Journal of Optimization Theory and Applications (2021) 191:1–30
https://doi.org/10.1007/s10957-021-01930-y

Superfast Second-Order Methods for Unconstrained Convex Optimization

Yurii Nesterov[1]

Received: 18 June 2020 / Accepted: 16 August 2021 / Published online: 29 August 2021

Abstract
In this paper, we present new second-order methods with convergence rate $O\left(k^{-4}\right)$, where k is the iteration counter. This is faster than the existing lower bound for this type of schemes (Agarwal and Hazan in Proceedings of the 31st conference on learning theory, PMLR, pp. 774–792, 2018; Arjevani and Shiff in Math Program 178(1–2):327–360, 2019), 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 Nesterov (Math Program 186:157–183, 2021) 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 (Bauschke et al. in Math Oper Res 42:330–348, 2016; Lu et al. in SIOPT 28(1):333–354, 2018). 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

Mathematics Subject Classification 90C25

1 Introduction

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

Communicated by Anil Aswani.

✉ Yurii Nesterov
Yurii.Nesterov@uclouvain.be

[1] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium

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 [6,7,9], given by the theoretical lower complexity bounds [1,2].

It is interesting that the first accelerated tensor methods were analyzed in the unpublished paper [3], where the author did not express any hope for their practical implementations in the future. In [3] 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 is 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 [6,7,9] 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 [4,10]. This auxiliary method has to minimize an augmented Taylor polynomial of degree three, computed at the current test point $x \in \mathbb{R}^n$:

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

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

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

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 Sect. 2, we introduce a convenient definition of the acceptable neighborhood of the exact tensor step. It differs from the previous ones (e.g. [5,8,13]) since for its verification it is necessary to call the 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 Sect. 3, we analyze the rate of convergence of the gradient method based on the relative smoothness condition [4,10], under the assumption that the gradient of the objective function is computed with a small absolute error. We need this analysis for replacing 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.

In Sect. 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 4.1).

In Sect. 5, we show how to ensure boundedness of the constants, essential for our minimization schemes. Finally, we conclude the paper with Sect. 6, containing a discussion of our results and directions for 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

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

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 our analysis, we use *Bregman divergence* of function $f(\cdot)$ defined as follows:

$$\beta_f(x, y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle, \quad x, y \in \mathrm{dom}\, f. \tag{1}$$

We often work with directional derivatives. For $p \geq 1$, denote by

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

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

$$\|D^p f(x)\| = \max_{h_1, \ldots, h_p} \left\{ \left| D^p f(x)[h_1, \ldots, h_p] \right| : \ \|h_i\| \leq 1, \ i = 1, \ldots, p \right\}. \tag{2}$$

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

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

For Hessian, this gives the *spectral norm* of self-adjoint linear operator (the maximal module of all eigenvalues computed with respect to operator B).

If all directions $h_1, \ldots, h_p$ 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

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

Note that, in general, we have (see, for example, Appendix 1 in [16])

$$\|D^p f(x)\| = \max_{h} \left\{ \left| D^p f(x)[h]^p \right| : \ \|h\| \leq 1 \right\}. \tag{3}$$

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

$$\|D^p f(x) - D^p f(y)\| = \max_{h} \left\{ \left| D^p f(x)[h]^p - D^p f(y)[h]^p \right| : \ \|h\| \leq 1 \right\}. \tag{4}$$

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:

$$\|D^p f(x) - D^p f(y)\| \leq L_p \|x - y\|, \quad x, y \in \operatorname{dom} f, \quad p \geq 1. \tag{5}$$

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

$$M_p(f) = \sup_{x \in \operatorname{dom} f} \|D^p f(x)\|. \tag{6}$$

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, $F : \operatorname{dom} F \to \mathbb{E}_2$. Then, by the well-known Taylor formula, we have

$$\begin{aligned} F(y) \; - \; & F(x) - \sum_{k=1}^{p} \frac{1}{k!} D^k F(x)[y - x]^k \\ = \; & \frac{1}{p!} \int_1^1 (1 - \tau)^p D^{p+1} F(x + \tau(y - x))[y - x]^{p+1} d\tau, \quad x, y \in \operatorname{dom} F. \end{aligned} \tag{7}$$

Hence, we can bound the following residual:

$$|f(y) - \Omega_{x,p}(y)| \leq \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad x, y \in \operatorname{dom} f. \tag{8}$$

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

$$\|\nabla f(y) - \nabla \Omega_{x,p}(y)\|_* \leq \frac{L_p}{p!} \|y - x\|^p, \tag{9}$$

$$\|\nabla^2 f(y) - \nabla^2 \Omega_{x,p}(y)\| \leq \frac{L_p}{(p-1)!} \|y - x\|^{p-1}, \tag{10}$$

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

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

$$\left(a^{1/p} + b^{1/p}\right)^p \leq 2^{p-1}(a + b), \tag{11}$$

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:

$$\min_{x \in \mathbb{E}} \; f(x), \tag{12}$$

where $f(\cdot)$ is a convex function with Lipschitz-continuous pth derivative:

$$\|D^p f(x) - D^p f(y)\| \le L_p \|x - y\|, \quad x, y \in \mathbb{E}, \quad p \ge 1. \tag{13}$$

In this section, we work only with Euclidean norms.

We are going to solve problem (12) by tensor methods. Their performance crucially depends on ability to achieve a significant improvement in the objective function at the current test point.

Definition 2.1 We say that point $T \in \mathbb{E}$ ensures $p\,th - order\ improvement$ of some point $x \in \mathbb{E}$ with factor $c > 0$ if it satisfies the following inequality:

$$\langle \nabla f(T), x - T \rangle \ge c \|\nabla f(T)\|_*^{\frac{p+1}{p}}. \tag{14}$$

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

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

By (8), for $H \ge L_p$, this function gives us an upper estimate for the objective. Moreover, for $H \ge pL_p$ 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:

$$\mathcal{N}^{\gamma}_{p,H}(x) = \{T \in \mathbb{E} : \; \|\nabla \hat{\Omega}_{x,p,H}(T)\|_* \le \gamma \|\nabla f(T)\|_*\}, \tag{15}$$

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 $\nabla \hat{\Omega}_{x,p,H}(x) = \nabla f(x)$. Hence, if $\nabla f(x) \ne 0$, then $x \notin \mathcal{N}^{\gamma}_{p,H}(x)$ for any $\gamma \in [0, 1)$.

These neighborhoods are important by the following reason.

Theorem 2.1 *Let $x \in \mathbb{E}$ and parameters γ, H satisfy the following condition:*

$$\gamma + \frac{L_p}{H} \le \frac{1}{p}. \tag{16}$$

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

$$c_{\gamma,H}(p) \stackrel{\text{def}}{=} \left[\frac{(1-\gamma)p!}{L_p + H}\right]^{\frac{1}{p}}. \tag{17}$$

Consequently, we have

$$f(x) - f(T) \geq c_{\gamma,H}(p)\|\nabla f(T)\|_*^{\frac{p+1}{p}}. \tag{18}$$

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

$$\begin{aligned}
&\|\nabla f(T)\|_*^2 + 2\frac{H}{p!}r^{p-1}\langle \nabla f(T), T - x\rangle + \left(\frac{H}{p!}\right)^2 r^{2p} \\
&= \|\nabla f(T) + \frac{H}{p!}r^{p-1}B(T-x)\|_*^2 \\
&= \|\nabla f(T) - \nabla \Omega_{x,p}(T) + \nabla \hat{\Omega}_{x,p,H}(T)\|_*^2 \\
&\overset{(9)}{\leq} \left(\frac{L_p}{p!}r^p + \gamma\|\nabla f(T)\|_*\right)^2.
\end{aligned}$$

Therefore,

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

In other words,

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

Function $\varkappa(r)$ is convex in $r \geq 0$. Its derivative in r is

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

Note that

$$\|\nabla f(T)\|_* = \|\nabla f(T) - \nabla \Omega_{x,p}(T) + \nabla \hat{\Omega}_{x,p,H}(T) - \frac{H}{p!}r^{p-1}B(T-x)\|_*$$

$$\leq \frac{L_p}{p!} r^p + \gamma \|\nabla f(T)\|_* + \frac{H}{p!} r^p.$$

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

$$\begin{aligned}
\varkappa'(r_*) &= -\frac{(1-\gamma^2)(p-1)p!\|\nabla f(T)\|_*^2}{2H} \cdot \frac{L_p + H}{(1-\gamma)p!\,\|\nabla f(T)\|_*} \\
&\quad + \frac{(p+1)(H^2 - L_p^2)}{2Hp!} \cdot \frac{(1-\gamma)p!\,\|\nabla f(T)\|_*}{L_p + H} - \gamma \frac{L_p}{H} \|\nabla f(T)\|_* \\
&= \|\nabla f(T)\|_* \left[-\frac{(1+\gamma)(p-1)}{2} \left(1 + \frac{L_p}{H}\right) \right. \\
&\quad \left. + \frac{(p+1)(1-\gamma)}{2} \left(1 - \frac{L_p}{H}\right) - \gamma \frac{L_p}{H} \right] \\
&= \|\nabla f(T)\|_* \left[1 - p\gamma - p\frac{L_p}{H} \right] \stackrel{(16)}{\geq} 0.
\end{aligned}$$

So by convexity of $\varkappa(\cdot)$ and $r \geq r_*$, we have $\varkappa(r) \geq \varkappa(r_*)$. Therefore,

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

Inequality (18) is valid since our function is convex:

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

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 corresponding methods.

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:

$$\gamma = \frac{1}{2p}, \quad H \;=\; 2pL_p. \tag{19}$$

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

$$\mathcal{N}_p(x) \stackrel{\text{def}}{=} \mathcal{N}^{1/(2p)}_{p,2pL_p}(x), \quad c_p \stackrel{\text{def}}{=} c_{1/(2p),2pL_p}(p) \;=\; \left[\frac{2p-1}{2p(2p+1)} \frac{p!}{L_p} \right]^{\frac{1}{p}}. \tag{20}$$

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

Corollary 2.1 *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:

$$\boxed{x_{k+1} \in \mathcal{N}_p(x_k), \quad k \geq 0.} \tag{21}$$

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

Theorem 2.2 *Let the sequence $\{x_k\}_{k \geq 0}$ be generated by method (21). Then, for any $k \geq 1$ we have*

$$\begin{aligned} f(x_k) - f^* &\leq \left[\frac{p+1}{k} \left(\frac{1}{c_p} R^{\frac{p+1}{p}}(x_0) + (f(x_0) - f^*)^{1/p} \right) \right]^p \\ &\overset{(11)}{\leq} \left(\frac{2(p+1)}{k} \right)^p \left[\frac{p(2p+1)}{(2p-1)p!} L_p R^{p+1}(x_0) + \frac{1}{2}(f(x_0) - f^*) \right]. \end{aligned} \tag{22}$$

Proof In view of inequality (18), we have $f(x_k) \leq f(x_0)$ for all $k \geq 0$. Therefore,

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

Consequently,

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\overset{(18)}{\geq} c_p \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}} \geq c_p \left(\frac{\langle \nabla f(x_{k+1}), x_{k+1} - x^* \rangle}{R(x_0)} \right)^{\frac{p+1}{p}} \\ &\geq c_p \left(\frac{f(x_{k+1}) - f^*}{R(x_0)} \right)^{\frac{p+1}{p}}. \end{aligned}$$

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 [13], 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 (22). □

Let us present a 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 $p+1$: for all $x, y \in \mathbb{E}$ we have

$$d_{p+1}(y) \geq d_{p+1}(x) + \langle \nabla d_{p+1}(x), y - x \rangle + \frac{1}{p+1} \left(\frac{1}{2}\right)^{p-1} \|y - x\|^{p+1} \quad (23)$$

(see, for example, Lemma 4.2.3 in [14]). Define the sequence

$$A_k = 2 \left(\frac{p+1}{2p} c_p\right)^p \left(\frac{k}{p+1}\right)^{p+1}, \quad a_{k+1} \stackrel{\text{def}}{=} A_{k+1} - A_k, \quad k \geq 0. \quad (24)$$

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

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

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

$$a_{k+1}^{\frac{p+1}{p}} \leq 2^{1/p} \frac{p+1}{2p} c_p \, A_{k+1}, \quad k \geq 0. \quad (25)$$

Inexact pth-Order Accelerated Tensor Method (ATMI$_p$)
Initialization. Choose $x_0 \in \mathbb{E}$. Define coefficients A_k by (24) and function $\psi_0(x) = d_{p+1}(x - x_0)$
Iteration $k \geq 0$. **1.** Compute $v_k = \arg\min_{x \in \mathbb{E}} \psi_k(x)$ 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 $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle]$.

(26)

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

$$\psi_k(x) \leq A_k f(x) + d_{p+1}(x - x_0), \quad x \in \mathbb{E}. \quad (27)$$

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

$$A_k f(x) + d_{p+1}(x - x_0) \stackrel{(27)}{\geq} \psi_k(x) \stackrel{(23)}{\geq} \psi_k^* + \frac{1}{p+1} \left(\frac{1}{2}\right)^{p-1} \|x - v_k\|^{p+1}. \quad (28)$$

Let us prove by induction the following relation:

$$\psi_k^* \geq A_k f(x_k), \quad k \geq 0. \quad (29)$$

For $k = 0$, we have $\psi_0^* = 0$ and $A_0 = 0$. Hence, (29) is valid. Assume it is valid for some $k \geq 0$. Then,

$$\begin{aligned}\psi_{k+1}^* &= \min_{x\in\mathbb{E}} \left\{\psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\right\} \\ &\overset{(28)}{\geq} \min_{x\in\mathbb{E}} \left\{\psi_k^* + \frac{1}{p+1}\left(\frac{1}{2}\right)^{p-1} \|x - v_k\|^{p+1} \right. \\ &\left. + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\right\}.\end{aligned}$$

Note that

$$\begin{aligned}&\psi_k^* + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\ &\overset{(29)}{\geq} A_k f(x_k) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\ &\geq A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), a_{k+1}(x - x_{k+1}) + A_k(x_k - x_{k+1})\rangle \\ &= A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), a_{k+1}(x - v_k) + A_{k+1}(y_k - x_{k+1})\rangle.\end{aligned}$$

Further, in view of inequality $\frac{\alpha}{p+1}\tau^{p+1} - \beta\tau \geq -\frac{p}{p+1}\alpha^{-1/p}\beta^{(p+1)/p}$, $\tau \geq 0$, for all $x \in \mathbb{E}$ we have

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

Finally, since $x_{k+1} \in \mathcal{N}_p(y_k)$, by Corollary 2.1 we get

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

Putting all these inequalities together, we obtain

$$\begin{aligned}\psi_{k+1}^* &\geq A_{k+1} f(x_{k+1}) - \frac{p}{p+1} 2^{\frac{p-1}{p}} \left(a_{k+1}\|\nabla f(x_{k+1})\|_*\right)^{\frac{p+1}{p}} \\ &\quad + A_{k+1} c_p \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}} \\ &= A_{k+1} f(x_{k+1}) + \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}} \left(A_{k+1} c_p - \frac{p}{p+1} 2^{\frac{p-1}{p}} a_{k+1}^{\frac{p+1}{p}}\right) \\ &\overset{(25)}{\geq} A_{k+1} f(x_{k+1}).\end{aligned}$$

Thus, we have proved the following theorem.

Theorem 2.3 *Let sequence $\{x_k\}_{k\geq 0}$ be generated by method (26). Then, for any $k \geq 1$, we have*

$$f(x_k) - f^* \leq \frac{2p+1}{2(2p-1)p!}\left(\frac{2p}{k}\right)^{p+1} \cdot L_p \|x^* - x_0\|^{p+1}. \tag{30}$$

Proof Indeed, in view of relations (27) and (29), we have

$$\begin{aligned} f(x_k) - f^* &\leq \frac{1}{A_k} d_{p+1}(x^* - x_0) \overset{(24)}{=} \frac{1}{2}\left(\frac{2p}{(p+1)c_p}\right)^p \left(\frac{p+1}{k}\right)^{p+1} \\ &\quad \cdot \frac{1}{p+1}\|x^* - x_0\|^{p+1} \\ &= \frac{1}{2}\left(\frac{2p}{c_p}\right)^p \left(\frac{1}{k}\right)^{p+1} \cdot \|x^* - x_0\|^{p+1} = \frac{(2p+1)L_p}{2(2p-1)p!}\left(\frac{2p}{k}\right)^{p+1} \\ &\quad \cdot \|x^* - x_0\|^{p+1}. \end{aligned}$$

□

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:

$$\min_{x \in \operatorname{dom}\psi} \left\{ F(x) \overset{\text{def}}{=} \varphi(x) + \psi(x) \right\}, \tag{31}$$

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 (31), and let $F^* = F(x^*)$.

Let $\varphi(\cdot)$ be non-degenerate with respect to some scaling function $d(\cdot)$:

$$\begin{aligned} \mu_d(\varphi)\beta_d(x,y) \leq \beta_\varphi(x,y) &\overset{(1)}{=} \varphi(y) - \varphi(x) - \langle \nabla\varphi(x), y - x\rangle \\ &\leq L_d(\varphi)\beta_d(x,y), \quad x, y \in \operatorname{dom}\psi, \end{aligned} \tag{32}$$

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 (32):

$$\mu_d(\varphi)\nabla^2 d(x) \preceq \nabla^2\varphi(x) \preceq L_d(\varphi)\nabla^2 d(x), \quad x \in \operatorname{dom}\psi. \tag{33}$$

We are going to solve problem (31) 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

$$\|g_\varphi(x) - \nabla\varphi(x)\|_* \le \delta, \tag{34}$$

where $\delta \ge 0$ 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 (31). For this, we need to assume that function $d(\cdot)$ is *uniformly convex* of degree $p+1$ with $p \ge 1$:

$$\beta_d(x, y) \ge \frac{1}{p+1}\sigma_{p+1}(d)\|x - y\|^{p+1}, \quad x, y \in \operatorname{dom}\psi. \tag{35}$$

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

Choose $x_0 \in \mathbb{E}$. **For** $k \ge 0$ **iterate:**

$$x_{k+1} = \arg\min_{y \in \operatorname{dom}\psi} \left\{\psi(y) + \langle g_\varphi(x_k), y - x_k\rangle + 2L_d(\varphi)\beta_d(x_k, y)\right\}. \tag{36}$$

Lemma 3.1 *Let the approximate gradient $g_\varphi(x_k)$ satisfy the condition (34). Then, for any $x \in \mathbb{E}$ and $k \ge 0$ we have*

$$\beta_d(x_{k+1}, x) \le \left(1 - \tfrac{1}{4}\gamma_d(\varphi)\right)\beta_d(x_k, x) + \tfrac{1}{2L_d(\varphi)}[F(x) - F(x_{k+1})] + \hat{\delta}, \tag{37}$$

where $\hat{\delta} \stackrel{\text{def}}{=} \frac{2p}{p+1}\delta^{\frac{p+1}{p}}\left(\frac{(p+1)(2+\gamma_d(\varphi))}{\sigma_{p+1}(d)\,\gamma_d(\varphi)}\right)^{\frac{1}{p}}$.

Proof The first-order optimality condition defining x_{k+1} is as follows:

$$\langle g_\varphi(x_k) + 2L_d(\varphi)(\nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1}\rangle + \psi(x) \ge \psi(x_{k+1}) \tag{38}$$

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

$$\begin{aligned}
&r_{k+1}(x) - r_k(x)\\
&\quad = \Big(d(x) - d(x_{k+1}) - \langle\nabla d(x_{k+1}), x - x_{k+1}\rangle\Big)\\
&\qquad -\Big(d(x) - d(x_k) - \langle\nabla d(x_k), x - x_k\rangle\Big)\\
&\quad = d(x_k) - \langle\nabla d(x_k), x_k - x_{k+1}\rangle - d(x_{k+1})\\
&\qquad +\langle\nabla d(x_k) - \nabla d(x_{k+1}), x - x_{k+1}\rangle\\
&\quad \stackrel{(38)}{\le} -\beta_d(x_k, x_{k+1}) + \frac{1}{2L_d(\varphi)}\Big[\langle g_\varphi(x_k), x - x_{k+1}\rangle + \psi(x) - \psi(x_{k+1})\Big].
\end{aligned}$$

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

$$\langle\nabla\varphi(x_k), x - x_{k+1}\rangle \stackrel{(32)}{\le} \langle\nabla\varphi(x_k), x_k - x_{k+1}\rangle + \varphi(x) - \varphi(x_k) - \mu_d(\varphi)\beta_d(x_k, x)$$

$$\overset{(32)}{\leq} L_d(\varphi) d(x_k, x_{k+1}) + \varphi(x) - \varphi(x_{k+1}) - \mu_d(\varphi)\beta_d(x_k, x).$$

Hence,

$$\begin{aligned} & r_{k+1}(x) - r_k(x) + \frac{1}{2L_d(\varphi)}[F(x_{k+1}) - F(x)] \\ & \leq \langle g_\varphi(x_k) - \nabla\varphi(x_k), x - x_{k+1}\rangle - \frac{1}{2}\beta_d(x_k, x_{k+1}) - \frac{1}{2}\gamma_d(\varphi)\beta_d(x_k, x) \\ & \overset{(35)}{\leq} -\frac{1}{4}\gamma_d(\varphi) r_k(x) + \langle g_\varphi(x_k) - \nabla\varphi(x_k), x - x_{k+1}\rangle \\ & -\frac{\sigma_{p+1}(d)}{2(p+1)}\Big(\|x_k - x_{k+1}\|^{p+1} + \frac{1}{2}\gamma_d(\varphi)\|x_k - x\|^{p+1}\Big). \end{aligned}$$

Since $\|x\| = \|-x\|$ for all x in $\mathbb{E}$, the minimum 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

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

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

$$\begin{aligned} \bar{\alpha}^{p+1} + \beta^p(1-\bar{\alpha})^{p+1} &= \bar{\alpha}\frac{\beta^p}{(1+\beta)^p} + \frac{\beta^p}{(1+\beta)^{p+1}} \\ &= \frac{\beta^p}{(1+\beta)^p} \overset{(11)}{\geq} \frac{\gamma_d(\varphi)}{2^{p-1}(2+\gamma_d(\varphi))}. \end{aligned}$$

Thus,

$$\begin{aligned} & r_{k+1}(x) - (1 - \frac{1}{4}\gamma_d(\varphi)) r_k(x) + \frac{1}{2L_d(\varphi)}[F(x_{k+1}) - F(x)] \\ & \leq \langle g_\varphi(x_k) - \nabla\varphi(x_k), x - x_{k+1}\rangle - \frac{\sigma_{p+1}(d)\,\gamma_d(\varphi)}{2^p(p+1)(2+\gamma_d(\varphi))}\|x - x_{k+1}\|^{p+1} \\ & \overset{(34)}{\leq} \frac{2p}{p+1}\delta^{\frac{p+1}{p}}\left(\frac{(p+1)(2+\gamma_d(\varphi))}{\sigma_{p+1}(d)\,\gamma_d(\varphi)}\right)^{\frac{1}{p}}. \square \end{aligned}$$

Applying inequality (37) with $x = x^*$ recursively to all $k = 0, \ldots, T-1$, we get the following relation:

$$\begin{aligned} & \beta_d(x_T, x^*) + \frac{1}{2L_d(\varphi)}\sum_{k=0}^{T-1}(1-\gamma)^{T-k-1}[F(x_{k+1}) - F(x^*)] \\ & \leq (1-\gamma)^T \beta_d(x_0, x^*) + S_T\hat{\delta}, \end{aligned} \tag{39}$$

where $\gamma = \frac{1}{4}\gamma_d(\varphi)$, and $S_T = \sum\limits_{k=0}^{T-1}(1-\gamma)^{T-k-1} = \frac{1}{\gamma}\Big(1-(1-\gamma)^T\Big)$.

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

$$F_T^* - F^* \overset{(39)}{\le} \frac{2\gamma(1-\gamma)^T}{1-(1-\gamma)^T} L_d(\varphi)\beta(x_0, x^*) + 2\hat{\delta}L_d(\varphi), \quad T \ge 1. \tag{40}$$

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 convergence rate

$$F_T^* - F^* \overset{(39)}{\le} 2L_d(\varphi)\left(\frac{1}{T}\beta(x_0, x^*) + 2\hat{\delta}\right), \quad T \ge 1. \tag{41}$$

□

In our main application, presented in Sect. 4, we need to generate points with small norm of the gradient. In order to achieve this goal with method (36), we need one more assumption on the scaling function $d(\cdot)$.

From now on, we consider the unconstrained minimization problems. This means that in (31) we have $\psi(x) = 0$ for all $x \in \mathbb{E}$.

Definition 3.1 We call the scaling function $d(\cdot)$ $\underline{norm - dominated}$ on the set $S \subseteq \mathbb{E}$ by some function $\theta_S(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ if there exists a convex function $\theta_S(\cdot)$ with $\theta_S(0) = 0$ such that

$$\beta_d(x, y) \le \theta_S(\|x - y\|) \tag{42}$$

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

Clearly, if function $d(\cdot)$ is norm-dominated by function $\theta_S(\cdot)$ and $\eta_S(\tau) \ge \theta_S(\tau)$ for all $\tau \ge 0$, 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 3.2 *Function $d_4(\cdot)$ is norm-dominated on the Euclidean ball*

$$B_R = \{x \in \mathbb{E} : \|x\| \le R\}$$

by the function

$$\theta_R(\tau) = \frac{1}{4}(\tau^2 + 2R\tau)^2 + \frac{1}{2}R^2\tau^2 \le \frac{1}{2}\tau^4 + \frac{5}{2}R^2\tau^2, \quad \tau \ge 0. \tag{43}$$

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

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

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

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 3.2 can be extended onto all convex polynomial scaling functions.

Norm-dominated scaling functions are important in view of the following.

Lemma 3.3 *Let scaling function $d(\cdot)$ be norm-dominated on the level set*

$$\mathcal{L}_\varphi(\bar{x}) = \{x \in \mathbb{E} : \ \varphi(x) \le \varphi(\bar{x})\}$$

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

$$\varphi(x) - \varphi(x^*) \ge L_d(\varphi)\,\theta^*\left(\frac{1}{L_d(\varphi)}\|\nabla\varphi(x)\|_*\right), \tag{44}$$

where $\theta^*(\tau) = \max_\lambda[\lambda\tau - \theta(\tau)]$.

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

$$
\begin{aligned}
\varphi(y) &\overset{(32)}{\le} \varphi(x) + \langle\nabla\varphi(x), y - x\rangle + L_d(\varphi)\beta_d(x, y) \\
&\overset{(42)}{\le} \varphi(x) + \langle\nabla\varphi(x), y - x\rangle + L_d(\varphi)\theta(\|y - x\|).
\end{aligned}
$$

Therefore,

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

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 Sect. 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

$$\min_{x \in \mathbb{E}} f(x), \tag{45}$$

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 (45) is Lipschitz continuity of the second derivative (see, for example, [12,17]). We are going to replace it by a stronger assumption, using the following fact.

Lemma 4.1 *Let constants $M_2(f)$ and $M_4(f)$ be finite. Then*

$$M_3(f) \le \sqrt{2M_2(f)M_4(f)}. \tag{46}$$

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

$$0 \preceq \nabla^2 f(x - \tau h) \overset{(7)}{=} \nabla^2 f(x) - \tau D^3 f(x)[h] + \tau^2 \int_0^1 (1-\lambda) D^4 f(x + \lambda h)[h]^2 d\lambda$$

$$\preceq \nabla^2 f(x) - \tau D^3 f(x)[h] + \frac{1}{2}\tau^2 M_4(f)\|h\|^2 B.$$

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 get (46). □

Thus, from now on, we assume that

$$L_3(f) \equiv M_4(f) < +\infty. \tag{47}$$

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

In our situation, we can apply to (45) the third-order tensor method ATMI$_3$ (see 26). 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]):

$$-\frac{1}{\xi}\nabla^2 f(x) - \frac{\xi}{2} M_4(f)\|h\|^2 B \preceq D^3 f(x)[h] \preceq \frac{1}{\xi}\nabla^2 f(x) + \frac{\xi}{2} M_4(f)\|h\|^2 B, \tag{48}$$

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 for all $x, y \in \mathbb{E}$, consider the following vector functions:

$$h_y^\tau(x) = \frac{2}{\tau^2}[\nabla f(y + \tau(x - y)) - \nabla f(y) - \tau\nabla^2 f(x)(y - x)] \in \mathbb{E}^*,$$
$$g_y^\tau(x) = \frac{1}{\tau^2}[\nabla f(y + \tau(x - y)) + \nabla f(y - \tau(x - y)) - 2\nabla f(y)] \in \mathbb{E}^*,$$

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

Lemma 4.2 *For any $x, y \in \mathbb{E}$, we have*

$$\|h_y^\tau(x) - D^3 f(y)[x - y]^2\|_* \le \frac{\tau}{3} M_4(f)\|x - y\|^3, \tag{49}$$
$$\|g_y^\tau(x) - D^3 f(y)[x - y]^2\|_* \le \frac{\tau}{3} M_4(f)\|x - y\|^3, \tag{50}$$
$$\|g_y^\tau(x) - D^3 f(y)[x - y]^2\|_* \le \frac{\tau^2}{12} L_4(f)\|x - y\|^4. \tag{51}$$

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

$$\nabla f(y + h) - \nabla f(y) - \nabla^2 f(y)h - \frac{1}{2}D^3 f(y)[h]^2$$
$$\overset{(7)}{=} \frac{1}{2}\int_0^1 (1 - \lambda)^2 D^4 f(y + \lambda h)[h]^3 d\lambda.$$

Applying a uniform upper bound for the fourth derivative to the right-hand side of this representation, we get inequality (49). Further,

$$\nabla f(y - h) - \nabla f(y) + \nabla^2 f(y)h - \frac{1}{2}D^3 f(y)[h]^2$$
$$\overset{(7)}{=} \frac{1}{2}\int_0^1 (1 - \lambda)^2 D^4 f(y - \lambda h)[-h]^3 d\lambda.$$

Adding these two representations, we get

$$g_y^\tau(x) - D^3 f(y)[x - y]^2$$
$$= \frac{\tau}{2}\int_0^1 (1 - \lambda)^2 \Big(D^4 f(y + \lambda\tau(x - y)) - D^4 f(y - \lambda\tau(x - y))\Big)[x - y]^3 d\lambda,$$

and we obtain inequality (50). If the fourth derivative derivative is Lipschitz continuous, then

$$\|g_y^\tau(x) - D^3 f(y)[x-y]^2\|_* \le \frac{\tau}{2}\int_0^1 (1-\lambda)^2 \cdot 2\lambda\tau\|x-y\|^4 L_4(f)d\lambda,$$

and this is inequality (51). □

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

$$\nabla\hat{\Omega}_{y,3,H}(x) = \nabla f(y) + \nabla^2 f(y)h + \frac{1}{2}D^3 f(y)[h]^2 + \frac{H}{6}\|h\|^2 Bh,$$

where $h = x - y$. 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 Sect. 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

$$\begin{aligned}\varphi_k(x) &= \hat{\Omega}_{0,3,6L_3}(x) \overset{\text{def}}{=} f(0) + \langle\nabla f(0), x\rangle + \frac{1}{2}\langle\nabla^2 f(0)x, x\rangle \\ &\quad + \frac{1}{6}D^3 f(0)[x]^3 + \frac{L_3}{4}\|x\|^4. \end{aligned} \tag{52}$$

Thus, our auxiliary problem is as follows:

$$\min_{x\in\mathbb{E}} \varphi_k(x). \tag{53}$$

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

$$\nabla\varphi_k(x) = \nabla f(0) + \nabla^2 f(0)x + \frac{1}{2}D^3 f(0)[x]^2 + L_3\|x\|^2 Bx, \tag{54}$$

$$\begin{aligned}\nabla^2\varphi_k(x) &= \nabla^2 f(0) + D^3 f(0)[x] + L_3\Big(\|x\|^2 B + 2Bxx^*B\Big) \\ &= \nabla^2 f(0) + D^3 f(0)[x] + L_3\nabla^2 d_4(x). \end{aligned} \tag{55}$$

Therefore,

$$\begin{aligned}\nabla^2\varphi_k(x) &\overset{(48)}{\preceq} \left(1 + \frac{1}{\xi}\right)\nabla^2 f(0) + \left(1 + \frac{\xi}{2}\right)L_3\nabla^2 d_4(x), \\ \nabla^2\varphi_k(x) &\overset{(48)}{\succeq} \left(1 - \frac{1}{\xi}\right)\nabla^2 f(0) + \left(1 - \frac{\xi}{2}\right)L_3\nabla^2 d_4(x), \end{aligned} \tag{56}$$

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

$$\rho_k(x) = \frac{1}{2}\langle\nabla^2 f(0)x, x\rangle + L_3 d_4(x), \quad x \in \mathbb{E}. \tag{57}$$

Indeed, applying the relations (56) with $\xi = \sqrt{2}$, we get

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

Thus, we can take

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

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

$$\gamma(\varphi) \stackrel{\text{def}}{=} \frac{\mu_{\rho_k}(\varphi_k)}{L_{\rho_k}(\varphi_k)} = \frac{1}{(1+\sqrt{2})^2} = \frac{1}{3+2\sqrt{2}} > \frac{1}{6}. \tag{58}$$

The second condition for applicability of method (36) is the uniform convexity of the Bregman distance. In our case, this is true since

$$\beta_{\rho_k}(x, y) \geq L_3 \beta_{d_4}(x, y) \stackrel{(23)}{\geq} \frac{1}{16} L_3 \|x - y\|^4, \quad x, y \in \mathbb{E}. \tag{59}$$

Thus, in terms of inequality (35), 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 4.3 *For any $x \in \mathcal{L}_k$, we have*

$$\|x\| \leq 2^{1/3} R_k, \quad \|x_k^*\| \leq R_k \stackrel{\text{def}}{=} 2 \left(\frac{2+\sqrt{2}}{L_3} \|\nabla f(0)\|_* \right)^{\frac{1}{3}}. \tag{60}$$

Proof Indeed,

$$\begin{aligned} \langle \nabla f(0), 0 - x_k^* \rangle &= \langle \nabla \varphi_k(0), 0 - x_k^* \rangle = \varphi_k(0) - \varphi_k^* + \beta_{\varphi_k}(0, x_k^*) \\ &= \beta_{\varphi_k}(x_k^*, 0) + \beta_{\varphi_k}(0, x_k^*) \geq \mu [\beta_{\rho_k}(x_k^*, 0) + \beta_{\rho_k}(0, x_k^*)] \\ &\geq \mu L_3 [\beta_{d_4}(x_k^*, 0) + \beta_{d_4}(0, x_k^*)] \stackrel{(23)}{\geq} 2\mu \frac{L_3}{16} \|x_k^*\|^4. \end{aligned}$$

Consequently, we have the following bound:

$$\|x_k^*\| \leq 2 \left[\frac{2+\sqrt{2}}{L_3} \|\nabla f(0)\|_* \right]^{\frac{1}{3}} = R_k. \tag{61}$$

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

$$\begin{aligned}\langle \nabla\varphi_k(0), 0 - x\rangle &= \varphi_k(0) - \varphi_k(x) + \beta_{\varphi_k}(0, x) \;\geq\; \beta_{\varphi_k}(0, x)\\ &\geq \mu L_3 \beta_{d_4}(0, x) \overset{(23)}{\geq} \frac{\mu L_3}{16}\|x\|^4.\end{aligned}$$

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

The third condition is the possibility of approximating the gradient of function $\varphi_k(\cdot)$. In our case, in view of Lemma 4.2, we can take

$$g_{\varphi_k,\tau}(x) = \nabla f(0) + \nabla^2 f(0)x + \frac{1}{2} g_0^\tau(x) + L_3\|x\|^2 Bx, \tag{62}$$

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

$$\|g_{\varphi_k,\tau}(x) - \nabla\varphi_k(x)\|_* \overset{(50)}{\leq} \frac{\tau}{3} L_3\|x\|^3, \quad x \in \mathbb{E}. \tag{63}$$

Thus, in order to ensure condition (34) 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 (60) 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 (53) by the following one:

$$\min_{x \in S_k} \varphi_k(x), \quad S_k \overset{\text{def}}{=} \{x \in \mathbb{E} : \|x\| \leq R_k\}. \tag{64}$$

In view of Lemma 4.3, the optimal solutions of problems (53) and (64) coincide.

Consider a variant of method (36) with $\psi \equiv 0$ and accuracy $\delta > 0$.

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

For $i \geq 0$ **iterate:**

1. Compute the approximate gradient $g_{\varphi_k,\tau}(x_i)$ by (62).

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

$$x_{i+1} = \arg\min_{x \in S_k} \left\{\langle g_{\varphi_k,\tau}(x_i), x\rangle + 2\left(1 + \frac{1}{\sqrt{2}}\right)\beta_{\rho_k}(x_i, x)\right\}. \tag{65}$$

Note that the auxiliary problem in this method has now an additional ball constraint (64). However, this does not increase significantly its complexity since the Euclidean norm is already present in the objective function.

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

$$\begin{aligned} \|g_{\varphi_k,\tau}(x_i) - \nabla\varphi_k(x_i)\|_* &\overset{(63)}{\leq} \frac{\tau}{3} L_3 R_k^3 \\ &= \frac{\delta L_3}{8(2+\sqrt{2})\|\nabla f(0)\|_*} \frac{8(2+\sqrt{2})}{L_3} \|\nabla f(0)\|_* \;=\; \delta. \end{aligned} \tag{66}$$

This means, in particular, that the sopping criterion at Step 2 of method (65) is correct: if it is satisfied, then

$$\|\nabla\varphi_k(x_i)\|_* \leq \|g_{\varphi_k,\tau}(x_i)\|_* + \delta \;\leq\; \frac{1}{6}\|\nabla f(x_i)\|_*,$$

which implies $x_i \in \mathcal{N}_3(0)$.

Moreover, we can apply Lemma 3.1 to the following objects:

$$d(\cdot) = \rho_k(\cdot), \quad L_{\rho_k}(\varphi_k) = 1 + \frac{1}{\sqrt{2}}, \quad \gamma_{\rho_k}(\varphi_k) = \frac{1}{6}, \quad \sigma_4(\rho_k) = \frac{1}{4}L_3. \tag{67}$$

Therefore, in our case, inequality (37) with $p = 3$ can be rewritten as

$$\begin{aligned} \beta_{\rho_k}(x_{i+1}, x) &\leq \left(1 - \frac{1}{24}\right)\beta_{\rho_k}(x_i, x) + \frac{1}{2+\sqrt{2}}[\varphi_k(x) - \varphi_k(x_{i+1})] + \hat{\delta}, \\ \hat{\delta} &= \frac{3}{2}\delta^{\frac{4}{3}}\left(\frac{208}{L_3}\right)^{\frac{1}{3}} \;<\; \hat{\delta}_+ \overset{\text{def}}{=} \frac{9\delta^{4/3}}{L_3^{1/3}}. \end{aligned} \tag{68}$$

In view of (57), $\beta_{\rho_k}(x_0, x) \leq \frac{1}{2}L_1 R_k^2 + \frac{1}{4}L_3 R_k^4$. Hence, by (40) we have

$$\min_{0\leq i\leq T} \varphi_k(x_i) - \varphi_k^* \leq (2+\sqrt{2}) \left\{ \frac{L_1 R_k^2 + \frac{1}{2}L_3 R_k^4}{6\left[\left(1+\frac{1}{23}\right)^T - 1\right]} + \hat{\delta}_+ \right\}, \quad T \geq 1, \tag{69}$$

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 (65), sufficient for obtaining the following inequality:

$$\varphi_k(\hat{x}_T) - \varphi_k^* \leq 2(2+\sqrt{2})\hat{\delta}_+, \tag{70}$$

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

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

Hence, we have the following bound:

$$T \le T_k(\delta) \stackrel{\text{def}}{=} 24 \ln \left(1 + \frac{2}{3} \left(\frac{1}{\delta} \right)^{4/3} L_3^{1/3} \left[L_1 R_k^2 + \frac{1}{2} L_3 R_k^4 \right] \right). \tag{71}$$

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

$$\|\nabla \varphi_k(\hat{x}_T)\|_* \le \frac{1}{6} \|\nabla f(\hat{x}_T)\|_*. \tag{72}$$

In order to derive this bound from inequality (70) with an appropriate value of $\hat{\delta}_+$, we use the fact that our scaling function $\rho_k(\cdot)$ is norm-dominated. Indeed, in view of Lemma 3.2 and representation (57), this function is norm-dominated on any Euclidean ball B_r by the following function:

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

Hence, in view of Lemma 4.3, our scaling function $\rho_k(\cdot)$ is norm-dominated on the set $\mathcal{L}_k$ by $\theta_{\hat{r}_k}(\cdot)$ with

$$\hat{r}_k = 2^{1/3} R_k. \tag{73}$$

Thus, in order to apply Lemma 3.3, we need to estimate from above the inverse to its conjugate function.

Lemma 4.4 *For any $r > 0$, we have*

$$\left(\theta_r^*\right)^{-1}(\xi) \le \sqrt{2(L_1 + 5L_3 r^2)\xi} + 2L_3^{1/4} \left(\frac{2}{3} \xi \right)^{3/4}, \quad \xi \ge 0. \tag{74}$$

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

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

We need to find $\lambda \ge 0$ 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

$$\lambda = a\tau + b\tau^3. \tag{75}$$

Therefore,

$$\xi = \theta^*(\lambda) \stackrel{(75)}{=} \frac{a}{2}\tau^2(\lambda) + \frac{3b}{4}\tau^4(\lambda)$$

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

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

Hence,

$$\lambda \overset{(75)}{\le} \sqrt{2a\xi} + b^{1/4}\left(\frac{4\xi}{3}\right)^{3/4}.$$

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

Now we can write down the condition for our parameter δ, which ensures the desired inequality (72). Indeed, in view of inequalities (70) and (44), after $T_k(\delta)$ inner steps (see 71) we can guarantee that

$$\|\nabla\varphi_k(\hat{x}_T)\|_* \le L \cdot (\theta^*_{\hat{r}_k})^{-1}\Big(\frac{2}{L}(2+\sqrt{2})\hat{\delta}_+\Big) \;=\; L \cdot (\theta^*_{\hat{r}_k})^{-1}\Big(4\hat{\delta}_+\Big), \tag{76}$$

where $L \overset{(67)}{=} 1 + \frac{1}{\sqrt{2}}$. In order to stop method (65) at this moment, we need to guarantee that the norm of the approximate gradient is small enough. Hence, our condition for parameter δ can be derived from the following reasoning. Since

$$\|g_{\varphi_k,\tau}(\hat{x}_T)\|_* \overset{(66)}{\le} \delta + \|\nabla\varphi_k(\hat{x}_T)\|_* \overset{(76)}{\le} \delta + L \cdot (\theta^*_{\hat{r}_k})^{-1}\Big(4\hat{\delta}_+\Big),$$

in order to satisfy condition $\|g_{\varphi_k,\tau}(\hat{x}_T)\|_* \le \frac{1}{6}\|\nabla f(\hat{x}_T)\| - \delta$, by Lemma 4.4, it is sufficient to satisfy inequality

$$2\delta + 2L\sqrt{2(L_1 + 5L_3\hat{r}_k^2)\hat{\delta}_+} + 2L_3^{1/4}\left(\frac{8}{3}\hat{\delta}_+\right)^{3/4} \le \frac{1}{6}\epsilon_g, \tag{77}$$

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

$$\hat{r}_k \overset{(73)}{=} 2^{4/3}\left(\frac{2+\sqrt{2}}{L_3}\|\nabla f(0)\|_*\right)^{\frac{1}{3}}, \quad \hat{\delta}_+ \overset{(68)}{=} \frac{9\delta^{4/3}}{L_3^{1/3}}.$$

Hence, this inequality can be rewritten in the following form:

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

Using the upper integer bounds on the coefficients, it can be strengthened:

$$24\delta + 21\delta^{2/3}\sqrt{\frac{1}{2L_3^{1/3}}\|\nabla^2 f(0)\| + 36\|\nabla f(0)\|_*^{2/3}} \le \frac{1}{6}\epsilon_g, \tag{78}$$

where we take $L_1 = \|\nabla^2 f(0)\|$ since this corresponds to the actual role of this constant in the complexity analysis of method (65).

This means that, in accordance to (78), we need to choose

$$\delta = O\left(\frac{\epsilon_g^{3/2}}{\|\nabla f(0)\|_*^{1/2} + \|\nabla^2 f(0)\|^{3/2}/L_3^{1/2}}\right). \tag{79}$$

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

Note that all coefficients in the condition (78) are known (provided that we have a good estimate for the Lipschitz constant L_3). Thus, we have

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

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 finiteness of these bounds is discussed in Sect. 5.

Let us write down our inexact algorithmic schemes (21) and (26), employing the inner procedure (65). These methods have only one parameter $\delta > 0$, which must be chosen in accordance to (78). They need also the constant L_3.

We start from the variant of Inexact Basic Tensor Method (21).

Inexact 3rd-Order Tensor Method

Initialization. Given $\delta > 0$, choose $x_0 \in \mathbb{E}$.

Iteration $k \ge 0$**.**

Compute $x_{k+1} \in \mathcal{N}_3(x_k)$ by method (65) with the following settings:

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

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

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

d) Function $\rho_k(x) = \frac{1}{2}\langle\nabla^2 f(x_k)(x - x_k), x - x_k\rangle + L_3 d_4(x - x_k)$. (80)

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 2.2, the rate of convergence of the main process

is as follows:

$$f(x_k) - f^* \le \left(\frac{8}{k}\right)^3 \left[\frac{7}{10} L_3 R^4(x_0) + \frac{1}{2}(f(x_0) - f^*)\right], \quad k \ge 1. \tag{81}$$

Thus, the analytical complexity bound of the method (80) is of the order

$$O\left(R(x_0) \cdot \left(\frac{L_3}{\epsilon_f}\right)^{1/3} \ln \frac{G+H}{\epsilon_g}\right), \tag{82}$$

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 $x_0 \in \mathbb{E}$ and define A_k by (24) with $p = 3$. Define function $\psi_0(x) = d_4(x - x_0)$.

Iteration $k \ge 0$.

1. Compute $v_k = \arg\min_{x \in \mathbb{E}} \psi_k(x)$ 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 (65) with the following settings:

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

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

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

d) 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 $\psi_{k+1}(x) = \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]$. (83)

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 2.3, the rate of convergence of the main process in (83) is as follows:

$$f(x_k) - f^* \le \frac{7}{60}\left(\frac{6}{k}\right)^4 \cdot L_3 \|x_0 - x^*\|^4, \quad k \ge 1. \tag{84}$$

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

$$O\left(\|x_0 - x^*\| \cdot \left(\frac{L_3}{\epsilon_f}\right)^{1/4} \ln \frac{G+H}{\epsilon_g}\right), \tag{85}$$

Recall that method (83) is a second-order scheme.

5 Bounds for the Derivatives

The complexity analysis in Sect. 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:

$$\mathcal{M}_{1,2,4} = \{f \in \mathbb{C}^4(\mathbb{E}) : \; M_1(f) < +\infty, \; M_2(f) < +\infty, \; M_4(f) < +\infty\}. \quad (86)$$

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 (47) and the properties of the minimization schemes.

Indeed, we can easily bound derivatives at test points from a bounded set. Let us present a trivial result, which follows from Taylor formula (7).

Lemma 5.1 *For any* $x \in B_D(x_0) \stackrel{\text{def}}{=} \{x \in \mathbb{E} : \; \|x - x_0\| \le D\}$, *we have*

$$\begin{aligned} \|\nabla f(x)\|_* &\le \|\nabla f(x_0)\|_* + \|\nabla^2 f(x_0)\| D + \frac{1}{2}\|D^3 f(x_0)\| D^2 + \frac{1}{6} M_4(f) D^3, \\ \|\nabla^2 f(x)\| &\le \|\nabla^2 f(x_0)\| + \|D^3 f(x_0)\| D + \frac{1}{2} M_4(f) D^2. \end{aligned} \quad (87)$$

We can use the right-hand sides of inequalities (87) 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 (78), 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 (80). 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 (80) is monotone, for any x_k generated by this scheme, we have

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

Thus, we can take in (87) $D = 2R(x_0)$.

Let us look now at Inexact Accelerated Tensor Method. Actually, for proving the boundedness of sequences of the test points $\{y_k\}_{k\ge 0}$, it is better to consider its monotone variant. The additional Step 4 of this method ensures monotonicity of the sequence $\{f(x_k)\}_{k\ge 0}$.

Monotone Inexact Accelerated 3rd-Order Tensor Method

Initialization. Choose $x_0 \in \mathbb{E}$. Define A_k by (24) with $p = 3$. Define function $\psi_0(x) = d_4(x - x_0)$.

Iteration $k \geq 0$.

1. Compute $v_k = \arg\min_{x \in \mathbb{E}} \psi_k(x)$ 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 (65) with the following settings:

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

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) 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 $x_{k+1} = \arg\min_x \left\{ f(x) : \; x \in \{x_0, \ldots, x_k, \hat{x}_{k+1}\} \right\}$.

(88)

Complexity analysis, presented in Sect. 2, remains also valid for the monotone variant (88). Indeed, in the right-hand side of the relation (29), we can replace point x_k by any point with better value of the objective function.

Lemma 5.2 *Let points* $\{y_k\}_{k \geq 0}$ *be generated by the method* (88)*. Then,*

$$\|y_k - x_0\| \leq (1 + \sqrt{2}) R(x_0), \quad k \geq 0. \tag{89}$$

Proof Indeed, choosing in the relation (28) $p = 3$ 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 (88),

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

Thus, for accelerated method (88) we can take $D = (1 + \sqrt{2})R(x_0)$. □

6 Conclusion

From our results, we conclude that the existing classification of the problem classes, optimization schemes, and complexity bounds is not 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 1st-order methods to functions with Lipschitz-continuous gradients. The 2nd-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 1st-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 1st-order methods only.

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 2nd- and the 3rd-order methods demonstrate the same rate of convergence. So, the correct classification of problem classes and optimization methods must be at least two-parametric. This is, of course, an interesting topic for the further research.

Another interesting question is related to the 1st-order schemes. Indeed, if we managed to accelerate the 2nd-order methods above their "natural" complexity limits, may be there exists a similar possibility for the 1st-order schemes? In our opinion, the answer is negative. Indeed, the lower complexity bounds for the 1st-order methods are supported by a worst-possible *quadratic function*. Quadratic functions already have zero high-order derivatives. Therefore, any assumptions on the high-order derivatives cannot eliminate this bad function from the problem class. For the 2nd-order methods, the worst-case function has *discontinuous* third derivative (see, for example, Section 4.3.1 in [14]). Therefore, assumptions on the fourth derivative can help.

Acknowledgements This paper has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 788368). It was also supported by Multidisciplinary Institute in Artificial intelligence MIAI@Grenoble Alpes (ANR-19-P3IA-0003). The author would like to thank Alexander Gasnikov for discussions. The comments of two anonymous referees were extremely useful.

References

1. Agarwal, N., Hazan, E.: Lower bounds for higher-order convex optimization. In: Proceedings of the 31st Conference On Learning Theory, PMLR, vol. 75, pp. 774–792 (2018)
2. Arjevani, O.S., Shiff, R.: Oracle complexity of second-order methods for smooth convex optimization. Math. Program. **178**(1–2), 327–360 (2019)
3. Baes, M.: Estimate sequence methods: extensions and approximations. Optimization (2009)
4. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first order methods revisited and applications. Math. Oper. Res. **42**, 330–348 (2016)
5. Birgin, E.G., Gardenghi, J.L., Martinez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. **163**, 359–368 (2017)

6. Bubeck, S., Jiang, Q., Lee, Y.T., Li, Y., Sidford, A.: Near-optimal method for highly nonsmooth convex optimization. In: COLT, pp. 492–507 (2019)
7. Gasnikov, A., Gorbunov, E., Kovalev, D., Mohhamed, A., Chernousova, E.: The global rate of convergence for optimal tensor methods in smooth convex optimization. arXiv:1809.00382 (2018)
8. Grapiglia, G.N., Nesterov, Yu.: On inexact solution of auxiliary problems in tensor methods for convex optimization. Optim. Methods Softw. **36**(1), 145–170 (2021)
9. Jiang, B., Wang, H., Zang, S.: An optimal high-order tensor method for convex optimization. In: Conference on Learning Theory, pp. 1799–1801 (2019)
10. Lu, H., Freund, R., Nesterov, Yu.: Relatively smooth convex optimization by first-order methods, and applications. SIOPT **28**(1), 333–354 (2018)
11. Monteiro, R.D.C., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to the second-order methods. SIOPT **23**(2), 1092–1125 (2013)
12. Nesterov, Y.: Accelerating the cubic regularization of Newtons method on convex problems. Math. Program. **112**(1), 159–181 (2008)
13. Nesterov, Y.: Inexact Basic Tensor Methods. CORE DP (# 2019/23) (2019)
14. Nesterov, Y.: Lectures on Convex Optimization. Springer, Berlin (2018)
15. Nesterov, Y.: Implementable tensor methods in unconstrained convex optimization. Math. Program. **186**, 157–183 (2021)
16. Nesterov, Y., Nemirovskii, A.: Interior Point Polynomial Methods in Convex Programming: Theory and Applications. SIAM, Philadelphia (1994)
17. Nesterov, Y., Polyak, B.: Cubic regularization of Newtons method and its global performance. Math. Program. **108**(1), 177–205 (2006)