,H}(p) \stackrel{\rm def}{=} \left[\frac{(1-\gamma)p!}{L_p+H}\right]^{\frac{1}{p}}. \tag{2.6}$$

Consequently, we have

$$f(x) - f(T) \geq c_{\gamma,H}(p)\|\nabla f(T)\|_*^{\frac{p+1}{p}}. \tag{2.7}$$

Proof:
Let $T \in \mathcal{N}^{\gamma}_{p,H}(x)$. Denote by $r = \|x-T\|$. Then

$$\begin{aligned}
&\|\nabla f(T)\|_*^2 + 2\frac{H}{p!}r^{p-1}\langle\nabla f(T), T-x\rangle + \left(\frac{H}{p!}\right)^2 r^{2p} \\
&= \|\nabla f(T) + \frac{H}{p!}r^{p-1}B(T-x)\|_*^2 \\
&= \|\nabla f(T) - \nabla\Omega_{x,p}(T) + \nabla\hat{\Omega}_{x,p,H}(T)\|_*^2 \\
&\stackrel{(1.8)}{\leq} \left(\frac{L_p}{p!}r^p + \gamma\|\nabla f(T)\|_*\right)^2.
\end{aligned}$$

Therefore,

$$2\frac{H}{p!}r^{p-1}\langle\nabla f(T), x-T\rangle \geq (1-\gamma^2)\|\nabla f(T)\|_*^2 + \frac{H^2-L_p^2}{(p!)^2}r^{2p} - 2\gamma\frac{L_p}{p!}r^p\|\nabla f(T)\|_*.$$

In other words,

$$\langle\nabla f(T), x-T\rangle \geq \frac{(1-\gamma^2)p!}{2Hr^{p-1}}\|\nabla f(T)\|_*^2 + \frac{H^2-L_p^2}{2Hp!}r^{p+1} - \gamma r\frac{L_p}{H}\|\nabla f(T)\|_* \stackrel{\rm def}{=} \varkappa(r).$$

The right-hand side of this inequality is a convex function in r. Its derivative in r is as follows:

$$\varkappa'(r) \;=\; -\tfrac{(1-\gamma^2)(p-1)p!}{2Hr^p}\|\nabla f(T)\|_*^2 + \tfrac{(p+1)(H^2-L_p^2)}{2Hp!}r^p - \gamma\tfrac{L_p}{H}\|\nabla f(T)\|_*.$$

Note that

$$\begin{aligned}\|\nabla f(T)\|_* \;&=\; \|\nabla f(T) - \nabla\Omega_{x,p}(T) + \nabla\hat{\Omega}_{x,p,H}(T) - \tfrac{H}{p!}r^{p-1}B(T-x)\|_*\\ &\leq\; \tfrac{L_p}{p!}r^p + \gamma\|\nabla f(T)\|_* + \tfrac{H}{p!}r^p.\end{aligned}$$

Thus, $r \geq r_* \stackrel{\mathrm{def}}{=} \left[\frac{(1-\gamma)p!\,\|\nabla f(T)\|_*}{L_p+H}\right]^{\frac{1}{p}}$. At the same time,

$$\begin{aligned}\varkappa'(r_*) \;&=\; -\tfrac{(1-\gamma^2)(p-1)p!\|\nabla f(T)\|_*^2\cdot}{2H}\cdot\tfrac{L_p+H}{(1-\gamma)p!\,\|\nabla f(T)\|_*}\\ &\quad+\tfrac{(p+1)(H^2-L_p^2)}{2Hp!}\cdot\tfrac{(1-\gamma)p!\,\|\nabla f(T)\|_*}{L_p+H} - \gamma\tfrac{L_p}{H}\|\nabla f(T)\|_*\\ &=\; \|\nabla f(T)\|_*\left[-\tfrac{(1+\gamma)(p-1)}{2}\left(1+\tfrac{L_p}{H}\right) + \tfrac{(p+1)(1-\gamma)}{2}\left(1-\tfrac{L_p}{H}\right) - \gamma\tfrac{L_p}{H}\right]\\ &=\; \|\nabla f(T)\|_*\left[1-p\gamma-p\tfrac{L_p}{H}\right] \;\stackrel{(2.5)}{\geq}\; 0.\end{aligned}$$

Therefore,

$$\begin{aligned}\langle\nabla f(T), x-T\rangle \;&\geq\; \varkappa(r_*) \;=\; r_*\left[\tfrac{(1-\gamma^2)p!}{2Hr_*^p}\|\nabla f(T)\|_*^2 + \tfrac{H^2-L_p^2}{2Hp!}r_*^p - \gamma\tfrac{L_p}{H}\|\nabla f(T)\|_*\right]\\ &=\; r_*\|\nabla f(T)\|_*\left[\tfrac{(1-\gamma^2)p!}{2H}\cdot\tfrac{L_p+H}{(1-\gamma)p!} + \tfrac{H^2-L_p^2}{2Hp!}\cdot\tfrac{(1-\gamma)p!}{L_p+H} - \gamma\tfrac{L_p}{H}\right]\\ &=\; r_*\|\nabla f(T)\|_*.\end{aligned}$$

Inequality (2.7) is valid since our function is convex:

$$f(x) \;\geq\; f(T) + \langle\nabla f(T), x-T\rangle. \qquad \square$$

Thus, we have proved that the pth-order improvement at point $x \in \mathbb{E}$ can be ensured by *inexact minimizers* of the augmented Taylor polynomials of degree $p \geq 1$. Let us present the efficiency estimates for the corresponding optimization schemes.

From now on, let us assume that the constant L_p is known. For the sake of notation, we fix the following values of the parameters:

$$\gamma \;=\; \tfrac{1}{2p}, \quad H \;=\; 2pL_p. \tag{2.8}$$

Then we can use a shorter notation for the following objects:

$$\mathcal{N}_p(x) \;\stackrel{\mathrm{def}}{=}\; \mathcal{N}^{1/(2p)}_{p,2pL_p}(x), \quad c_p \;\stackrel{\mathrm{def}}{=}\; c_{1/(2p),2pL_p}(p) \;=\; \left[\tfrac{2p-1}{2p(2p+1)}\,\tfrac{p!}{L_p}\right]^{\frac{1}{p}}. \tag{2.9}$$

As a consequence of all these specifications, we have the following result.

Corollary 1 *For any $x \in \mathbb{E}$, all points from the neighborhood $\mathcal{N}_p(x)$ ensure the pth-order improvement of x with factor c_p.*

Let us start from the simplest Inexact Basic Tensor Method:

$$\boxed{x_{k+1} \;\; \in \;\; \mathcal{N}_p(x_k), \quad k \geq 0.} \tag{2.10}$$

Denote $R(x_0) = \max\limits_{y \in \mathbb{E}} \{ \| y - x^* \| : \; f(y) \leq f(x_0) \}$.

Theorem 2 *Let the sequence $\{x_k\}_{k \geq 0}$ be generated by method (2.10). Then for any $k \geq 1$ we have*

$$\begin{array}{rcl} f(x_k) - f^* & \leq & \left[\frac{p+1}{k} \left(\frac{1}{c_p} R^{\frac{p+1}{p}}(x_0) + (f(x_0) - f^*)^{1/p} \right) \right]^p \\ & \overset{(1.10)}{\leq} & \left(\frac{2(p+1)}{k} \right)^p \left[\frac{p(2p+1)}{(2p-1)p!} L_p R^{p+1}(x_0) + \frac{1}{2}(f(x_0) - f^*) \right]. \end{array} \tag{2.11}$$

Proof:
In view of inequality (2.7), we have $f(x_k) \leq f(x_0)$ for all $k \geq 0$. Therefore

$$\| x_k - x^* \| \;\; \leq \;\; R_0 \overset{\mathrm{def}}{=} R(x_0), \quad k \geq 0.$$

Consequently,

$$\begin{array}{rcl} f(x_k) - f(x_{k+1}) & \overset{(2.7)}{\geq} & c_p \| \nabla f(x_{k+1}) \|_*^{\frac{p+1}{p}} \;\geq\; c_p \left(\frac{\langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle}{R(x_0)} \right)^{\frac{p+1}{p}} \\ & \geq & c_p \left(\frac{f(x_{k+1}) - f^*}{R(x_0)} \right)^{\frac{p+1}{p}}. \end{array}$$

Denoting $\xi_k = \frac{c_p^p}{R_0^{p+1}}(f(x_k) - f^*)$, we get inequality $\xi_k - \xi_{k+1} \geq \xi_{k+1}^{\frac{p+1}{p}}$. Hence, in view of Lemma 11 in [16], we have

$$\xi_k \;\; \leq \;\; \frac{1}{k^p} \left[(p+1)(1 + \xi_0^{1/p}) \right]^p, \quad k \geq 1.$$

This is exactly the estimate (2.11). $\square$

Let us present now the convergence analysis for Inexact Accelerated Tensor Method. We need to choose the degree of the method and define the prox-function

$$d_{p+1}(x) \;\; = \;\; \frac{1}{p+1} \| x \|^{p+1}, \quad x \in \mathbb{E}.$$

This is a uniformly convex function of degree $p+1$:

$$d_{p+1}(y) \;\; \geq \;\; d_{p+1}(x) + \langle \nabla d_{p+1}(x), y - x \rangle + \frac{1}{p+1} \left(\frac{1}{2} \right)^{p-1} \| y - x \|^{p+1}, \quad x, y \in \mathbb{E} \tag{2.12}$$

(see, for example, Lemma 4.2.3 in [17]).

Define the sequence

$$A_k \;=\; 2\left(\frac{p+1}{2p}c_p\right)^p \left(\frac{k}{p+1}\right)^{p+1}, \quad a_{k+1} \stackrel{\mathrm{def}}{=} A_{k+1} - A_k, \quad k \geq 0. \tag{2.13}$$

Note that for all values $B_k = \left(\frac{k}{p+1}\right)^{p+1}$ with $k \geq 0$ we have

$$\frac{(B_{k+1}-B_k)^{\frac{p+1}{p}}}{B_{k+1}} \;=\; \left(\frac{k+1}{p+1} - \frac{k}{p+1}\left[\frac{k}{k+1}\right]^p\right)^{\frac{p+1}{p}} \;\leq\; \left(\frac{k+1}{p+1} - \frac{k}{p+1}\left[1 - \frac{p}{k+1}\right]\right)^{\frac{p+1}{p}} \;\leq\; 1.$$

Therefore, the elements of sequence $\{A_k\}_{k\geq 0}$ satisfy the following inequality:

$$a_{k+1}^{\frac{p+1}{p}} \;\leq\; 2^{1/p}\frac{p+1}{2p}c_p\, A_{k+1}, \quad k \geq 0. \tag{2.14}$$

Inexact pth-Order Accelerated Tensor Method (ATMI$_p$)
Initialization. Choose $x_0 \in \mathbb{E}$. Define coefficients A_k by (2.13) and function $\psi_0(x) = d_{p+1}(x - x_0)$
Iteration $k \geq 0$.   **1.** Compute $v_k = \arg\min\limits_{x\in\mathbb{E}} \psi_k(x)$ and choose $y_k = \frac{A_k}{A_{k+1}}x_k + \frac{a_k}{A_{k+1}}v_k$.   **2.** Compute $x_{k+1} \in \mathcal{N}_p(y_k)$ and update $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle].$

(2.15)

First of all, note that by induction it is easy to see that

$$\psi_k(x) \;\leq\; A_k f(x) + d_{p+1}(x - x_0), \quad x \in \mathbb{E}. \tag{2.16}$$

In particular, for $\psi_k^* \stackrel{\mathrm{def}}{=} \min\limits_{x\in\mathbb{E}} \psi_k(x)$ and all $x \in \mathbb{E}$, we have

$$A_k f(x) + d_{p+1}(x - x_0) \;\stackrel{(2.16)}{\geq}\; \psi_k(x) \;\stackrel{(2.12)}{\geq}\; \psi_k^* + \frac{1}{p+1}\left(\frac{1}{2}\right)^{p-1}\|x - v_k\|^{p+1}. \tag{2.17}$$

Let us prove by induction the following relation:

$$\psi_k^* \;\geq\; A_k f(x_k), \quad k \geq 0. \tag{2.18}$$

For $k = 0$, we have $\psi_0^* = 0$ and $A_0 = 0$. Hence, (2.18) is valid. Assume it is valid for some $k \geq 0$. Then

$$\begin{aligned}
\psi_{k+1}^* \;&=\; \min_{x\in\mathbb{E}}\left\{\psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\right\} \\
&\stackrel{(2.17)}{\geq} \min_{x\in\mathbb{E}}\left\{\psi_k^* + \frac{1}{p+1}\left(\frac{1}{2}\right)^{p-1}\|x - v_k\|^{p+1} + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\right\}.
\end{aligned}$$

Note that

$$
\begin{aligned}
& \psi_k^* + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\
\overset{(2.18)}{\geq} \quad & A_k f(x_k) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\
\geq \quad & A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), a_{k+1}(x - x_{k+1}) + A_k(x_k - x_{k+1})\rangle \\
= \quad & A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), a_{k+1}(x - v_k) + A_{k+1}(y_k - x_{k+1})\rangle.
\end{aligned}
$$

Further, for all $x \in \mathbb{E}$ we have

$$
\begin{aligned}
& \tfrac{1}{p+1}\left(\tfrac{1}{2}\right)^{p-1}\|x - v_k\|^{p+1} + a_{k+1}\langle \nabla f(x_{k+1}), x - v_k\rangle \\
\geq \quad & -\tfrac{p}{p+1} 2^{\frac{p-1}{p}} \left(a_{k+1}\|\nabla f(x_{k+1})\|_*\right)^{\frac{p+1}{p}}.
\end{aligned}
$$

Finally, since $x_{k+1} \in \mathcal{N}_p(y_k)$, by Corollary 1 we get

$$
\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \quad \geq \quad c_p \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}}.
$$

Putting all these inequalities together, we obtain

$$
\begin{aligned}
\psi_{k+1}^* \quad \geq \quad & A_{k+1} f(x_{k+1}) - \tfrac{p}{p+1} 2^{\frac{p-1}{p}} \left(a_{k+1}\|\nabla f(x_{k+1})\|_*\right)^{\frac{p+1}{p}} + A_{k+1} c_p \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}} \\
= \quad & A_{k+1} f(x_{k+1}) + \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}} \left(A_{k+1} c_p - \tfrac{p}{p+1} 2^{\frac{p-1}{p}} a_{k+1}^{\frac{p+1}{p}}\right) \\
\overset{(2.14)}{\geq} \quad & A_{k+1} f(x_{k+1}).
\end{aligned}
$$

Thus, we have proved the following theorem.

Theorem 3 *Let sequence $\{x_k\}_{k \geq 0}$ be generated by method (2.15). Then, for any $k \geq 1$, we have*

$$
f(x_k) - f^* \quad \leq \quad \tfrac{2p+1}{2(2p-1)p!} \left(\tfrac{2p}{k}\right)^{p+1} \cdot L_p \|x^* - x_0\|^{p+1}. \tag{2.19}
$$

Proof:
Indeed, in view of relations (2.16) and (2.18), we have

$$
\begin{aligned}
f(x_k) - f^* \quad \leq \quad & \tfrac{1}{A_k} d_{p+1}(x^* - x_0) \overset{(2.13)}{=} \tfrac{1}{2}\left(\tfrac{2p}{(p+1)c_p}\right)^p \left(\tfrac{p+1}{k}\right)^{p+1} \cdot \tfrac{1}{p+1}\|x^* - x_0\|^{p+1} \\
= \quad & \tfrac{1}{2}\left(\tfrac{2p}{c_p}\right)^p \left(\tfrac{1}{k}\right)^{p+1} \cdot \|x^* - x_0\|^{p+1} \quad = \quad \tfrac{(2p+1)L_p}{2(2p-1)p!}\left(\tfrac{2p}{k}\right)^{p+1} \cdot \|x^* - x_0\|^{p+1}.
\end{aligned}
$$

$\square$

3 Relative non-degeneracy and approximate gradients

In this section, we measure distances in $\mathbb{E}$ by general norms. Consider the following composite minimization problem:

$$\min_{x \in \mathrm{dom}\,\psi} \left\{ F(x) \stackrel{\mathrm{def}}{=} \varphi(x) + \psi(x) \right\}, \tag{3.1}$$

where the convex function $\varphi(\cdot)$ is differentiable, and $\psi(\cdot)$ is a simple closed convex function. The most important example of function $\psi(\cdot)$ is an indicator function for a closed convex set. Denote by x^* one of the optimal solutions of problem (3.1), and let $F^* = F(x^*)$.

We assume that $\varphi(\cdot)$ is non-degenerate with respect to some scaling function $d(\cdot)$:

$$\begin{array}{rcl} \mu_d(\varphi)\beta_d(x,y) & \leq & \beta_\varphi(x,y) \stackrel{\mathrm{def}}{=} \varphi(y) - \varphi(x) - \langle \nabla\varphi(x), y - x\rangle \\ & \leq & L_d(\varphi)\beta_d(x,y), \quad x, y \in \mathrm{dom}\,\psi, \end{array} \tag{3.2}$$

where $0 \leq \mu_d(\varphi) \leq L_d(\varphi)$. Denote by $\gamma_d(\varphi) = \frac{\mu_d(\varphi)}{L_d(\varphi)} \leq 1$ the *condition number* of function $\varphi(\cdot)$ with respect to the scaling function $d(\cdot)$. Sometimes it is more convenient to work with the second-order variant of the condition (3.2):

$$\mu_d(\varphi)\nabla^2 d(x) \ \preceq \ \nabla^2\varphi(x) \ \preceq \ L_d(\varphi)\nabla^2 d(x), \quad x \in \mathrm{dom}\,\psi. \tag{3.3}$$

We are going to solve problem (3.1) using an approximate gradient of the smooth part of the objective function. Namely, at each point $x \in \mathbb{E}$ we use a vector $g_\varphi(x)$ such that

$$\| g_\varphi(x) - \nabla\varphi(x) \|_* \ \leq \ \delta, \tag{3.4}$$

where $\delta \geq 0$ is an accuracy parameter.

Our first goal is to describe the influence of parameter δ onto the quality of the computed approximate solutions to problem (3.1). For this, we need to assume that function $d(\cdot)$ is *uniformly convex* of degree $p+1$ with $p \geq 1$:

$$\beta_d(x,y) \ \geq \ \tfrac{1}{p+1}\sigma_{p+1}(d)\|x-y\|^{p+1}, \quad x, y \in \mathrm{dom}\,\psi. \tag{3.5}$$

Consider the following Bregman-Distance Gradient Method (BDGM), working with inexact information.

$$\boxed{\begin{array}{l} \textbf{Choose } x_0 \in \mathbb{E}. \textbf{ For } k \geq 0 \textbf{ iterate:} \\[1em] x_{k+1} = \arg\min\limits_{x \in \mathrm{dom}\,\psi} \left\{ \psi(y) + \langle g_\varphi(x), y - x\rangle + 2L_d(\varphi)\beta_d(x,y) \right\}. \end{array}} \tag{3.6}$$

Lemma 1 *Let the approximate gradient $g_\varphi(x_k)$ satisfy the condition (3.4). Then for any $x \in \mathbb{E}$ we have*

$$\begin{array}{rcl} \beta_d(x_{k+1},x) & \leq & \left(1 - \frac{1}{4}\gamma_d(\varphi)\right)\beta_d(x_k,x) + \frac{1}{2L_d(\varphi)}[F(x) - F(x_{k+1})] \\ & & + \frac{2p}{p+1}\delta^{\frac{p+1}{p}} \left(\frac{(p+1)(2+\gamma_d(\varphi))}{\sigma_{p+1}(d)\,\gamma_d(\varphi)} \right)^{\frac{1}{p}}, \quad k \geq 0. \end{array} \tag{3.7}$$

Proof:
The first-order optimality condition for the optimization problem defining the point x_{k+1} is as follows:

$$\langle g_\varphi(x_k) + 2L_d(\varphi)(\nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1}\rangle + \psi(x) \quad \geq \quad \psi(x_{k+1}) \tag{3.8}$$

for all $x \in \mathrm{dom}\, \psi$. Therefore, denoting $r_k(x) = \beta_d(x_k, x)$, we have

$$\begin{aligned} r_{k+1}(x) - r_k(x) \quad &= \quad \Big(d(x) - d(x_{k+1}) - \langle \nabla d(x_{k+1}), x - x_{k+1}\rangle\Big) \\ &\quad\quad -\Big(d(x) - d(x_k) - \langle \nabla d(x_k), x - x_k\rangle\Big) \\ &= \quad d(x_k) - \langle \nabla d(x_k), x_k - x_{k+1}\rangle - d(x_{k+1}) \\ &\quad\quad +\langle \nabla d(x_k) - \nabla d(x_{k+1}), x - x_{k+1}\rangle \\ &\overset{(3.8)}{\leq} \quad -\beta_d(x_k, x_{k+1}) + \tfrac{1}{2L_d(\varphi)}\Big[\langle g_\varphi(x_k), x - x_{k+1}\rangle + \psi(x) - \psi(x_{k+1})\Big]. \end{aligned}$$

Note that $\langle g_\varphi(x_k), x - x_{k+1}\rangle = \langle g_\varphi(x_k) - \nabla\varphi(x_k), x - x_{k+1}\rangle + \langle \nabla\varphi(x_k), x - x_{k+1}\rangle$, and

$$\begin{aligned} \langle \nabla\varphi(x_k), x - x_{k+1}\rangle \quad &\overset{(3.2)}{\leq} \quad \langle \nabla\varphi(x_k), x_k - x_{k+1}\rangle + \varphi(x) - \varphi(x_k) - \mu_d(\varphi) d(x_k, x) \\ &\overset{(3.2)}{\leq} \quad L_d(\varphi) d(x_k, x_{k+1}) + \varphi(x) - \varphi(x_{k+1}) - \mu_d(\varphi) d(x_k, x). \end{aligned}$$

Hence,

$$\begin{aligned} & r_{k+1}(x) - r_k(x) + \tfrac{1}{2L_d(\varphi)}[F(x_{k+1}) - F(x)] \\ \leq \quad & \langle g_\varphi(x_k) - \nabla\varphi(x_k), x - x_{k+1}\rangle - \tfrac{1}{2}\beta_d(x_k, x_{k+1}) - \tfrac{1}{2}\gamma_d(\varphi) d(x_k, x) \\ \overset{(3.5)}{\leq} \quad & -\tfrac{1}{4}\gamma_d(\varphi) r_k(x) + \langle g_\varphi(x_k) - \nabla\varphi(x_k), x - x_{k+1}\rangle \\ & -\tfrac{\sigma_{p+1}(d)}{2(p+1)}\Big(\|x_k - x_{k+1}\|^{p+1} + \tfrac{1}{2}\gamma_d(\varphi)\|x_k - x\|^{p+1}\Big). \end{aligned}$$

Since $\|x\| = \|-x\|$ for all x in $\mathbb{E}$, the maximum in x_k of the expression in brackets is attained at some $x_k = (1-\alpha)x_{k+1} + \alpha x$ with $\alpha \in (0,1)$. On the other hand, the minimum of the function

$$\alpha^{p+1} + \tfrac{1}{2}\gamma_d(\varphi)(1-\alpha)^{p+1}, \quad \alpha \in [0,1],$$

is attained at $\bar\alpha = \frac{\beta}{1+\beta}$ with $\beta = \left(\frac{1}{2}\gamma_d(\varphi)\right)^{\frac{1}{p}}$. This is

$$\bar\alpha^{p+1} + \beta^p(1-\bar\alpha)^{p+1} \quad = \quad \bar\alpha\tfrac{\beta^p}{(1+\beta)^p} + \tfrac{\beta^p}{(1+\beta)^{p+1}} \quad = \quad \tfrac{\beta^p}{(1+\beta)^p} \quad \overset{(1.10)}{\geq} \quad \tfrac{\gamma_d(\varphi)}{2^{p-1}(2+\gamma_d(\varphi))}.$$

Thus,

$$
\begin{aligned}
& r_{k+1}(x) - (1 - \tfrac{1}{4}\gamma_d(\varphi)) r_k(x) + \tfrac{1}{2L_d(\varphi)}[F(x_{k+1}) - F(x)] \\
\leq\ & \langle g_\varphi(x_k) - \nabla\varphi(x_k), x - x_{k+1}\rangle - \tfrac{\sigma_{p+1}(d)\,\gamma_d(\varphi)}{2^p(p+1)(2+\gamma_d(\varphi))}\|x - x_{k+1}\|^{p+1} \\
\stackrel{(3.4)}{\leq}\ & \tfrac{2p}{p+1}\delta^{\frac{p+1}{p}} \left(\tfrac{(p+1)(2+\gamma_d(\varphi))}{\sigma_{p+1}(d)\,\gamma_d(\varphi)}\right)^{\frac{1}{p}}. \qquad \Box
\end{aligned}
$$

Applying inequality (3.7) with $x = x^*$ recursively to all $k = 0, \dots, T-1$, we get the following relation:

$$
\begin{aligned}
& \beta_d(x_T, x^*) + \tfrac{1}{2L_d(\varphi)} \sum_{k=0}^{T-1} (1-\gamma)^{T-k-1}[F(x_{k+1}) - F(x^*)] \\
& \leq\ (1-\gamma)^T \beta_d(x_0, x^*) + S_T \hat{\delta},
\end{aligned}
\tag{3.9}
$$

where $\gamma = \frac{1}{4}\gamma_d(\varphi)$, $\hat{\delta} = \frac{2p}{p+1}\delta^{\frac{p+1}{p}} \left(\frac{(p+1)(2+\gamma_d(\varphi))}{\sigma_{p+1}(d)\,\gamma_d(\varphi)}\right)^{\frac{1}{p}}$, and

$$
S_T \;=\; \sum_{k=0}^{T-1} (1-\gamma)^{T-k-1} \;=\; \tfrac{1}{\gamma}\Big(1 - (1-\gamma)^T\Big).
$$

Thus, denoting $F_T^* = \min\limits_{0 \leq k \leq T} F(x_k)$, we get the following bound:

$$
F_T^* - F^* \;\stackrel{(3.9)}{\leq}\; \tfrac{2\gamma(1-\gamma)^T}{1-(1-\gamma)^T} L_d(\varphi)\beta(x_0, x^*) + 2\hat{\delta}L_d(\varphi), \quad T \geq 1. \tag{3.10}
$$

Note that $\lim\limits_{\gamma \downarrow 0} \frac{\gamma(1-\gamma)^T}{1-(1-\gamma)^T} = \frac{1}{T}$. Hence, for $\mu_d(\varphi) = 0$ we get the following rate of convergence:

$$
F_T^* - F^* \;\stackrel{(3.9)}{\leq}\; 2L_d(\varphi)\left(\tfrac{1}{T}\beta(x_0, x^*) + 2\hat{\delta}\right), \quad T \geq 1. \tag{3.11}
$$

In our main application, presented in Section 4, we need to generate points with small norm of the gradient. In order to achieve this goal with method (3.6), we need one more assumption on the scaling function $d(\cdot)$.

From now on, we consider the unconstrained minimization problems. This means that in (3.1) we have $\psi(x) = 0$ for all $x \in \mathbb{E}$.

Definition 2 *We call the scaling function $d(\cdot)$ norm-dominated on the set $S \subseteq \mathbb{E}$ by some function $\theta_S(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ if there exist a convex function $\theta_S(\cdot)$ with $\theta_S(0) = 0$ such that*

$$
\beta_d(x, y) \;\leq\; \theta_S(\|x - y\|) \tag{3.12}
$$

for all $x \in S$ and $y \in \mathbb{E}$.

Clearly, if function $d(\cdot)$ is norm-dominated by function $\theta_S(\cdot)$ and $\eta_S(\tau) \geq \theta_S(\tau)$ for all $\tau \geq 0$, then $d(\cdot)$ is also norm-dominated by function $\eta_S(\cdot)$.

Let us give an important example of a norm-dominated scaling function.

Lemma 2 *Function $d_4(\cdot)$ is norm-dominated on the Euclidean ball*

$$B_R \quad = \quad \{x \in \mathbb{E} : \; \|x\| \leq R\}$$

by the function

$$\theta_R(\tau) \quad = \quad \tfrac{1}{4}(\tau^2 + 2R\tau)^2 + \tfrac{1}{2}R^2\tau^2 \;\leq\; \tfrac{1}{2}\tau^4 + \tfrac{5}{2}R^2\tau^2, \quad \tau \geq 0. \tag{3.13}$$

Proof:
Let $x \in B_R$ and $y = x + h \in \mathbb{E}$. Then

$$\begin{aligned}
\beta_{d_4}(x,y) &= \tfrac{1}{4}\|y\|^4 - \tfrac{1}{4}\|x\|^4 - \|x\|^2\langle Bx, y-x\rangle \\
&= \tfrac{1}{4}[\|x\|^2 + 2\langle Bx, h\rangle + \|h\|^2]^2 - \tfrac{1}{4}\|x\|^4 - \|x\|^2\langle Bx, h\rangle \\
&= \tfrac{1}{4}[\|x\|^4 + 4\langle Bx, h\rangle^2 + \|h\|^4 + 4(\|x\|^2 + \|h\|^2)\langle Bx, h\rangle + 2\|x\|^2\|h\|^2] \\
&\quad -\tfrac{1}{4}\|x\|^4 - \|x\|^2\langle Bx, h\rangle \\
&= \tfrac{1}{4}(\|h\|^2 + 2\langle Bx, h\rangle)^2 + \tfrac{1}{2}\|x\|^2\|h\|^2.
\end{aligned}$$

Thus, we can take $\theta_R(\tau) = \frac{1}{4}(\tau^2 + 2R\tau)^2 + \frac{1}{2}R^2\tau^2$. □

Note that the statement of Lemma 2 can be extended onto all convex polynomial scaling functions.

The importance of norm-dominated scaling functions is related to the following fact.

Lemma 3 *Let scaling function $d(\cdot)$ be norm-dominated on the level set*

$$\mathcal{L}(\bar{x}) \quad = \quad \{x \in \mathbb{E} : \; \varphi(x) \leq \varphi(\bar{x})\}$$

by some function $\theta(\cdot)$. Then, for any $x \in \mathcal{L}_\varphi(\bar{x})$ we have:

$$\varphi(x) - \varphi(x^*) \quad \geq \quad L_d(\varphi)\, \theta^*\left(\tfrac{1}{L_d(\varphi)}\|\nabla\varphi(x)\|_*\right), \tag{3.14}$$

where $\theta^*(\tau) = \max\limits_{\lambda}[\lambda\tau - \theta(\tau)]$.

Proof:
Indeed, for any $x \in \mathcal{L}_\varphi(\bar{x})$ and $y \in \mathbb{E}$ we have

$$\begin{aligned}
\varphi(y) &\overset{(3.2)}{\leq} \varphi(x) + \langle\nabla\varphi(x), y-x\rangle + L_d(\varphi)\beta_d(x,y) \\
&\overset{(3.12)}{\leq} \varphi(x) + \langle\nabla\varphi(x), y-x\rangle + L_d(\varphi)\theta(\|y-x\|).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\varphi^* &= \min_{y\in\mathbb{E}} \varphi(y) \;\le\; \min_{y\in\mathbb{E}} \Big\{ \varphi(x) + \langle \nabla\varphi(x), y-x\rangle + L_d(\varphi)\theta(\|y-x\|) \Big\} \\
&= \min_{r\ge 0} \min_{y:\|y-x\|=r} \Big\{ \varphi(x) + \langle \nabla\varphi(x), y-x\rangle + L_d(\varphi)\theta(r) \Big\} \\
&= \varphi(x) + \min_{r\ge 0} \Big\{ -r\|\nabla\varphi(x)\|^* + L_d(\varphi)\theta(r) \Big\} \\
&= \varphi(x) - L_d(\varphi)\,\theta^*\left(\tfrac{1}{L_d(\varphi)} \|\nabla\varphi(x)\|_* \right). \qquad \Box
\end{aligned}$$

Thus, for norm-dominated scaling functions, the rate of convergence in function value can be transformed into the rate of decrease of the norm of the gradient of function $\varphi(\cdot)$. This feature is very important for practical implementations of Inexact Tensor Methods presented in Section 2. In the next section, we discuss in details how it works for inexact third-order methods.

4 Second-order implementations of the third-order methods

In this section, we are going to solve the unconstrained minimization problem

$$\min_{x\in\mathbb{E}} \; f(x), \tag{4.1}$$

where the objective function is convex and smooth, using the second-order implementations of the *third-order methods*. For the pure second-order methods, the standard assumption on the objective function in (4.1) is Lipschitz continuity of the second derivative (see, for example, [13, 14]). We are going to replace it by a stronger assumption, using the following fact.

Lemma 4 *Let constants $M_2(f)$ and $M_4(f)$ be finite. Then*

$$M_3(f) \;\le\; \sqrt{2M_2(f)M_4(f)}. \tag{4.2}$$

Proof:
Let $x \in \operatorname{dom} f$. Then for any direction $h \in \mathbb{E}$ and $\tau > 0$ small enough, we have $x + \tau h \in \operatorname{dom} f$ and

$$\begin{aligned}
0 \;&\preceq\; \nabla^2 f(x+\tau h) \overset{(1.6)}{=} \nabla^2 f(x) + \tau D^3 f(x)[h] + \tau^2 \int_0^1 (1-\lambda) D^4 f(x+\lambda h)[h]^2 d\lambda \\
&\preceq\; \nabla^2 f(x) + \tau D^3 f(x)[h] + \tfrac{1}{2}\tau^2 M_4(f)\|h\|^2 B.
\end{aligned}$$

Thus, $D^3 f(x)[h]^3 \le \frac{1}{\tau}\langle \nabla^2 f(x)h, h\rangle + \frac{\tau}{2} M_4(f)\|h\|^4$. Minimizing this inequality in $\tau > 0$ and taking the supremum of the result in $h \in \mathbb{E}$, we obtain inequality (4.2). $\Box$

Thus, from now on, we assume that

$$L_3(f) \;\equiv\; M_4(f) \;<\; +\infty. \tag{4.3}$$

Assumption $M_2(f) < +\infty$ is not so necessary. We will discuss different variants of its replacements in Section 5.

In our situation, we can apply to problem (4.1) the third-order tensor method ATMI$_3$ (see (2.15)). At each iteration of this method we need to minimize the augmented third-order Taylor polynomial $\hat{\Omega}_{x,3,H}(\cdot)$. As it was shown in [15], this can be done by an auxiliary scheme based on the relative smoothness condition. This approach is based on the following matrix inequality (see Lemma 3 in [15]):

$$-\tfrac{1}{\xi}\nabla^2 f(x) - \tfrac{\xi}{2}M_4(f)\|h\|^2 B \;\preceq\; D^3 f(x)[h] \;\preceq\; \tfrac{1}{\xi}\nabla^2 f(x) + \tfrac{\xi}{2}M_4(f)\|h\|^2 B, \tag{4.4}$$

which is valid for all $x \in \mathrm{dom}\, f$, $h \in \mathbb{E}$ and $\xi > 0$.

As compared with [15], our situation is more complicated. Firstly, we are not going to use the exact minimum of function $\hat{\Omega}_{x,3,H}(\cdot)$. And secondly, we are going to minimize this function using its *approximate gradients.*

Let us start from discussion of the second issue. Let us fix a parameter $\tau > 0$ and consider the following vector functions:

$$\begin{array}{rcl} h_y^\tau(x) & = & \frac{2}{\tau^2}[\nabla f(y + \tau(x-y)) - \nabla f(y) - \tau\nabla^2 f(x)(y-x)] \;\in\; \mathbb{E}^*, \\ \\ g_y^\tau(x) & = & \frac{1}{\tau^2}[\nabla f(y+\tau(x-y)) + \nabla f(y - \tau(x-y)) - 2\nabla f(y)] \;\in\; \mathbb{E}^*, \quad x, y \in \mathbb{E}. \end{array}$$

They are the finite-difference approximations of the third derivative along direction $[x-y]^2$.

Lemma 5 *For any $x, y \in \mathbb{E}$, we have*

$$\|h_y^\tau(x) - D^3 f(y)[x-y]^2\|_* \;\leq\; \tfrac{\tau}{3}M_4(f)\|x-y\|^3, \tag{4.5}$$

$$\|g_y^\tau(x) - D^3 f(y)[x-y]^2\|_* \;\leq\; \tfrac{\tau}{3}M_4(f)\|x-y\|^3, \tag{4.6}$$

$$\|g_y^\tau(x) - D^3 f(y)[x-y]^2\|_* \;\leq\; \tfrac{\tau^2}{12}L_4(f)\|x-y\|^4. \tag{4.7}$$

Proof:
Denote $h = \tau(x-y)$. Then, by Taylor formula we have

$$\nabla f(y+h) - \nabla f(y) - \nabla^2 f(y) h - \tfrac{1}{2}D^3 f(y)[h]^2 \;\overset{(1.6)}{=}\; \tfrac{1}{2}\int\limits_0^1 (1-\lambda)^2 D^4 f(y + \lambda h)[h]^3 d\lambda.$$

Applying a uniform upper bound for the forth derivative to the right-hand side of this representation, we get inequality (4.5). Further,

$$\nabla f(y-h) - \nabla f(y) + \nabla^2 f(y) h - \tfrac{1}{2}D^3 f(y)[h]^2 \;\overset{(1.6)}{=}\; \tfrac{1}{2}\int\limits_0^1 (1-\lambda)^2 D^4 f(y - \lambda h)[-h]^3 d\lambda.$$

Adding these two representations, we get

$$\begin{array}{rl} & g_y^\tau(x) - D^3 f(y)[x-y]^2 \\ = & \frac{\tau}{2}\int\limits_0^1 (1-\lambda)^2 \Big(D^4 f(y+\lambda\tau(x-y)) - D^4 f(y-\lambda\tau(x-y)) \Big) [x-y]^3 d\lambda, \end{array}$$

and we obtain inequality (4.6). If the forth derivative derivative is Lipschitz continuous, then

$$\| g_y^\tau(x) - D^3 f(y)[x-y]^2 \|_* \quad \leq \quad \frac{\tau}{2}\int\limits_0^1 (1-\lambda)^2 \cdot 2\lambda\tau \|x-y\|^4 L_4(f) d\lambda,$$

and this is inequality (4.7). □

In this paper, we usually employ the approximation $g_y^\tau(\cdot)$. Note that

$$\nabla \hat{\Omega}_{y,3,H}(x) \quad = \quad \nabla f(y) + \nabla^2 f(y)(x-y) + \tfrac{1}{2} D^3 f(y)[x-y]^2 + \tfrac{H}{6}\|x-y\|^2 B(x-y).$$

Thus, we can easily compute approximate gradients of function $\hat{\Omega}_{y,3,H}(\cdot)$ using the first-order information on function $f(\cdot)$. Let us show that this can help us to minimize the augmented Taylor polynomial of degree three by the machinery presented in Section 3.

At each iteration k of ATMI$_3$, we need to find point $x_{k+1} \in \mathcal{N}_3(y_k)$. For the sake of notation, let us assume that $y_k = 0$. We need to find a point $x_+ \in \mathcal{N}_3(0)$ by minimizing the function

$$\varphi_k(x) \; = \; \hat{\Omega}_{0,3,6L_3}(x) \quad = \quad f(0) + \langle \nabla f(0), x\rangle + \tfrac{1}{2}\langle \nabla^2 f(0) x, x\rangle + \tfrac{1}{6} D^3 f(0)[x]^3 + \tfrac{L_3}{4}\|x\|^4.$$

Thus, our auxiliary problem is as follows:

$$\min_{x\in\mathbb{E}} \; \varphi_k(x). \tag{4.8}$$

Denote $x_k^* = \arg\min\limits_{x\in\mathbb{E}} \varphi_k(x)$ and $\varphi_k^* = \varphi_k(x_k^*)$. Note that

$$\nabla \varphi_k(x) \quad = \quad \nabla f(0) + \nabla^2 f(0) x + \tfrac{1}{2} D^3 f(0)[x]^2 + L_3 \|x\|^2 Bx, \tag{4.9}$$

$$\begin{array}{rcl} \nabla^2 \varphi_k(x) & = & \nabla^2 f(0) + D^3 f(0)[x] + L_3 \Big(\|x\|^2 B + 2Bxx^* B \Big) \\ & = & \nabla^2 f(0) + D^3 f(0)[x] + L_3 \nabla^2 d_4(x). \end{array} \tag{4.10}$$

Therefore,

$$\begin{array}{rcl} \nabla^2 \varphi_k(x) & \overset{(4.4)}{\preceq} & \left(1+\frac{1}{\xi}\right)\nabla^2 f(0) + \left(1+\frac{\xi}{2}\right) L_3 \nabla^2 d_4(x), \\ \nabla^2 \varphi_k(x) & \overset{(4.4)}{\succeq} & \left(1-\frac{1}{\xi}\right)\nabla^2 f(0) + \left(1-\frac{\xi}{2}\right) L_3 \nabla^2 d_4(x), \end{array} \tag{4.11}$$

Now it is clear that in our case a good scaling function is as follows:

$$\rho_k(x) \quad = \quad \tfrac{1}{2}\langle \nabla^2 f(0) x, x\rangle + L_3 d_4(x), \quad x \in \mathbb{E}. \tag{4.12}$$

Indeed, applying the relations (4.11) with $\xi = \sqrt{2}$, we get

$$\left(1 - \tfrac{1}{\sqrt{2}}\right) \nabla^2 \rho_k(x) \;\; \preceq \;\; \nabla^2 \varphi_k(x) \;\preceq\; \left(1 + \tfrac{1}{\sqrt{2}}\right) \nabla^2 \rho_k(x), \quad x \in \mathbb{E}.$$

Thus, we can take

$$\mu \;\; \equiv \;\; \mu_{\rho_k}(\varphi_k) \; = 1 - \tfrac{1}{\sqrt{2}} \;\; = \;\; \tfrac{1}{2+\sqrt{2}}, \quad L \;\equiv\; L_{\rho_k}(\varphi_k) \;\; = \;\; 1 + \tfrac{1}{\sqrt{2}},$$

and obtain for function $\varphi_k(\cdot)$ the condition number bounded by an absolute constant:

$$\gamma(\varphi) \;\; \stackrel{\mathrm{def}}{=} \;\; \tfrac{\mu_{\rho_k}(\varphi_k)}{L_{\rho_k}(\varphi_k)} \;\; = \;\; \tfrac{1}{(1+\sqrt{2})^2} \;\; = \;\; \tfrac{1}{3+2\sqrt{2}} \;\; > \;\; \tfrac{1}{6}. \tag{4.13}$$

The second condition for applicability of method (3.6) is the uniform convexity of the Bregman distance. In our case, this is true since

$$\beta_{\rho_k}(x, y) \;\; \geq \;\; L_3 \beta_{d_4}(x, y) \;\; \stackrel{(2.12)}{\geq} \;\; \tfrac{1}{16} L_3 \|x - y\|^4, \quad x, y \in \mathbb{E}. \tag{4.14}$$

Thus, in terms of inequality (3.5), we have $\sigma_4(\rho_k) = \frac{1}{4} L_3$. This property is important for bounding the size of the set

$$\mathcal{L}_k \;\; = \;\; \{x \in \mathbb{E} : \; \varphi_k(x) \leq \varphi_k(0)\}.$$

Lemma 6 *For any $x \in \mathcal{L}_k$ we have*

$$\|x\| \;\; \leq \;\; 2^{1/3} R_k, \quad \|x_k^*\| \;\leq\; R_k \;\stackrel{\mathrm{def}}{=}\; 2 \left(\tfrac{2+\sqrt{2}}{L_3} \|\nabla f(0)\|_*\right)^{\frac{1}{3}}. \tag{4.15}$$

Proof:
Indeed,

$$\begin{aligned}
\langle \nabla f(0), 0 - x_k^* \rangle \;\; &= \;\; \langle \nabla \varphi_k(0), 0 - x_k^* \rangle \;\; = \;\; \varphi_k(0) - \varphi_k^* + \beta_{\varphi_k}(0, x_k^*) \\
&= \;\; \beta_{\varphi_k}(x_k^*, 0) + \beta_{\varphi_k}(0, x_k^*) \;\; \geq \;\; \mu[\beta_{\rho_k}(x_k^*, 0) + \beta_{\rho_k}(0, x_k^*)] \\
&\geq \;\; \mu L_3 [\beta_{d_4}(x_k^*, 0) + \beta_{d_4}(0, x_k^*)] \;\; \stackrel{(2.12)}{\geq} \;\; 2\mu \tfrac{L_3}{16} \|x_k^*\|^4.
\end{aligned}$$

Consequently, we have the following bound:

$$\|x_k^*\| \;\; \leq \;\; 2 \left[\tfrac{2+\sqrt{2}}{L_3} \|\nabla f(0)\|_*\right]^{\frac{1}{3}} \;\; = \;\; R_k. \tag{4.16}$$

Further, for $x \in \mathcal{L}_k$, we have

$$\begin{aligned}
\langle \nabla \varphi_k(0), 0 - x \rangle \;\; &= \;\; \varphi_k(0) - \varphi_k(x) + \beta_{\varphi_k}(0, x) \;\; \geq \;\; \beta_{\varphi_k}(0, x) \\
&\geq \;\; \mu L_3 \beta_{d_4}(0, x) \;\; \stackrel{(2.12)}{\geq} \;\; \tfrac{\mu L_3}{16} \|x\|^4.
\end{aligned}$$

Thus, $\|x\| \leq \left[\frac{16}{\mu L_3} \|\nabla f(0)\|*\right]^{\frac{1}{3}} = 2^{1/3} R_k$. $\square$

The third condition is the possibility of approximating the gradient of function $\varphi_k(\cdot)$. In our case, in view of Lemma 5, we can take

$$g_{\varphi_k,\tau}(x) \;=\; \nabla f(0) + \nabla^2 f(0)x + \tfrac{1}{2} g_0^\tau(x) + L_3\|x\|^2 Bx, \tag{4.17}$$

where $g_0^\tau(x) = \frac{1}{\tau^2}[\nabla f(\tau x) + \nabla f(-\tau x) - 2\nabla f(0)]$. In this case,

$$\|g_{\varphi_k,\tau}(x) - \nabla \varphi_k(x)\|_* \;\overset{(4.6)}{\leq}\; \tfrac{\tau}{3} L_3 \|x\|^3, \quad x \in \mathbb{E}. \tag{4.18}$$

Thus, in order to ensure condition (3.4) and keep τ separated from zero (this is necessary for stability of the process), we need to guarantee the boundedness of the minimizing sequence for function $\varphi_k(\cdot)$. However, since we know an explicit upper bound (4.15) on the size of the optimal point, it is possible to ensure this by introducing an additional constraint on the size of variables. Let us replace the problem (4.8) by the following one:

$$\min_{x \in S_k} \varphi_k(x), \quad S_k \overset{\text{def}}{=} \{x \in \mathbb{E} : \; \|x\| \leq R_k\}. \tag{4.19}$$

In view of Lemma 6, the optimal solutions of problems (4.8) and (4.19) coincide.

Consider the following variant of method (3.6) with accuracy parameter $\delta > 0$.

Initialization. Choose $x_0 = 0$ and set $\tau = \frac{3\delta}{8(2+\sqrt{2})\|\nabla f(0)\|_*}$.

For $i \geq 0$ iterate:

1. Compute the approximate gradient $g_{\varphi_k,\tau}(x_i)$ by (4.17).

2. If $\|g_{\varphi_k,\tau}(x_i)\|_* \leq \frac{1}{6}\|\nabla f(x_i)\|_* - \delta$, then STOP.

3. Else, compute the new point

$$x_{i+1} = \arg\min_{x \in S_k} \left\{ \langle g_{\varphi_k,\tau}(x_i), x\rangle + 2\left(1 + \tfrac{1}{\sqrt{2}}\right)\beta_{\rho_k}(x_i, x)\right\}.$$

(4.20)

Let us mention the main properties of this minimization process. First of all, since all points x_i belong to S_k, for all $i \geq 0$ we have

$$\|g_{\varphi_k,\tau}(x_i) - \nabla\varphi_k(x_i)\|_* \;\overset{(4.18)}{\leq}\; \tfrac{\tau}{3} L_3 R_k^3 \;=\; \tfrac{\delta L_3}{8(2+\sqrt{2})\|\nabla f(0)\|_*} \cdot \tfrac{8(2+\sqrt{2})}{L_3}\|\nabla f(0)\|_* \;=\; \delta.$$

This means, in particular, that the sopping criterion at Step 2 of method (4.20) is correct: if it is satisfied, then $x_i \in \mathcal{N}_3(0)$.

Moreover, all conditions of Lemma 1 are satisfied with

$$d(\cdot) \;=\; \rho_k(\cdot), \quad L_{\rho_k}(\varphi_k) \;=\; 1 + \tfrac{1}{\sqrt{2}}, \quad \gamma_{\rho_k}(\varphi_k) \;=\; \tfrac{1}{6}, \quad \sigma_4(\rho_k) \;=\; \tfrac{1}{4}L_3.$$

Therefore, in our case, inequality (3.7) with $p = 3$ can be rewritten in the following form:

$$\beta_{\rho_k}(x_{i+1}, x) \quad \leq \quad \left(1 - \tfrac{1}{24}\right) \beta_{\rho_k}(x_i, x) + \tfrac{1}{2+\sqrt{2}}[\varphi_k(x) - \varphi_k(x_{i+1})] + \hat{\delta}, \tag{4.21}$$

where $\hat{\delta} = \frac{3}{2}\delta^{\frac{4}{3}}\left(\frac{208}{L_3}\right)^{\frac{1}{3}} < \hat{\delta}_+ \stackrel{\mathrm{def}}{=} \frac{9\delta^{4/3}}{L_3^{1/3}}$. Consequently, in view of inequality (3.10), we have

$$\min_{0 \leq i \leq T} \varphi_k(x_i) - \varphi_k^* \quad \leq \quad (2 + \sqrt{2}) \left\{ \frac{L_1 R_k^2 + \frac{1}{2} L_3 R_k^4}{6\left[\left(1+\frac{1}{23}\right)^T - 1\right]} + \hat{\delta}_+ \right\}, \quad T \geq 1, \tag{4.22}$$

where L_1 is any upper estimate for the value $\|\nabla^2 f(0)\|$.

From this bound, we have a natural limit for the number of iterations of method (4.20), sufficient for obtaining the following inequality:

$$\varphi_k(\hat{x}_T) - \varphi_k^* \quad \leq \quad 2(2 + \sqrt{2})\hat{\delta}_+, \tag{4.23}$$

where $\hat{x}_T = \arg\min\limits_x \left\{ \varphi_k(x) : \; x \in \{0, x_1, \dots, x_T\} \right\} \in \mathcal{L}_k$. Indeed, for this it is enough to have

$$1 + \tfrac{6}{\hat{\delta}_+}[L_1 R_k^2 + \tfrac{1}{2} L_3 R_k^4] \quad \leq \quad e^{T/24} \quad \left(\leq \; \left(1 + \tfrac{1}{23}\right)^T \right).$$

Hence, we have the following bound:

$$T \quad \leq \quad T_k(\delta) \stackrel{\mathrm{def}}{=} 24 \ln \left(1 + \tfrac{2}{3} \left(\tfrac{1}{\delta}\right)^{4/3} L_3^{1/3} \left[L_1 R_k^2 + \tfrac{1}{2} L_3 R_k^4 \right] \right). \tag{4.24}$$

However, the upper-level method ATMI$_3$ needs a point with a small norm of the gradient:

$$\|\nabla \varphi_k(\hat{x}_T)\|_* \quad \leq \quad \tfrac{1}{6}\|\nabla f(0)\|_*. \tag{4.25}$$

In order to derive this bound from inequality (4.23) with an appropriate value of $\hat{\delta}_+$, we use the fact that out scaling function $\rho_k(\cdot)$ is norm-dominated. Indeed, in view of Lemma 2 and representation (4.12), this function is norm-dominated on any Euclidean ball B_r by the following function:

$$\theta_r(\tau) \quad = \quad \tfrac{1}{2}(L_1 + 5L_3 r^2)\tau^2 + \tfrac{1}{2} L_3 \tau^4.$$

Hence, in view of Lemma 6, our scaling function $\rho_k(\cdot)$ is norm-dominated on the set $\mathcal{L}_k$ by the function $\theta_{\hat{r}_k}(\cdot)$ with $\hat{r}_k = 2^{1/3} R_k$. Therefore, for applying Lemma 3, we need to estimate from above the inverse to its conjugate function.

Lemma 7 *For any $r > 0$, we have*

$$\left(\theta_r^*\right)^{-1}(\xi) \quad \leq \quad \sqrt{2(L_1 + 5L_3 r^2)\xi} + 2L_3^{1/4}\left(\tfrac{2}{3}\xi\right)^{3/4}, \quad \xi \geq 0. \tag{4.26}$$

Proof:
Consider the primal function $\theta(\tau) = \frac{a\tau^2}{2} + \frac{b\tau^4}{4}$ with $a, b \geq 0$. Then its conjugate function is defined as follows:

$$\theta^*(\lambda) \quad = \quad \max_\tau \left[\lambda\tau - \tfrac{a\tau^2}{2} - \tfrac{b\tau^4}{4} \right], \quad \lambda \geq 0.$$

We need to find $\lambda \geq 0$ from the equation $\xi = \theta^*(\lambda)$.

Note that the optimal solution $\tau = \tau(\lambda)$ in the above maximization problem can be found from the equation

$$\lambda \;=\; a\tau + b\tau^3. \tag{4.27}$$

Therefore,

$$\xi \;=\; \theta^*(\lambda) \;\overset{(4.27)}{=}\; \tfrac{a}{2}\tau^2(\lambda) + \tfrac{3b}{4}\tau^4(\lambda)$$

Thus, we can write down $\tau(\lambda)$ as a function of ξ:

$$\tau^2(\lambda) \;=\; \frac{4\xi}{a+\sqrt{a^2+12b\xi}} \;\leq\; \min\left\{2\frac{\xi}{a}, \sqrt{\frac{4\xi}{3b}}\right\}.$$

Hence,

$$\lambda \;\overset{(4.27)}{\leq}\; \sqrt{2a\xi} + b^{1/4}\left(\frac{4\xi}{3}\right)^{3/4}.$$

It remains to use the actual values $a = L_1 + 5L_3r^2$ and $b = 2L_3$. $\square$

Now we can write down the condition for our parameter δ, which ensures the desired inequality (4.25). Indeed, in view of inequalities (4.23) and (3.14), after $T_k(\delta)$ inner steps we can guarantee that

$$\|\nabla\varphi_k(\hat{x}_T)\|_* \;\leq\; L\cdot(\theta^*_{\hat{r}_k})^{-1}\Big(\tfrac{2}{L}(2+\sqrt{2})\hat{\delta}_+\Big) \;=\; L\cdot(\theta^*_{\hat{r}_k})^{-1}\Big(4\hat{\delta}_+\Big),$$

where $L = 1 + \frac{1}{\sqrt{2}}$. In order to stop method (4.20) at this moment we need to guarantee that the norm of the corresponding approximate gradient is small enough. Hence, our condition for the parameter δ can be derived from the following target inequalities:

$$\begin{aligned}\|g_{\varphi_k,\tau}(\hat{x}_T)\|_* \;&\leq\; \delta + \|\nabla\varphi_k(\hat{x}_T)\|_* \;\leq\; \delta + L\cdot(\theta^*_{\hat{r}_k})^{-1}\Big(4\hat{\delta}_+\Big)\\ &\overset{(?)}{\leq}\; \tfrac{1}{6}\|\nabla f(\hat{x}_T)\| - \delta.\end{aligned}$$

In view of Lemma 7, this inequality will be satisfied if

$$2\delta + 2L\sqrt{2(L_1+5L_3\hat{r}_k^2)\hat{\delta}_+} + 2L_3^{1/4}\left(\tfrac{8}{3}\hat{\delta}_+\right)^{3/4} \;\leq\; \tfrac{1}{6}\epsilon_g, \tag{4.28}$$

where $\epsilon_g > 0$ is a lower bound for the norm of the gradients of the objective function during the whole minimization process. Recall that

$$\hat{r}_k \;=\; 2^{4/3}\left(\tfrac{2+\sqrt{2}}{L_3}\|\nabla f(0)\|_*\right)^{\frac{1}{3}}, \quad \hat{\delta}_+ \;=\; \frac{9\delta^{4/3}}{L_3^{1/3}}.$$

Hence, this inequality can be rewritten in the followingform:

$$2(1+(24)^{3/4})\delta + 6L\delta^{2/3}\sqrt{\frac{2L_1}{L_3^{1/3}} + 10\Big(16(2+\sqrt{2})\|\nabla f(0)\|_*\Big)^{2/3}} \;\leq\; \tfrac{1}{6}\epsilon_g.$$

Using the closest upper integer bounds on the coefficients, it can be strengthen as follows:

$$24\delta + 21\delta^{2/3}\sqrt{\frac{1}{2L_3^{1/3}}\|\nabla^2 f(0)\| + 36\|\nabla f(0)\|_*^{2/3}} \quad \leq \quad \tfrac{1}{6}\epsilon_g, \tag{4.29}$$

where we replaced L_1 by $\|\nabla^2 f(0)\|$ since this corresponds to the actual role of this constant in the complexity analysis of method (4.20).

This means that, in accordance to (4.29), we need to choose

$$\delta \quad = \quad O\left(\frac{\epsilon_g^{3/2}}{\|\nabla f(0)\|_*^{1/2} + \|\nabla^2 f(0)\|^{3/2}/L_3^{1/2}}\right). \tag{4.30}$$

Since $\|\nabla f(0)\|_* \geq \epsilon_g$, we always have $\delta \leq O(\epsilon_g)$.

Note that all coefficients in the condition (4.29) are known (provided that we have a good estimate for the Lipschitz constant L_3). Thus, we have

$$T_k(\delta) \quad = \quad O\left(\ln \tfrac{G+H}{\epsilon_g}\right),$$

where G and H are the uniform upper bounds for the norms of the gradients and Hessians computed at the points generated by the main process. Validity of the assumption on the finiteness of these bounds will be discussed in Section 5.

Let us write down our inexact algorithmic schemes (2.10) and (2.15), employing the inner procedure (4.20). These methods have only one accuracy parameter $\delta > 0$, which must be chosen in accordance to condition (4.29). They need also the constant L_3.

We start from the variant of Inexact Basic Tensor Method (2.10).

Inexact 3rd-Order Tensor Method

Initialization. Choose $x_0 \in \mathbb{E}$.

Iteration $k \geq 0$**.**

Compute $x_{k+1} \in \mathcal{N}_3(x_k)$ by method (4.20) with the following settings:

a) Starting point $x_{k,0} = x_k$. Approximation step $\tau = \frac{3\delta}{8(2+\sqrt{2})\|\nabla f(x_k)\|_*}$.

b) Objective function $\varphi_k(x) = \hat{\Omega}_{x_k,3,6L_3}(x)$.

c) Feasible set $S_k = \left\{ x : \|x - x_k\| \leq 2\left[\frac{2+\sqrt{2}}{L_3}\|\nabla f(x_k)\|_*\right]^{\frac{1}{3}} \right\}$.

d) Scaling function $\rho_k(x) = \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k\rangle + L_3 d_4(x - x_k)$.

(4.31)

At each iteration of this method, we have $O\left(\ln \frac{G+H}{\epsilon_g}\right)$ iterations of the inner scheme. Each of them needs three calls of oracle of the main objective function (twice for computing the approximate gradient of function $\varphi_k(\cdot)$ and once for verifying the stopping criterion). In view of Theorem 2, the rate of convergence of the main process is as follows:

$$f(x_k) - f^* \quad \leq \quad \left(\tfrac{8}{k}\right)^3 \left[\tfrac{7}{10} L_3 R^4(x_0) + \tfrac{1}{2}(f(x_0) - f^*)\right], \quad k \geq 1. \tag{4.32}$$

Thus, the analytical complexity bound of the method (4.31) is of the order

$$O\left(R(x_0) \cdot \left(\tfrac{L_3}{\epsilon_f}\right)^{1/3} \ln \tfrac{G+H}{\epsilon_g} \right), \tag{4.33}$$

where $\epsilon_f > 0$ is the desired accuracy in the function value. Note that this method uses only the second-order oracle.

Let us look now at the accelerated scheme.

Inexact Accelerated 3rd-Order Tensor Method

Initialization. Choose $x_0 \in \mathbb{E}$. Define coefficients A_k by (2.13) with $p = 3$. Define function $\psi_0(x) = d_4(x - x_0)$.

Iteration $k \geq 0$**.** (4.34)

1. Compute $v_k = \arg\min\limits_{x \in \mathbb{E}} \psi_k(x)$ and choose $y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_k$.

2. Compute $x_{k+1} \in \mathcal{N}_3(y_k)$ by method (4.20) with the following settings:

- **a)** Starting point $x_{k,0} = y_k$. Approximation step $\tau = \frac{3\delta}{8(2+\sqrt{2})\|\nabla f(y_k)\|_*}$.
- **b)** Objective function $\varphi_k(x) = \hat{\Omega}_{y_k,3,6L_3}(x)$.
- **c)** Feasible set $S_k = \left\{ x : \|x - y_k\| \leq 2\left[\frac{2+\sqrt{2}}{L_3}\|\nabla f(y_k)\|_*\right]^{\frac{1}{3}} \right\}$.
- **d)** Scaling function $\rho_k(x) = \frac{1}{2}\langle \nabla^2 f(y_k)(x - y_k), x - y_k\rangle + L_3 d_4(x - y_k)$.

3. Update $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]$.

As before, each iteration of this method needs at most $O\left(\ln \frac{G+H}{\epsilon_g}\right)$ iterations of the inner scheme. In view of Theorem 3, the rate of convergence of the main process in (4.34) is as follows:

$$f(x_k) - f^* \quad \leq \quad \tfrac{7}{60}\left(\tfrac{6}{k}\right)^4 \cdot L_3 \|x_0 - x^*\|^4, \quad k \geq 1. \tag{4.35}$$

Thus, the analytical complexity bound of this method is of the order

$$O\left(\|x_0 - x^*\| \cdot \left(\tfrac{L_3}{\epsilon_f}\right)^{1/4} \ln \tfrac{G+H}{\epsilon_g} \right), \tag{4.36}$$

Recall that method (4.34) is a second-order scheme.

5 Bounds for the derivatives

The complexity analysis in Section 4 is valid only if we can guarantee the finiteness of the constants G and H. The simplest way of doing this consists in considering the following class of functions:

$$\mathcal{M}_{1,2,4} \quad = \quad \{ f \in \mathbb{C}^4(\mathbb{E}) : \; M_1(f) < +\infty, \; M_2(f) < +\infty, \; M_4(f) < +\infty \}. \tag{5.1}$$

This is a nontrivial class, but it is quite restrictive. In this section, we show that it is possible to derive the finiteness of G and H from our main assumption (4.3) and the properties of the minimization schemes.

Indeed, we can easily bound the derivatives at test points belonging to a bounded set. Let us present a trivial result, which follows directly from the Taylor formula (1.6).

Lemma 8 *For any $x \in B_D(x_0) \stackrel{\mathrm{def}}{=} \{ x \in \mathbb{E} : \; \| x - x_0 \| \leq D \}$, we have*

$$\begin{array}{rcl} \| \nabla f(x) \|_* & \leq & \| \nabla f(x_0) \|_* + \| \nabla^2 f(x_0) \| D + \frac{1}{2} \| D^3 f(x_0) \| D^2 + \frac{1}{6} M_4(f) D^3, \\ \| \nabla^2 f(x) \| & \leq & \| \nabla^2 f(x_0) \| + \| D^3 f(x_0) \| D + \frac{1}{2} M_4(f) D^2. \end{array} \tag{5.2}$$

We can use the right-hand sides of inequalities (5.2) as our constants G and H provided that the distance between x_0 and the test points does not exceed some $D < +\infty$. Note that we do not use D, G, and H in our methods. They appear only in the bounds for the number of inner steps and stay inside the logarithm. The important criterion (4.29), defining an appropriate value of the parameter $\delta > 0$, is based on the available information about the first and second derivatives at the current test point.

Thus, we need to prove that the sequences of test points in our methods are bounded. Let us start from Inexact Basic Tensor Method (4.31). For this method, the situation is very simple. We have already assumed that the size of the level set $R(x_0)$ is finite. Since the method (4.31) is monotone, for any x_k generated by this scheme, we have

$$\| x_k - x_0 \| \;\; \leq \;\; \| x_k - x^* \| + \| x^* - x_0 \| \; \leq \; 2R(x_0), \quad k \geq 0.$$

Thus, we can take in (5.2) $D = 2R(x_0)$.

Let us look now at Inexact Accelerated Tensor Method. Actually, for proving the boundedness of sequences of the test points $\{ y_k \}_{k \geq 0}$, it is better to consider its monotone variant.

Monotone Inexact Accelerated 3rd-Order Tensor Method

Initialization. Choose $x_0 \in \mathbb{E}$. Define coefficients A_k by (2.13) with $p = 3$. Define function $\psi_0(x) = d_4(x - x_0)$.

Iteration $k \geq 0$**.**

1. Compute $v_k = \arg\min\limits_{x \in \mathbb{E}} \psi_k(x)$ and choose $y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_k$.

2. Compute $\hat{x}_{k+1} \in \mathcal{N}_3(y_k)$ by method (4.20) with the following settings:

a) Starting point $x_{k,0} = y_k$. Approximation step $\tau = \frac{3\delta}{8(2+\sqrt{2})\|\nabla f(y_k)\|_*}$.

b) Objective function $\varphi_k(x) = \hat{\Omega}_{y_k,3,6L_3}(x)$.

c) Feasible set $S_k = \left\{ x : \|x - y_k\| \leq 2 \left[\frac{2+\sqrt{2}}{L_3} \|\nabla f(y_k)\|_* \right]^{\frac{1}{3}} \right\}$.

d) Scaling function $\rho_k(x) = \frac{1}{2}\langle \nabla^2 f(y_k)(x - y_k), x - y_k \rangle + L_3 d_4(x - y_k)$.

3. Update $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(\hat{x}_{k+1}) + \langle \nabla f(\hat{x}_{k+1}), x - \hat{x}_{k+1} \rangle]$.

4. Choose $x_{k+1} = \arg\min\limits_{x} \left\{ f(x) : \; x \in \{x_0, \dots, x_k, \hat{x}_{k+1}\} \right\}$.

(5.3)

The additional Step 4 of this method ensures monotonicity of the sequence $\{f(x_k)\}_{k \geq 0}$. The complexity analysis, presented in the end of Section 2, remains also valid for the monotone variant (5.3). Indeed, in the right-hand side of the relation (2.18), we can replace the point x_k by any point with better value of the objective function.

Lemma 9 *Let points* $\{y_k\}_{k \geq 0}$ *be generated by the method (5.3). Then*

$$\|y_k - x_0\| \;\; \leq \;\; (1 + \sqrt{2}) R(x_0), \quad k \geq 0. \tag{5.4}$$

Proof:
Indeed, choosing in the relation (2.17) $p = 3$ and $x = x^*$, we get

$$\frac{1}{16}\|v_k - x^*\|^4 \;\; \leq \;\; \frac{1}{4}\|x^* - x_0\|^4$$

At the same time, since $f(x_k) \leq f(x_0)$, we have $\|x_k - x^*\| \leq R(x_0)$. Hence, in view of the definition of y_k at Step 1 in (5.3),

$$\begin{aligned} \|y_k - x_0\| \;\; &\leq \;\; \max\{\|x_k - x_0\|, \|v_k - x_0\|\} \; \leq \; \max\{2R(x_0), (1 + \sqrt{2})R(x_0)\} \\ &= \;\; (1 + \sqrt{2}) R(x_0). \end{aligned}$$

□

Thus, for the accelerated monotone method (5.3) we can take $D = (1+\sqrt{2})R(x_0)$.

6 Conclusion

The main conclusion, which we can derive from our results is that the existing classification of the problem classes, optimization schemes, and complexity bounds is far from being perfect. Traditionally, we put in one-to-one correspondence the type of numerical schemes (classified by its order) and the problem classes (classified by the Lipschitz condition for the highest derivative). In this way, we attach the first-order methods to functions with Lipschitz-continuous gradients. The second-order methods correspond to the functions with Lipschitz-continuous Hessian, etc.

This picture allows us to speak about the *optimal methods*. For example, we say that the Fast Gradient Methods (FGM) with the convergence rate $O\left(k^{-2}\right)$ are the optimal first-order methods. However, the only reason why FGM could be called optimal is that they implement the lower bound for a certain *problem class*, which is considered to be the natural field of application for the first-order methods only.

However, now it is clear the above over-simplified picture of the world must be replaced by something more elaborated. We have seen that there exist problem classes for which the second- and the third-order methods demonstrate the same rate of convergence. So, the correct classification of problem classes and optimization methods must be at least two-dimensional. This is, of course, an interesting topic for the further research.

Acknowledgments. The author would like to thank Alexander Gasnikov for careful reading and comments on the text.

References

[1] N. Agarwal, E. Hazan. Lower Bounds for Higher-Order Convex Optimization. arXiv:1710.10329v1 [math.OC] (2017).

[2] Arjevani, O. Shamir, and R. Shiff. Oracle Complexity of Second-Order Methods for Smooth Convex Optimization. arXiv:1705.07260 [math.OC] (2017).

[3] H.H. Bauschke, J. Bolte, M. Teboulle.. A descent lemma beyond Lipschitz gradient continuity: first order methods revisited and applications. *Math. Oper. Res.* **42**, 330–348 (2016).

[4] M. Baes. Estimate sequence methods: extensions and approximations. *Optimization Online* (2009).

[5] S. Bubeck, Q. Jiang, Y.T. Lee, Y. Li, and A. Sidford. Nera-optimal method for highly nonsmooth convex optimization. arXiv: 1812.08026v2 [math.OC] 23 Jun 2019.

[6] E. G. Birgin, J. L. Gardenghi, J. M. Martinez, S. A. Santos, and Ph. L. Toint. Worstcase evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, **163**, 359-368 (2017).

[7] A. Gasnikov, E. Gorbunov, D. Kovalev, A. Mohhamed, and E. Chernousova. The global rate of convergence for optimal tensor methods in smooth convex optimization. arXiv: 1809.00382, (2018).

[8] G. N. Grapiglia and Yu. Nesterov. On inexact solution of auxiliary problems in tensor methods for convex optimization. *CORE Discussion Paper* #2019/30 (Dec. 2019)

[9] B. Jiang, H. Wang, and S. Zang. An optimal high-order tensor method fo convex optimization. arXiv: 1812.06557, (2018).

[10] H. Lu, R. Freund, and Yu. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIOPT*, **28**(1), 333-354 (2018).

[11] R.D.C. Monteiro and B.F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to the second-order methods, *SIOPT*, **23**(2), 1092-1125 (2013).

[12] Yu. Nesterov, A. Nemirovskii. *Interior point polynomial methods in convex programming: Theory and Applications,*, SIAM, Philadelphia, 1994.

[13] Yu. Nesterov, B. Polyak. "Cubic regularization of Newton's method and its global performance". *Mathematical Programming*, **108**(1), 177-205 (2006).

[14] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, **112**(1) 159-181 (2008)

[15] Yu. Nesterov. Implementable Tensot Methods in Unconstrained Convex Optimization. *Mathematical Programming*, DOI 10.1007/s10107-019-01449-1 (2019)

[16] Yu. Nesterov. Inexact Basic Tensor Methods. *CORE DP* #2019/23, (2019),

[17] Yu. Nesterov. Lectures on Convex Optimization. *Springer* (2018).