IMPLEMENTABLE TENSOR METHODS IN UNCONSTRAINED CONVEX OPTIMIZATION

Yurii Nesterov

REPRINT | 3246

UCLouvain

CORE | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email:lidam-library@uclouvain.be

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

Mathematical Programming (2021) 186:157–183
https://doi.org/10.1007/s10107-019-01449-1

FULL LENGTH PAPER

Series A

Implementable tensor methods in unconstrained convex optimization

Yurii Nesterov[1]

Received: 29 March 2018 / Accepted: 4 November 2019 / Published online: 21 November 2019

Abstract

In this paper we develop new tensor methods for unconstrained convex optimization, which solve at each iteration an auxiliary problem of minimizing convex multivariate polynomial. We analyze the simplest scheme, based on minimization of a regularized local model of the objective function, and its accelerated version obtained in the framework of estimating sequences. Their rates of convergence are compared with the worst-case lower complexity bounds for corresponding problem classes. Finally, for the third-order methods, we suggest an efficient technique for solving the auxiliary problem, which is based on the recently developed relative smoothness condition (Bauschke et al. in Math Oper Res 42:330–348, 2017; Lu et al. in SIOPT 28(1):333–354, 2018). With this elaboration, the third-order methods become implementable and very fast. The rate of convergence in terms of the function value for the accelerated third-order scheme reaches the level $O\left(\frac{1}{k^4}\right)$, where k is the number of iterations. This is very close to the lower bound of the order $O\left(\frac{1}{k^5}\right)$, which is also justified in this paper. At the same time, in many important cases the computational cost of one iteration of this method remains on the level typical for the second-order methods.

Keywords High-order methods · Tensor methods · Convex optimization · Worst-case complexity bounds · Lower complexity bounds

Mathematics Subject Classification 90C25 · 90C06 · 65K05

Research results presented in this paper were obtained in the framework of ERC Advanced Grant 788368 and Russian Science Foundation (Grant 17-11-01027).

✉ Yurii Nesterov
Yurii.Neterov@uclouvain.be

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

1 Introduction

Motivation In the last decade, we observe an increasing interest to the complexity analysis of the high-order methods. Starting from the paper [31] containing the first global rate of convergence of Cubic Regularization of Newton Method, it became more and more common to provide the second-order methods with the worst-case complexity bounds on different problem classes (see, for example, [5,11,12]). New efficiency measurements in this field naturally generated a new spectrum of questions, starting from the possibilities to accelerate the second-order methods (see [27]) up to the lower complexity bounds (see [1,2,13,18]) and attempts of constructing the optimal methods [24].

Another possibility of accelerating the minimization processes consists in the increase of the power of oracle. The idea of using the high-order approximations in Optimization is not new. Initially, such approximations were employed in the optimality conditions (see, for example [22]). However, it seems that the majority of attempts of using the high-order tensors in optimization methods failed by the standard obstacle related to the enormous complexity of minimization of nonconvex multivariate polynomials. To the best of our knowledge, the only theoretical analysis of such schemes for convex problems can be found in an unpublished preprint [3], which is concluded by a pessimistic comment on practical applicability of these methods. For nonconvex problems, several recent papers [8–10,14] contain the complexity analysis for high-order methods designed for generating points with small norm of the gradient. For the auxiliary nonconvex optimization problem, these methods need to guarantee a sufficient level for the first-order optimality condition and the local decrease of the objective function. However, for nonconvex functions even this moderate goal is difficult to achieve.

The key observation, which underlies all results of this paper, is that an appropriately regularized Taylor approximation of convex function is a convex multivariate polynomial. This is indeed a very natural property since this regularized approximation usually belongs to the epigraph of convex function. Thus, the auxiliary optimization problem in the high-order (or *tensor*) methods becomes generally solvable by many powerful methods of Convex Optimization. This fact explains our interest to complexity analysis of the simplest tensor scheme (Sect. 2), based on the *convex* regularized Taylor approximation, and to its accelerated version (Sect. 3). The latter method is obtained by the technique of estimating functions (see [25–27]). Therefore it is similar to Algorithm 4.2 in [3]. The main difference consists in the correct choice of parameters ensuring convexity of the auxiliary problem. We show that this algorithm converges with the rate $O((\frac{1}{k})^{p+1})$, where k is the number of iterations and p is the degree of the tensor.

In the next Sect. 4, we derive lower complexity bounds for the tensor methods. We show that the lower bound for the rate of convergence is of the order $O\left((\frac{1}{k})^{\frac{3p+1}{2}}\right)$. This result is better than the bound in [1] and coincide with the bound in [2]. However, it seems that our justification is simpler.

For practical implementations, the most important results are included in Sect. 5, where we discuss an efficient scheme for minimizing the regularized Taylor approximation of degree three. This auxiliary convex problem can be treated in the framework of *relative smoothness condition*. The first element of this approach was introduced in [4], for generalizing the Lipschitz condition for the norm of the gradient. In [21] it was shown that the same extension can be applied to the condition of strong convexity. This second step is important since it leads to linearly convergent methods for functions with nonstandard growth properties. The auxiliary problem with the third-order tensor is a good application of this technique. We show that the corresponding method converges linearly, with the rate depending on an absolute constant. In the end of the section, we argue that the complexity of one iteration of the resulting third-order scheme is often of the same order as that of the second-order methods.

In the last Sect. 6 we discuss the presented results and mention the open problems.

Notations 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 $x \in \mathbb{E}$. Using a self-adjoint positive-definite operator $B : \mathbb{E} \to \mathbb{E}^*$ (notation $B = B^* \succ 0$), we can endow these spaces with *conjugate Euclidean norms*:

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

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

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

the directional derivative of function 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 in the standard way:

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

For example, 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 (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 follows:

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

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

Note that in general, we have (see, for example, Appendix 1 in [30])

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

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

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

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:

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

Sometimes, if an ambiguity can arise, we us notation $L_p(f)$.

Assuming that $f \in \mathcal{F}_p$ and $L_p < +\infty$, by the standard integration arguments we can bound the residual between function value and its Taylor approximation:

$$|f(y) - \Phi_{x,p}(y)| \le \tfrac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad x, y \in \mathrm{dom}\, f. \tag{1.5}$$

If $p \ge 2$, then applying the same reasoning for functions $\langle \nabla f(\cdot), h \rangle$ and $\langle \nabla^2 f(\cdot) h, h \rangle$ with direction $h \in \mathbb{E}$ being fixed, we get the following guarantees:

$$\|\nabla f(y) - \nabla \Phi_{x,p}(y)\|_* \leq \frac{L_p}{p!}\|y - x\|^p, \tag{1.6}$$

$$\|\nabla^2 f(y) - \nabla^2 \Phi_{x,p}(y)\| \leq \frac{L_p}{(p-1)!}\|y - x\|^{p-1}, \tag{1.7}$$

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

2 Convex tensor approximations

In our methods, we use the following *power prox function*

$$d_p(x) = \frac{1}{p}\|x\|^p, \quad p \geq 2. \tag{2.1}$$

Note that

$$\begin{aligned} \nabla d_p(x) &= \|x\|^{p-2} Bx, \\ \nabla^2 d_p(x) &= (p-2)\|x\|^{p-4} Bxx^*B + \|x\|^{p-2} B \\ &\geq \|x\|^{p-2} B. \end{aligned} \tag{2.2}$$

All results of this paper are based on the following observation. From now on, for the sake of simplicity, we assume that dom $f = \mathbb{E}$.

Theorem 1 *Let $f \in \mathcal{F}_p$ with $p \geq 2$ and $L_p < +\infty$. Then for any $x, y \in \mathbb{E}$ we have*

$$\nabla^2 f(y) \preceq \nabla^2 \Phi_{x,p}(y) + \frac{L_p}{(p-1)!}\|y - x\|^{p-1} B. \tag{2.3}$$

Moreover, for any $M \geq L_p$ and any $x \in \mathbb{E}$, function[1]

$$\Omega_{x,p,M}(y) = \Phi_{x,p}(y) + \frac{M}{(p-1)!} d_{p+1}(y - x) \tag{2.4}$$

is convex and

$$f(y) \leq \Omega_{x,p,M}(y), \quad y \in \mathbb{E}. \tag{2.5}$$

Proof Let us fix arbitrary x and y from dom f. Then for any direction $h \in \mathbb{E}$ we have

$$\begin{aligned} \langle (\nabla^2 f(y) - \nabla^2 \Phi_{x,p}(y))h, h\rangle &\leq \|\nabla^2 f(y) - \nabla^2 \Phi_{x,p}(y)\| \cdot \|h\|^2 \\ &\overset{(1.7)}{\leq} \frac{L_p}{(p-1)!}\|y - x\|^{p-1}\|h\|^2, \end{aligned}$$

[1] In our notation, the approximation function in [3] was chosen as $\Phi_{x,p}(y) + \frac{M}{p!} d_{p+1}(y - x)$. Thus, we cannot guarantee that this polynomial is convex.

and this is (2.3). Further,

$$\begin{aligned} 0 \preceq \nabla^2 f(y) &\overset{(2.3)}{\preceq} \nabla^2 \Phi_{x,p}(y) + \tfrac{L_p}{(p-1)!}\|y-x\|^{p-1} B \\ &\preceq \nabla^2 \Phi_{x,p}(y) + \tfrac{M}{(p-1)!}\|y-x\|^{p-1} B \\ &\overset{(2.2)}{\preceq} \nabla^2 \Phi_{x,p}(y) + \tfrac{M}{(p-1)!}\nabla^2 d_{p+1}(y-x) \\ &\overset{(2.4)}{=} \nabla^2 \Omega_{x,p,M}(y). \end{aligned}$$

Thus, function $\Omega_{x,p,M}(\cdot)$ is convex. Finally,

$$\begin{aligned} f(y) &\overset{(1.5)}{\leq} \Phi_{x,p}(y) + \tfrac{L_p}{(p+1)!}(p+1) d_{p+1}(y-x) \\ &\leq \Phi_{x,p}(y) + \tfrac{M}{p!} d_{p+1}(y-x) \\ &\leq \Omega_{x,p,M}(y). \end{aligned}$$

□

The statements of Theorem 1 explain our interest to the following point:

$$T_{p,M}(x) \in \operatorname{Arg}\min_{y \in \mathbb{E}} \Omega_{x,p,M}(y) \tag{2.6}$$

with $M \geq L_p$. We are going to use such points for solving the problem

$$f_* = \min_{x \in \mathbb{E}} f(x), \tag{2.7}$$

starting from some point $x_0 \in \mathbb{E}$, where $f \in \mathcal{F}_p$ and $L_p < +\infty$. We assume that there exists at least one solution x_* of this problem and that the level sets of f are bounded:

$$\begin{aligned} \max_{x \in \mathcal{L}(x_0)} \|x - x_*\| &\leq D < +\infty, \\ \mathcal{L}(x_0) &\overset{\text{def}}{=} \{x \in \mathbb{E} : \ f(x) \leq f(x_0)\}. \end{aligned} \tag{2.8}$$

In this case, in view of (2.5), the level sets of function $\Omega_{x,p,M}(\cdot)$ are also bounded. Therefore, point $T = T_{p,M}(x)$ is well defined. It satisfies the following first-order optimality condition:

$$\nabla \Phi_{x,p}(T) + \tfrac{M}{(p-1)!}\|T-x\|^{p-1} B(T-x) = 0. \tag{2.9}$$

Multiplying this equality by $T - x$, we get

$$\frac{M}{(p-1)!}\|T - x\|^{p+1} = \langle \nabla \Phi_{x,p}(T), x - T \rangle. \tag{2.10}$$

Denote $r_{p,M}(x) = \|x - T_{p,M}(x)\|$.

Lemma 1 *For any $x \in \mathbb{E}$ and $M \geq L_p$ we have*

$$f(T_{p,M}(x)) \leq \min_{y \in \mathbb{E}} \left\{ f(y) + \frac{pM + L_p}{(p+1)!}\|y - x\|^{p+1} \right\}, \tag{2.11}$$

$$\|\nabla f(T_{p,M}(x))\|_* \leq \frac{pM + L_p}{p!}\|x - T_{p,M}(x)\|^p, \tag{2.12}$$

$$\langle \nabla f(T_{p,M}(x)), x - T_{p,M}(x) \rangle \geq \frac{(p-1)!}{2M r_{p,M}^{p-1}(x)}\|\nabla f(T_{p,M}(x))\|_*^2 + \frac{M^2 - L_p^2}{2M(p-1)!} r_{p,M}^{p+1}(x). \tag{2.13}$$

First two inequalities of this lemma are already known (see, for example, [8]). We provide them with a simple proof for the reader convenience.

Proof Denote $T = T_{p,M}(x)$ and $r = \|x - T\|$. Then,

$$\begin{aligned} f(T) &\overset{(2.5)}{\leq} \min_{y \in \mathbb{E}} \Omega_{x,p,M}(y) \;=\; \min_{y \in \mathbb{E}} \left\{ \Phi_{p,x}(y) + \frac{pM}{(p+1)!}\|y - x\|^{p+1} \right\} \\ &\overset{(1.5)}{\leq} \min_{y \in \mathbb{E}} \left\{ f(y) + \frac{pM + L_p}{(p+1)!}\|y - x\|^{p+1} \right\}. \end{aligned}$$

and this is inequality (2.11). Further,

$$\|\nabla f(T) - \nabla \Phi_{x,p}(T)\|_* \overset{(1.6)}{\leq} \frac{L_p}{p!} r^p. \tag{2.14}$$

Therefore, we get the following bound:

$$\begin{aligned} \|\nabla f(T)\|_* \;&\leq\; \|\nabla f(T) - \nabla \Phi_{x,p}(T)\|_* + \|\nabla \Phi_{x,p}(T)\|_* \\ &\overset{(2.14)}{\leq} \frac{L_p}{p!} r^p + \|\nabla \Phi_{x,p}(T)\|_* \\ &\overset{(2.9)}{=} \left(\frac{L_p}{p!} + \frac{M}{(p-1)!} \right) r^p, \end{aligned}$$

which leads to (2.12). Finally,

$$\|\nabla f(T) + \tfrac{M r^{p-1}}{(p-1)!} B(T - x)\|_* \overset{(2.9)}{=} \|\nabla f(T) - \nabla \Phi_{x,p}(T)\|_* \overset{(2.14)}{\leq} \frac{L_p}{(p-1)!} r^p.$$

Squaring both sides of this bound, we get:

$$\|\nabla f(T)\|_*^2 + \frac{2Mr^{p-1}}{(p-1)!}\langle \nabla f(T), T - x\rangle + \frac{M^2 r^{2p}}{[(p-1)!]^2} \le \frac{L_{p+1}^2 r^{2p}}{[(p-1)!]^2},$$

and this is (2.13). □

Corollary 1 *For any $x \in \mathbb{E}$ and $M \ge L_p$, we have*

$$\langle \nabla f(T_{p,M}(x), x - T_{p,M}(x)\rangle \ge \frac{c(p)}{M}[M^2 - L_p^2]^{\frac{p-1}{2p}} \|\nabla f(T_{p,M}(x))\|_*^{\frac{p+1}{p}}, \quad (2.15)$$

where $c(p) = \frac{p}{p-1}\left[\frac{p-1}{p+1}\right]^{\frac{1-p}{2p}}[(p-1)!]^{\frac{1}{p}}$.

Proof Indeed, in view of inequality (2.13), we have the following bound:

$$\langle \nabla f(T), x - T\rangle \ge \frac{a}{\tau} + b\tau^\alpha,$$

where $a = \frac{(p-1)!}{2M}\|\nabla f(T_{p,M}(x))\|_*^2$, $b = \frac{M^2 - L_p^2}{2M(p-1)!}$, $\tau = r_{p,M}^{p-1}(x)$, and $\alpha = \frac{p+1}{p-1}$. Note that

$$\min_{\tau > 0}\left\{\frac{a}{\tau} + b\tau^\alpha\right\} = (1+\alpha)\left(\frac{a}{\alpha}\right)^{\frac{\alpha}{1+\alpha}} b^{\frac{1}{1+\alpha}}.$$

It remains to substitute in this bound the values of our parameters a, b, and α. □

Let us estimate now the rate of convergence of the following process:

$$\boxed{x_{t+1} = T_{p,M}(x_t),\ t \ge 0} \quad (2.16)$$

where $M \ge L_p$. Thus, in view of Theorem 1, point x_{t+1} is a solution to the auxiliary convex problem (2.6).

Theorem 2 *Let sequence $\{x_t\}_{t\ge 0}$ be generated by method (2.16) as applied to problem (2.7). Then for all $t \ge 0$ we have $f(x_{t+1}) \le f(x_t)$. At the same time,*

$$f(x_t) - f_* \le \frac{(pM+L_p)D^{p+1}}{(p+1)!\left(1+(t-1)\left(\frac{1}{p+1}\right)^{\frac{p+1}{p}}\right)^p} \le \frac{(pM+L_p)D^{p+1}}{p!}\left(\frac{p+1}{t}\right)^p, \quad t \ge 1. \quad (2.17)$$

Proof In view of inequality (2.5), method (2.16) is monotone. Hence, for all $t \ge 0$ we have

$$\|x_t - x_*\| \le D. \quad (2.18)$$

Let us prove the first inequality in (2.17). First of all, let us estimate the difference $f(x_1) - f_*$. We have

$$f(x_1) \overset{(2.11)}{\leq} \min_{y \in \mathbb{E}} \left\{ f(y) + \tfrac{pM+L_p}{(p+1)!} \|y - x_0\|^{p+1} \right\} \overset{(2.18)}{\leq} f_* + \tfrac{(pM+L_p)D^{p+1}}{(p+1)!},$$

and this is (2.17) for $t = 1$.

Further, for any $t \geq 1$, we have

$$\begin{aligned} f(x_{t+1}) &\overset{(2.11)}{\leq} \min_{y \in \mathbb{E}} \left\{ f(y) + \tfrac{pM+L_p}{(p+1)!} \|y - x_t\|^{p+1} \right\} \\ &\overset{(2.18)}{\leq} \min_{\alpha \in [0,1]} \left\{ f(x_t + \alpha(x_* - x_t)) + \tfrac{(pM+L_p)D^{p+1}}{(p+1)!} \alpha^{p+1} \right\} \\ &\leq \min_{\alpha \in [0,1]} \left\{ f(x_t) - \alpha(f(x_t) - f_*) + \tfrac{(pM+L_p)D^{p+1}}{(p+1)!} \alpha^{p+1} \right\}. \end{aligned}$$

The minimum of the above objective in $\alpha \geq 0$ is achieved for

$$\alpha_* = \left(\tfrac{(f(x_t)-f_*)p!}{(pM+L_p)D^{p+1}} \right)^{\frac{1}{p}} \leq \left(\tfrac{(f(x_1)-f_*)p!}{(pM+L_p)D^{p+1}} \right)^{\frac{1}{p}} \overset{(2.17)}{\leq} \left(\tfrac{1}{p+1} \right)^{\frac{1}{p}} < 1.$$

Thus, we conclude that

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \alpha_* \left(f(x_t) - f_* - \tfrac{(pM+L_p)D^{p+1}}{(p+1)!} \alpha_*^p \right) \\ &= f(x_t) - \tfrac{p\,\alpha_*}{p+1} (f(x_t) - f_*). \end{aligned}$$

Denoting $\delta_t = f(x_t) - f_*$, we get the following estimate:

$$\delta_t - \delta_{t+1} \geq C \delta_t^{\frac{p+1}{p}}, \quad t \geq 1,$$

where $C = \frac{p}{p+1} \left(\frac{p!}{(pM+L_p)D^{p+1}} \right)^{\frac{1}{p}}$. Thus, for $\mu_t = C^p \delta_t$, the recursive inequality is as follows:

$$\mu_t - \mu_{t+1} \geq \mu_t^{\frac{p+1}{p}}, \quad t \geq 1.$$

Then,

$$\begin{aligned}\frac{1}{\mu_{t+1}^{1/p}} - \frac{1}{\mu_t^{1/p}} = \frac{\mu_t^{1/p} - \mu_{t+1}^{1/p}}{\mu_{t+1}^{1/p}\mu_t^{1/p}} &= \frac{1}{\mu_{t+1}^{1/p}\mu_t^{1/p}}\left(\mu_t^{1/p} - \mu_t^{1/p}\left(1 + \frac{\mu_{t+1}-\mu_t}{\mu_t}\right)^{1/p}\right) \\ &\geq \frac{1}{\mu_{t+1}^{1/p}\mu_t^{1/p}}\left(\mu_t^{1/p} - \mu_t^{1/p}\left(1 + \frac{\mu_{t+1}-\mu_t}{p\mu_t}\right)\right) = \frac{\mu_t - \mu_{t+1}}{p\mu_t\mu_{t+1}^{1/p}} \\ &\geq \frac{\mu_t^{1/p}}{p\mu_{t+1}^{1/p}} \geq \frac{1}{p}.\end{aligned}$$

This means that $\frac{1}{\mu_t} \geq \left(\frac{1}{\mu_1^{1/p}} + \frac{t-1}{p}\right)^p$. Note that

$$\frac{1}{\mu_1^{1/p}} = \frac{1}{C\delta_1^{1/p}} = \frac{p+1}{p}\left(\frac{(pM+L_p)D^{p+1}}{p!(f(x_1)-f_*)}\right)^{\frac{1}{p}} \overset{(2.17)}{\geq} \frac{1}{p}(p+1)^{\frac{p+1}{p}}.$$

Therefore,

$$\begin{aligned}\delta_t = C^{-p}\mu_t &= \left(\frac{p+1}{p}\right)^p \frac{(pM+L_p)D^{p+1}}{p!}\mu_t \\ &\leq \left(\frac{p+1}{p}\right)^p \frac{(pM+L_p)D^{p+1}}{p!}\left(\frac{1}{p}(p+1)^{\frac{p+1}{p}} + \frac{t-1}{p}\right)^{-p} \\ &= \frac{(pM+L_p)D^{p+1}}{p!}\left((p+1)^{1/p} + \frac{t-1}{p+1}\right)^{-p},\end{aligned}$$

and this is (2.17). □

3 Accelerated tensor methods

In order to accelerate method (2.16), we apply a variant of the *estimating sequences technique*, which becomes a standard tool for accelerating the usual Gradient and Second-Order Methods (see, for example, [25–27]). In our situation, this idea can be applied to tensor methods in the following way.

For solving the problem (2.7), we choose a constant $M \geq L_p$ and recursively update the following sequences.

- Sequence of estimating functions

$$\psi_k(x) = \ell_k(x) + \frac{C}{p!}d_{p+1}(x - x_0), \quad k = 1, 2, \ldots, \tag{3.1}$$

  where $\ell_k(x)$ are linear functions in $x \in \mathbb{E}$, and C is a positive parameter.
- Minimizing sequence $\{x_k\}_{k=1}^{\infty}$.

- Sequence of scaling parameters $\{A_k\}_{k=1}^{\infty}$:

$$A_{k+1} \stackrel{\text{def}}{=} A_k + a_k, \quad k = 1, 2, \ldots.$$

For these objects, we are going to maintain the following relations:

$$\left.\begin{array}{l} \mathcal{R}_k^1 : A_k f(x_k) \le \psi_k^* \equiv \min\limits_{x \in \mathbb{E}} \psi_k(x), \\ \\ \mathcal{R}_k^2 : \psi_k(x) \le A_k f(x) + \frac{pM+L_p+C}{p!} d_{p+1}(x - x_0), \ \forall x \in \mathbb{E}. \end{array}\right\}, \quad k \ge 1. \quad (3.2)$$

Let us ensure that relations (3.2) hold for $k = 1$. We choose

$$x_1 = T_{p,M}(x_0), \quad \ell_1(x) \equiv f(x_1), \ x \in \mathbb{E}, \quad A_1 = 1. \quad (3.3)$$

Then $\psi_1^* = f(x_1)$, so $\mathcal{R}_1^1$ holds. On the other hand, in view of definition (3.1), we get

$$\begin{aligned} \psi_1(x) &= f(x_1) + \tfrac{C}{p!} d_{p+1}(x - x_0) \\ &\overset{(2.11)}{\le} \min_{y \in \mathbb{E}} \left[f(y) + \tfrac{pM+L_p}{(p+1)!} \|y - x_0\|^{p+1} \right] + \tfrac{C}{p!} d_{p+1}(x - x_0), \end{aligned}$$

and $\mathcal{R}_1^2$ follows.

Assume now that relations (3.2) hold for some $k \ge 1$. Denote

$$v_k = \arg\min_{x \in \mathbb{E}} \psi_k(x).$$

Let us choose some $a_k > 0$ and $M \ge L_p$. Define

$$\begin{aligned} &\alpha_k = \tfrac{a_k}{A_k + a_k}, \quad y_k = (1 - \alpha_k) x_k + \alpha_k v_k, \quad x_{k+1} = T_{p,M}(y_k), \\ &\psi_{k+1}(x) = \psi_k(x) + a_k [f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle]. \end{aligned} \quad (3.4)$$

In view of $\mathcal{R}_k^2$ and convexity of f, for any $x \in \mathbb{E}$ we have

$$\begin{aligned} \psi_{k+1}(x) &\le A_k f(x) + \frac{pM + L_p + C}{p!} d_{p+1}(x - x_0) \\ &\quad + a_k [f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle] \\ &\le (A_k + a_k) f(x) + \frac{pM + L_p + C}{p!} d_{p+1}(x - x_0), \end{aligned}$$

and this is $\mathcal{R}_{k+1}^2$. Let us show now that, for the appropriate choices of a_k, C and M, relation $\mathcal{R}_{k+1}^1$ is also valid.

Indeed, in view of $\mathcal{R}_k^1$ and Lemma 4 in [27], for any $x \in \mathbb{E}$, we have

$$\begin{aligned}\psi_k(x) \equiv \ell_k(x) + \tfrac{C}{p!} d_{p+1}(x - x_0) \;&\geq\; \psi_k^* + \tfrac{C}{(p+1)!} \cdot (\tfrac{1}{2})^{p-1} \|x - v_k\|^{p+1} \\ &\geq A_k f(x_k) + \tfrac{C}{(p+1)!} \cdot (\tfrac{1}{2})^{p-1} \|x - v_k\|^{p+1}.\end{aligned} \tag{3.5}$$

Denote $\gamma_p = \frac{C}{p!} \cdot (\frac{1}{2})^{p-1}$. Then,

$$\begin{aligned}\psi_{k+1}^* &= \min_{x \in \mathbb{E}} \{\psi_k(x) + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\} \\ &\overset{(3.5)}{\geq} \min_{x \in \mathbb{E}} \left\{ A_k f(x_k) + \frac{\gamma_p}{p+1} \|x - v_k\|^{p+1} + a_k[f(x_{k+1}) \right. \\ &\quad \left. + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \right\} \\ &\geq \min_{x \in \mathbb{E}} \left\{ (A_k + a_k) f(x_{k+1}) + A_k \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \right. \\ &\quad \left. + a_k \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \frac{\gamma_p}{p+1} \|x - v_k\|^{p+1} \right\} \\ &\overset{(3.4)}{=} \min_{x \in \mathbb{E}} \left\{ A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), A_{k+1} y_k - a_k v_k - A_k x_{k+1}\rangle \right. \\ &\quad \left. + a_k \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \frac{\gamma_p}{p+1} \|x - v_k\|^{p+1} \right\} \\ &= \min_{x \in \mathbb{E}} \left\{ A_{k+1} f(x_{k+1}) + A_{k+1} \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \right. \\ &\quad \left. + a_k \langle \nabla f(x_{k+1}), x - v_k\rangle + \frac{\gamma_p}{p+1} \|x - v_k\|^{p+1} \right\}.\end{aligned}$$

Further, if we choose $M \geq L_p$, then by inequality (2.15) we have

$$\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \;\geq\; \tfrac{c(p)}{M} [M^2 - L_p^2]^{\frac{p-1}{2p}} \, \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}}.$$

Hence, our choice of parameters must ensure the following inequality:

$$\begin{aligned}&A_{k+1} \frac{c(p)}{M} [M^2 - L_p^2]^{\frac{p-1}{2p}} \, \|\nabla f(x_{k+1})\|_*^{\frac{p+1}{p}} + a_k \langle \nabla f(x_{k+1}), x - v_k\rangle \\ &\quad + \frac{\gamma_p}{p+1} \|x - v_k\|^{p+1} \geq 0,\end{aligned}$$

for all $x \in \mathbb{E}$. Minimizing this expression in $x \in \mathbb{E}$, we come to the following condition:

$$A_{k+1} \tfrac{c(p)}{M} [M^2 - L_p^2]^{\frac{p-1}{2p}} \geq \tfrac{p}{p+1} \left(\tfrac{1}{\gamma_p}\right)^{\frac{1}{p}} a_k^{\frac{p+1}{p}}.$$

Substituting in this inequality expressions for corresponding constants, we obtain

$$A_{k+1}\tfrac{p}{p+1}\left[\tfrac{p-1}{p+1}\right]^{\frac{1-p}{2p}}[(p-1)!]^{\frac{1}{p}}\tfrac{1}{M}[M^2-L_p^2]^{\frac{p-1}{2p}} \geq \tfrac{p}{p+1}\left[\tfrac{p!}{C}2^{p-1}\right]^{\frac{1}{p}} a_k^{\frac{p+1}{p}}.$$

After cancellations, we get

$$A_{k+1}\sqrt{1-\tfrac{L_p^2}{M^2}}\left(\tfrac{C^2}{M^2-L_p^2}\right)^{\frac{1}{2p}} \geq 2a_k^{\frac{p+1}{p}}\left(\tfrac{p}{2}\sqrt{\tfrac{p+1}{p-1}}\right)^{\frac{1}{p}}\sqrt{\tfrac{p+1}{p-1}}. \tag{3.6}$$

For the sake of notation, let us choose

$$C = \tfrac{p}{2}\sqrt{\tfrac{(p+1)}{(p-1)}(M^2-L_p^2)}. \tag{3.7}$$

Then, inequality (3.6) becomes simpler:

$$A_{k+1} \geq 2\sqrt{\tfrac{(p+1)M^2}{(p-1)(M^2-L_p^2)}}\, a_k^{\frac{p+1}{p}}. \tag{3.8}$$

Let us choose some $\alpha > 0$ and define

$$\begin{aligned} A_k &= \tfrac{1}{\alpha}k^{p+1}, \\ a_k &= A_{k+1} - A_k = \tfrac{1}{\alpha}((k+1)^{p+1} - k^{p+1}). \end{aligned} \tag{3.9}$$

Note that for any $k \geq 0$, using trivial inequality $(1-\tau)^p \geq 1 - p\tau, 0 \leq \tau \leq 1$, we get

$$\begin{aligned} a_k \cdot A_{k+1}^{-\frac{p}{p+1}} &= \alpha^{-\frac{1}{p+1}} \cdot \tfrac{(k+1)^{p+1}-k^{p+1}}{(k+1)^p} = \alpha^{-\frac{1}{p+1}}\left(k+1-k\left(1-\tfrac{1}{k+1}\right)^p\right) \\ &\leq \alpha^{-\frac{1}{p+1}}\left(1+\tfrac{kp}{k+1}\right) \leq \alpha^{-\frac{1}{p+1}}(1+p). \end{aligned}$$

Thus, $A_{k+1} \geq \alpha^{\frac{1}{p}}\left(\frac{1}{p+1}\right)^{\frac{p+1}{p}} a_k^{\frac{p+1}{p}}$. Now, if we choose

$$\alpha = (p+1)^{p+1}\left[\tfrac{4(p+1)M^2}{(p-1)(M^2-L_p^2)}\right]^{\frac{p}{2}}, \tag{3.10}$$

then $\alpha^{-\frac{1}{p+1}}(1+p) = \left[\frac{(p-1)(M^2-L_p^2)}{4(p+1)M^2}\right]^{\frac{p}{2(p+1)}}$, and inequality (3.8) is satisfied for all $k \geq 0$.

Now we are ready to write down the accelerated tensor method. Define

$$A_k = \left[\frac{(p-1)(M^2-L_p^2)}{4(p+1)M^2}\right]^{\frac{p}{2}} \left(\frac{k}{p+1}\right)^{p+1}, \quad a_k = A_{k+1} - A_k, \quad k \geq 0. \tag{3.11}$$

Accelerated Tensor Method

Initialization: Choose $x_0 \in \mathbb{E}$ and $M > L_p$. Compute $x_1 = T_{p,M}(x_0)$.

Define $C = \frac{p}{2}\sqrt{\frac{(p+1)}{(p-1)}(M^2 - L_p^2)}$ and $\psi_1(x) = f(x_1) + \frac{C}{p!}d_{p+1}(x - x_0)$.

Iteration k**,** $(k \geq 1)$**:**

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 $x_{k+1} = T_{p,M}(y_k)$ and update

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

(3.12)

The above discussion proves the following theorem.

Theorem 3 *Let the sequence* $\{x_k\}_{k=1}^{\infty}$ *be generated by method* (3.12) *as applied to the problem* (2.7). *Then for any* $k \geq 1$ *we have:*

$$f(x_k) - f(x^*) \leq \frac{pM+L_p+C}{(p+1)!}\left[\frac{4(p+1)M^2}{(p-1)(M^2-L_p^2)}\right]^{\frac{p}{2}} \left(\frac{p+1}{k}\right)^{p+1} \|x_0 - x^*\|^{p+1}. \tag{3.13}$$

Proof Indeed, we have shown that

$$A_k f(x_k) \overset{\mathcal{R}_k^1}{\leq} \psi_k^* \overset{\mathcal{R}_k^2}{\leq} A_k f(x^*) + \frac{pM+L_p+C}{(p+1)!}\|x_0 - x^*\|^{p+1}.$$

Thus, (3.13) follows from (3.11). □

Note that the point v_k can be found in (3.12) by a closed-form expression. Consider

$$s_k = \nabla \ell_k(x).$$

Since function $\ell_k(\cdot)$ is linear, this vector is the same for all $x \in \mathbb{E}$. Therefore, the point v_k is a solution of the following minimization

$$\min_{x \in \mathbb{E}} \left\{ \langle s_k, x \rangle + \tfrac{C}{(p+1)!} \|x - x_0\|^{p+1} \right\}.$$

The first-order optimality condition for this problem is as follows:

$$s_k + \tfrac{C}{p!} \|x - x_0\|^{p-1} B(x - x_0) = 0.$$

Thus, we get the following closed-form expression for its solution:

$$v_k = x_0 - \left(\frac{p!}{C \|s_k\|_*^{p-1}} \right)^{\frac{1}{p}} \cdot B^{-1} s_k.$$

4 Lower complexity bounds for tensor methods

For constructing functions, which are difficult for all tensor methods, it is convenient to assume that $\mathbb{E} = \mathbb{E}^* = \mathbb{R}^n$, and $B = I_n$, the identity $n \times n$-matrix. Thus, in this section we work with the standard Euclidean norm

$$\|x\| = \left[\sum_{i=1}^n (x^{(i)})^2 \right]^{1/2}, \quad x \in \mathbb{R}^n.$$

For an integer parameter $p \geq 1$, define the following function:

$$\eta_{p+1}(x) = \tfrac{1}{p+1} \sum_{i=1}^n |x^{(i)}|^{p+1}, \quad x \in \mathbb{R}^n.$$

Clearly, $\eta_{p+1} \in \mathcal{F}_p$. On the other hand, for any x and $h \in \mathbb{R}^n$, we have

$$D^k \eta_{p+1}(x)[h]^k = \tfrac{p!}{(p+1-k)!} \sum_{i=1}^n |x^{(i)}|^{p+1-k} (h^{(i)})^k, \quad \text{if } k \text{ is even},$$

$$D^k \eta_{p+1}(x)[h]^k = \tfrac{p!}{(p+1-k)!} \sum_{i=1}^n |x^{(i)}|^{p-k} x^{(i)} (h^{(i)})^k, \text{ if } k \text{ is odd}.$$

Therefore, for all x, y, and h from $\mathbb{R}^n$, by Cauchy-Schwartz inequality we have

$$\begin{aligned} |D^p \eta_{p+1}(x)[h]^p - D^p \eta_{p+1}(y)[h]^p| &\leq p! \|x - y\| \left[\sum_{i=1}^n (h^{(i)})^{2p} \right]^{1/2} \\ &\leq p! \, \|x - y\| \, \|h\|^p. \end{aligned} \tag{4.1}$$

Thus, $L_p(\eta_{p+1}) = p!$.

For integer parameter $k, 2 \le k \le n$, let us define the following $k \times k$ upper triangular matrix with two nonzero diagonals:

$$U_k = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ & & \dots & \dots & \\ 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \quad U_k^{-1} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 1 \\ & & \dots & \dots & \\ 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

Now we can introduce $n \times n$ upper triangular matrix A_k with the following structure:

$$A_k = \begin{pmatrix} U_k & 0 \\ 0 & I_{n-k} \end{pmatrix}.$$

Note that

$$\begin{aligned} \|A_k\|^2 &= \max_{x \in \mathbb{R}^n} \{\|Ax\|^2 : \ \|x\| \le 1\} \\ &= \max_{x \in \mathbb{R}^n} \left\{ \sum_{i=1}^{k-1} (x^{(i)} - x^{(i+1)})^2 + \sum_{i=k}^{n} (x^{(i)})^2 : \ \|x\| \le 1 \right\} \\ &\le \max_{x \in \mathbb{R}^n} \left\{ \sum_{i=1}^{k-1} \left[2(x^{(i)})^2 + 2(x^{(i+1)})^2 \right] + \sum_{i=k}^{n} (x^{(i)})^2 : \ \|x\| \le 1 \right\} \\ &\le 4. \end{aligned} \tag{4.2}$$

Our parametric family of difficult functions is defined in the following way:

$$f_k(x) = \eta_{p+1}(A_k x) - \langle e_1, x \rangle, \quad 2 \le k \le p, \tag{4.3}$$

where $e_1 = (1, 0, \dots, 0)^T$. Let us compute its optimal solution from the first-order optimality condition

$$A_k^T \nabla \eta_{p+1}(A_k x_k^*) = e_1.$$

Thus, $A_k x_k^* = y_k^*$, where y_k^* satisfies equation $\nabla \eta_{p+1}(y_k^*) = A_k^{-T} e_1 = \hat{e}_k \stackrel{\text{def}}{=} \begin{pmatrix} \bar{e}_k \\ 0_{n-k} \end{pmatrix}$, with $\bar{e}_k \in \mathbb{R}^k$ being the vector of all ones, and 0_{n-k} beng the origin in $\mathbb{R}^{n-k}$.

Thus, in coordinate form, vector y_k^* can be found from the following equations:

$$|(y_k^*)^{(i)}|^{p-2} (y_k^*)^{(i)} = 1, \quad i = 1, \dots, k,$$

$$|(y_k^*)^{(i)}|^{p-2} (y_k^*)^{(i)} = 0, \quad i = k+1, \dots, n.$$

In other words, $y_k^* = \hat{e}_k$ and vector $x_k^* = A_k^{-1}\hat{e}_k$ has the following coordinates:

$$(x_k^*)^{(i)} = (k - i + 1)_+, \quad i = 1, \ldots, n, \tag{4.4}$$

where $(\tau)_+ = \max\{0, \tau\}$. Consequently,

$$\begin{aligned} f_k^* &= \eta_{p+1}(\hat{e}_k) - \langle e_1, x_k^* \rangle \; = \; \tfrac{k}{p+1} - k \; = \; -\tfrac{kp}{p+1}, \\ \|x_k^*\|^2 &= \sum_{i=1}^{k} i^2 \; = \; \tfrac{k(k+1)(2k+1)}{6} \; \leq \; \tfrac{(k+1)^3}{3}. \end{aligned} \tag{4.5}$$

Let us describe now the abilities of tensor methods of degree $p \geq 2$ in generating new test points. We assume that the response of oracle at point $\bar{x} \in \mathbb{R}^n$ consists in the following collection of multi-linear forms:

$$f(\bar{x}), \quad D^k f(\bar{x})[h]^k, \quad k = 1, \ldots, p.$$

Therefore, we assume that the method is able to compute stationary points of the following polynomial functions

$$\phi_{a,\gamma,m}(h) = \sum_{i=1}^{p} a^{(i)} D^i f(\bar{x})[h]^i + \gamma \|h\|^m \tag{4.6}$$

with coefficients $a \in \mathbb{R}^p$, $\gamma > 0$ and $m > 1$. Denote by $\Gamma_{\bar{x},f}(a, \gamma, m)$ the set of all stationary points of this function. Then we can define the linear subspace

$$\mathcal{S}_f(\bar{x}) = \mathrm{Lin}\left(\Gamma_{\bar{x},f}(a, \gamma, m) : \; a \in \mathbb{R}^p, \; \gamma > 0, \; m > 1\right).$$

Our assumption about the rules of tensor methods is as follows.

Assumption 1 The method generates a sequence of test points $\{x_k\}_{k \geq 0}$ satisfying recursive condition

$$x_{k+1} \in x_0 + \sum_{i=0}^{k} \mathcal{S}_f(x_i), \quad k \geq 0. \tag{4.7}$$

Note that for the absolute majority of the first-order, second-order, and tensor methods this assumption is satisfied.

Let us look at the consequences Assumption 1 in the case of minimization of function $f_k(\cdot)$. Denote

$$\mathbb{R}_k^n = \{x \in \mathbb{R}^n : \; x^{(i)} = 0, \; i = k+1, \ldots, n\}, \quad 1 \leq k \leq n-1.$$

Lemma 2 *Any tensor method satisfying Assumption 1 and minimizing function $f_t(\cdot)$ starting from the point $x_0 = 0$, generates test points $\{x_k\}_{k\geq 0}$ satisfying condition*

$$x_{k+1} \in \sum_{i=0}^{k} \mathcal{S}_{f_t}(x_i) \subseteq \mathbb{R}^n_{k+1}, \quad 0 \leq k \leq t-1. \tag{4.8}$$

Proof Let us prove first that inclusion $x \in \mathbb{R}^n_k$ with $k \geq 1$ implies $\mathcal{S}_{f_t}(x) \subseteq \mathbb{R}^n_{k+1}$. Indeed, since matrix A_t is upper triangular, this inclusion ensures that $y \stackrel{\text{def}}{=} A_t x \in \mathbb{R}^n_k$. Therefore, all derivatives of function $f_t(\cdot)$ along direction $h \in \mathbb{R}^n$ have the following form:

$$Df_t(x)[h] = D\eta_{p+1}(y)[A_t h] - h^{(1)} = \sum_{i=1}^{k} d_{i,1}\langle e_i, A_t h\rangle - h^{(1)},$$

$$D^i f_t(x)[h]^k = D^k \eta_{p+1}(y)[A_t h]^k = \sum_{i=1}^{k} d_{i,k}\langle e_i, A_t h\rangle^k, \quad 2 \leq k \leq p,$$

with certain coefficients $d_{i,k}, i = 1, \ldots, n, k = 1, \ldots, p$. This means that the gradients of these derivatives in h are as follows

$$\nabla Df_t(x)[h] = \sum_{i=1}^{k} d_{i,1} A_t^T e_i - e_1,$$

$$\nabla D^k f_t(x)[h]^k = \sum_{i=1}^{k} k d_{i,k}\langle e_i, A_t h\rangle^{k-1} A_t^T e_i, \quad 2 \leq k \leq p,$$

Thus $\nabla D^k f_t(x)[h]^k \in \mathbb{R}^n_{k+1}$ for all k, $1 \leq k \leq p$. Hence, since the regularization term in definition (4.6) is formed by the standard Euclidean norm, all stationary points of this function belong to $\mathbb{R}^n_{k+1}$.

Now we can prove statement of the lemma by induction. For $k = 0$, we have $x_0 = 0$, and therefore

$$\nabla f_t(x_0) = -e_1, \quad D^i f(x_0)[h]^i = 0, \quad i = 1, \ldots, p,$$

for all $h \in \mathbb{R}^n$. Consequently, stationary points of all possible functions $\phi_{a,\gamma,m}(\cdot)$ belong to $\mathbb{R}^n_1$ implying $\mathcal{S}_{f_t}(x_0) = \mathbb{R}^n_1$. Thus, x_1 belongs to $\mathbb{R}^n_1$ by Assumption 1.

Assume now that all $x_i \in \mathbb{R}^n_k$, $i = 1, \ldots, k$, for some $k \geq 1$. Then, as we have already seen, $\mathcal{S}_{f_t}(x_i) \subseteq \mathbb{R}^n_{k+1}$. Hence, the inclusion (4.8) follows from Assumption 1. □

Now we can prove the main statement of this section.

Theorem 4 *Let some tensor method $\mathcal{M}$ of degree p satisfies Assumption 1. Assume that this method ensures for any function $f \in \mathcal{F}_p$ with $L_p((f) < +\infty$ the following rate of convergence:*

$$\min_{0\le k\le t} f(x_k) - f_* \le \frac{L_p \|x_0 - x^*\|^{p+1}}{(p+1)!\,\kappa(t)}, \quad t \ge 1, \tag{4.9}$$

where $\{x_k\}_{k\ge 0}$ is the sequence of test points, generated by method $\mathcal{M}$ and x^* is the solution of the problem (2.7). Then for all $t \ge 1$ such that $2t+1 \le n$ we have

$$\kappa(t) \le \frac{1}{3p} 2^{p+1} (2t+2)^{\frac{3p+1}{2}}. \tag{4.10}$$

Proof Let us use method $\mathcal{M}$ for minimizing function $f(x) = f_{2t+1}(x)$. In view of Lemma 2, we have $x_i \in \mathbb{R}^n_t$ for all $i, 0 \le i \le t$. However,

$$f_{2t+1}(x) \equiv f_t(x), \quad \forall x \in \mathbb{R}^n_t.$$

At the same time, for all $x, y, h \in \mathbb{R}^n$ we have

$$\begin{aligned}
&|D^p f_{2t+1}(x)[h]^p - D^p f_{2t+1}(y)[h]^p| \\
&\quad = |D^p \eta_{p+1}(x)[A_{2t+1}h]^p - D^p \eta_{p+1}(y)[A_{2t+1}h]^p| \\
&\quad \overset{(4.1)}{\le} p!\|x-y\|\,\|A_{2t+1}h\|^p \\
&\quad \overset{(4.2)}{\le} 2^p p!\|x-y\|.
\end{aligned}$$

Therefore, $L_p(f_{2t+1}) \le 2^p p!$, and we have

$$(p+1)!\,\kappa(t) \overset{(4.9)}{\le} \frac{L_p(f_{2t+1})\|x_0 - x^*_{2t+1}\|^{p+1}}{\min\limits_{0\le k\le t} f(x_k) - f^*_{2t+1}} \le \frac{2^p p!(2t+2)^{\frac{3}{2}(p+1)}}{3(f^*_t - f^*_{2t+1})}$$

$$= \frac{2^p p!(2t+2)^{\frac{3}{2}(p+1)}}{3(t+1)} \cdot \frac{p+1}{p}. \qquad \square$$

5 Third-order methods: implementation details

Tensor optimization methods, presented in Sects. 2 and 3, are based on the solution of the auxiliary optimization problem (2.6). In the existing literature on the tensor methods [6,7,20,32], it was solved by the standard local technique of Nonconvex Optimization. However, now we know that by Theorem 1, this problem is convex. Hence, it is solvable by the standard and very efficient methods of Convex Optimization.

Since we need to solve this problem at each iteration of the methods, its complexity significantly affects the total computational time. Since the objective function in the problem (2.6) is a *convex multivariate polynomial*, we there could exist some special

efficient algorithms for finding its solution. Unfortunately, at this moment the authors failed to find such methods in the literature. Therefore, we present in this section a special approach for solving the problem (2.6) with the third degree Taylor approximation, which is based on the recently developed optimization framework of *relatively smooth functions* (see [4,21]).

Let us fix an arbitrary $x \in \mathbb{E}$. Consider the following multivariate polynomial of degree three:[2]

$$\Phi_x(h) = \langle \nabla f(x), h \rangle + \tfrac{1}{2}\langle \nabla^2 f(x)h, h \rangle + \tfrac{1}{6} D^3 f(x)[h]^3.$$

Since in this section we work only with third-order approximations, we drop the corresponding index.

Recall that $D^3 f(x)[h_1, h_2, h_3]$ is a symmetric trilinear form. Hence, $D^3 f(x)[h_1, h_2] \in \mathbb{E}^*$ is a symmetric bilinear vector function, and $D^3 f(x)[h]$ is a linear function of $h \in \mathbb{E}$, whose values are self-adjoint linear operators from $\mathbb{E}$ to $\mathbb{E}^*$ (as Hessians).

Denote $p(h) = \frac{1}{6} D^3 f(x)[h]^3$. Then we can define its gradient and Hessian as follows:

$$\nabla p(h) = \tfrac{1}{2} D^3 f(x)[h, h], \quad \nabla^2 p(h) = D^3 f(x)[h]. \tag{5.1}$$

In this section, our main class of functions is $\mathcal{F}_3$, composed by convex functions, which are three times continuously differentiable, and for which the constant L_3 is finite. As we have shown in Theorem 1, our assumptions imply interesting relations between derivatives. Let us derive a consequence of the matrix inequality (2.3).

Lemma 3 *For any $h \in \mathbb{E}$ and $\tau > 0$, we have*

$$-\tfrac{1}{\tau}\nabla^2 f(x) - \tfrac{\tau}{2} L_3 \|h\|^2 B \preceq D^3 f(x)[h] \preceq \tfrac{1}{\tau}\nabla^2 f(x) + \tfrac{\tau}{2} L_3 \|h\|^2 B. \tag{5.2}$$

Consequently, for any $h, u \in \mathbb{E}$, we get

$$D^3 f(x)[h, u, u] \le \sqrt{2L_3}\langle \nabla^2 f(x)u, u \rangle^{1/2} \|u\| \|h\|. \tag{5.3}$$

Proof Let us fix arbitrary directions $u, h \in \mathbb{E}$. Then, in view of relation (2.3) we have

$$\begin{aligned} 0 \le \langle \nabla^2 f(x+h)u, u \rangle &\le \langle \nabla^2 \Phi_x(h)u, u \rangle + \tfrac{1}{2} L_3 \|h\|^2 \|u\|^2 \\ &= \langle (\nabla^2 f(x) + D^3 f(x)[h])u, u \rangle + \tfrac{1}{2} L_3 \|h\|^2 \|u\|^2 \end{aligned}$$

Thus, replacing h by τh with $\tau > 0$ and dividing the resulting inequality by τ, we get

$$-\langle D^3 f(x)[h]u, u \rangle \le \tfrac{1}{\tau}\langle \nabla^2 f(x)u, u \rangle + \tfrac{\tau}{2} L_3 \|h\|^2 \|u\|^2.$$

[2] We drop the index in this notation since from now on we always have $p = 3$.

And this is equivalent to the left-hand side of matrix inequality (5.2). Its right-hand side can be obtained by replacing h by $-h$, which gives

$$\langle D^3 f(x)[h]u, u\rangle \le \frac{1}{\tau}\langle \nabla^2 f(x)u, u\rangle + \frac{\tau}{2}L_3\|h\|^2\|u\|^2.$$

Minimizing the right-hand side of this inequality in τ, we get (5.3). □

Let us look now at our auxiliary minimization problem:

$$\Omega_{x,M}(h) \stackrel{\text{def}}{=} \Phi_x(h) + \frac{M}{2}d_4(h) \;\to\; \min_{h\in\mathbb{E}}, \tag{5.4}$$

where $d_4(h) = \frac{1}{4}\|h\|^4$. In view of Theorem 1, function $\Omega_{x,M}(\cdot)$ is convex for

$$M = \tau^2 L_3 \tag{5.5}$$

with $\tau \ge 1$. For any $h \in \mathbb{E}$, we have

$$\begin{aligned}
\nabla^2\Omega_{x,M}(h) &= \nabla^2 f(x) + D^3 f(x)[h] + \tfrac{M}{2}\nabla^2 d_4(h) \\
&\stackrel{(5.2)}{\succeq} (1-\tfrac{1}{\tau})\nabla^2 f(x) + \tfrac{M}{2}\nabla^2 d_4(h) - \tfrac{\tau}{2}L_3\|h\|^2 B \\
&\stackrel{(2.2)}{\succeq} (1-\tfrac{1}{\tau})\nabla^2 f(x) + \tfrac{M-\tau L_3}{2}\nabla^2 d_4(h).
\end{aligned}$$

Let $\rho_x(h) = \frac{1}{2}(1-\frac{1}{\tau})\langle\nabla^2 f(x)h, h\rangle + \frac{M-\tau L_3}{2}d_4(h)$. Then, we have proved that

$$\nabla^2\Omega_{x,M}(h) \succeq \nabla^2\rho_x(h), \quad h \in \mathbb{E}. \tag{5.6}$$

On the other hand,

$$\begin{aligned}
\nabla^2\Omega_{x,M}(h) &= \nabla^2 f(x) + D^3 f(x)[h] + \tfrac{M}{2}\nabla^2 d_4(h) \\
&\stackrel{(5.2)}{\preceq} (1+\tfrac{1}{\tau})\nabla^2 f(x) + \tfrac{M}{2}\nabla^2 d_4(h) + \tfrac{\tau}{2}L_3\|h\|^2 B \\
&\stackrel{(2.2)}{\preceq} (1+\tfrac{1}{\tau})\nabla^2 f(x) + \tfrac{M+\tau L_3}{2}\nabla^2 d_4(h) \\
&\stackrel{(5.5)}{=} \left(\tfrac{1+\tau}{\tau-1}\right)\left((1-\tfrac{1}{\tau})\nabla^2 f(x) + \tfrac{\tau(\tau-1)L_3}{2}\nabla^2 d_4(h)\right) \\
&= \tfrac{\tau+1}{\tau-1}\nabla^2\rho_x(h).
\end{aligned}$$

Thus, we have proved the following lemma.

Lemma 4 *Let $M = \tau^2 L_3$ with $\tau > 1$. Then function $\Omega_{x,M}(\cdot)$ satisfies the strong relative smoothness condition*

$$\nabla^2 \rho_x(h) \preceq \nabla^2 \Omega_{x,M}(h) \preceq \kappa(\tau) \nabla^2 \rho_x(h), \quad h \in \mathbb{E}, \tag{5.7}$$

with respect to function $\rho_x(\cdot)$, where $\kappa(\tau) = \frac{\tau+1}{\tau-1}$.

As it is shown in [21], condition (5.7) allows to solve problem (5.4) very efficiently by a kind of primal gradient method. In accordance to this approach, we need to define the *Bregman distance* of function $\rho_x(\cdot)$:

$$\begin{aligned} \beta_{\rho_x}(u, v) &= \rho_x(v) - \rho_x(u) - \langle \nabla \rho_x(u), v - u \rangle \\ &= \tfrac{1}{2}(1 - \tfrac{1}{\tau}) \langle \nabla^2 f(x)(v - u), v - u \rangle + \tfrac{\tau(\tau-1)}{2} L_3 \beta_{d_4}(u, v), \end{aligned}$$

and iterate the process

$$h_{k+1} = \arg\min_{h \in \mathbb{E}} \left\{ \langle \nabla \Omega_{x,M}(h_k), h - h_k \rangle + \kappa(\tau) \beta_{\rho_x}(h_k, h) \right\}.$$

In our case, this method has the following form:

$$\begin{aligned} h_0 &= 0, \\ h_{k+1} &= \arg\min_{h \in \mathbb{E}} \Bigg\{ \langle \nabla \Omega_{x,M}(h_k), h - h_k \rangle \\ &\quad + \frac{\tau+1}{2} \left[\frac{1}{\tau} \langle \nabla^2 f(x)(h - h_k), h - h_k \rangle + \tau L_3 \beta_{d_4}(h_k, h) \right] \Bigg\}, \quad k \geq 0. \end{aligned} \tag{5.8}$$

In accordance to Theorem 3.1 in [21], the rate of convergence of this method is as follows:

$$\Omega_{x,M}(h_k) - \Omega_{x,M}(h_*) \leq \frac{\beta_{\rho_x}(h_0, h_*)}{\left(\frac{\kappa(\tau)}{\kappa(\tau)-1}\right)^k - 1} = \frac{\frac{\tau-1}{2}\left[\frac{1}{\tau}\langle \nabla^2 f(x) h_*, h_* \rangle + \frac{\tau}{4} L_3 \|h_*\|^4\right]}{\left(\frac{\tau+1}{2}\right)^k - 1}, \tag{5.9}$$

where h_* is the unique optimal solution to problem (5.4).

As we can see, the algorithm (5.8) is very fast. Its linear rate of convergence depends only on absolute constant $\tau > 1$, which can be chosen reasonably close to one for allowing faster convergence of the main tensor methods (2.16) and (3.12). Let us discuss two potentially expensive operations in the implementation of method (5.8).

1. **Computation of the gradient** $\nabla \Omega_{x,M}(h)$. Note that

$$\nabla \Omega_{x,M}(h) = \nabla f(x) + \nabla^2 f(x) h + \tfrac{1}{2} D^3 f(x)[h]^2.$$

In this formula, only the computation of the third derivative may be dangerous. However, this difficulty can be resolved using the technique of automatic differentiation (see, for example, [19]). Indeed, assume we have a sequence of

operations for computing the function value $f(x)$ with computational complexity T. Let us fix a direction $h \in \mathbb{E}$. Then by forward differentiation, we can generate automatically a sequence of operations for computing the value

$$g_h(x) = \langle \nabla^2 f(x)h, h \rangle$$

with computational complexity $O(T)$. Now, by backward differentiation in x, we can compute the gradient of this function:

$$\nabla g(x) = D^3 f(x)[h, h]$$

with computational complexity $O(T)$. Thus, the oracle complexity of method (5.8) is proportional to the complexity of computing the function value $f(x)$.
Another example of simple computation of the third derivative is provided by a separable objective function. Assume that $\mathbb{E} = \mathbb{R}^n$ and

$$f(x) = \sum_{i=1}^{N} f_i(b_i - \langle a_i, x \rangle),$$

where $a_i \in \mathbb{R}^n$ and univariate functions $f_i(\cdot)$ are three times continuously differentiable, $i = 1, \ldots, N$. Then vector $D^3 f[h]^2$ has the following representation:

$$D^3 f(x)[h]^2 = -\sum_{i=1}^{N} a_i \, f_i'''(b_i - \langle a_i, x \rangle) \langle a_i, h \rangle^2.$$

Thus, for solving the problem (5.4), we need to compute in advance all values

$$f_i'''(b_i - \langle a_i, x \rangle), \quad i = 1, \ldots, N$$

(this needs $O(nN)$ operations). After that, each computation of vector $D^3 f(x)[h]^2 \in \mathbb{R}^n$ also needs $O(nN)$ operations. This computation will be cheaper for the sparse data.

2. **Solution of the auxiliary problem** At all iterations of method (5.8), we need to solve an auxiliary problem in the following form:

$$\min_{h \in \mathbb{E}} \left\{ \langle c, h \rangle + \tfrac{1}{2} \langle Ah, h \rangle + \tfrac{\gamma}{4} \|h\|^4 \right\}, \tag{5.10}$$

where $A \succeq 0$ and $\gamma > 0$. Note that at all these iterations only the vector c and coefficients γ are changing, and matrix $A = \nabla^2 f(x)$ remains the same. Therefore, before the algorithm (5.8) starts working, it is reasonable to transform this matrix in a tri-diagonal form:

$$A = UTU^T,$$

where $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix: $UU^T = I$, and $T \in \mathbb{R}^{n\times n}$ is tri-diagonal.

Denoting now $\tilde{c} = U^T c$, we have:

$$\begin{aligned}
&\min_{h\in\mathbb{E}} \left\{ \langle c, h\rangle + \frac{1}{2}\langle UTU^T h, h\rangle + \frac{\gamma}{4}\|h\|^4 \right\} \\
&\quad = \min_{h\in\mathbb{E}} \max_{\tau>0} \left\{ \langle \tilde{c}, U^T h\rangle + \frac{1}{2}\langle TU^T h, U^T h\rangle + \frac{\gamma}{2}\tau\|U^T h\|^2 - \frac{1}{2}\tau^2 \right\} \\
&\quad = \max_{\tau>0} \min_{h\in\mathbb{E}} \left\{ \langle \tilde{c}, U^T h\rangle + \frac{1}{2}\langle TU^T h, U^T h\rangle + \frac{\gamma}{2}\tau\|U^T h\|^2 - \frac{1}{2}\tau^2 \right\} \\
&\quad = -\min_{\tau>0} \left\{ \frac{1}{2}\tau^2 + \frac{1}{2}\langle (\gamma\tau I + T)^{-1}\tilde{c}, \tilde{c}\rangle \right\}. \qquad (5.11)
\end{aligned}$$

Thus, the solution of the primal problem can be retrieved from a solution to the univariate dual problem. The complexity of computing its function value and derivative is linear in n. Moreover, since its objective function is strongly convex and infinitely times differentiable, all reasonable one-dimensional methods have global linear rate of convergence and the quadratic convergence in the end.

Let us estimate the total computational complexity of the method (5.8), assuming that the computational time of the value of the objective function is T_f. Assume also that its gradient, the product of its Hessian by a vector, and the value of its third derivative on two identical vectors can be computed using the fast backward differentiation (then the complexity of all these operations is $O(T_f)$). Then, the most expensive operations in this method are as follows.

- Computation of the Hessian $\nabla^2 f(x)$ and its tri-diagonal factorization: $O(nT_f + n^3)$ operations.
- We need $O(\ln \frac{1}{\epsilon})$ iterations of method (5.8), in order to get ϵ-solution of the auxiliary problem. At each iteration of this method we need:
  - Compute the gradient $\nabla\Omega_{x,M}(h_k)$: $O(T_f)$ operations.
  - Compute the vector $\tilde{c}$ for the univariate problem in (5.11): $O(n^2)$ operations.
  - Solve the dual problem in (5.11) up to accuracy δ: $O(n \ln \frac{1}{\delta})$ operations.
  - Compute an approximate solution $h = -U(\gamma\tau I + T)^{-1}\tilde{c}$ of the problem (5.10), using an approximate solution τ of the dual problem: $O(n^2)$ operations.

Thus, we come to the following estimate:

$$O\left(nT_f + n^3 + \left[T_f + n^2 + n\ln\tfrac{1}{\delta}\right]\ln\tfrac{1}{\epsilon}\right).$$

This is the same order of complexity as that of one iteration in Trust Region Methods [15] and usual Cubic Regularization [27,31]. However, we can expect that the third-order methods converge much faster.

For the readers, which are not interested in all these computational details, we just mention that the Galahad Optimization Library [16] has special subroutines for solving the auxiliary problems in the form (5.10).

6 Discussion

In this paper, we did an important step towards practical implementation of tensor methods in unconstrained convex optimization. We have shown that the auxiliary optimization problems in these scheme can be reduced to minimization of a convex multivariate polynomial. In the important case of third-order tensor, we have proved that this problem can be efficiently solved by a special optimization scheme derived from the relative smoothness condition.

Our results highlight several interesting questions. One of the direct consequences of our approach is a systematic way of generating convex multivariate polynomials. Is it possible to minimize them by some tools of Algebraic Geometry (see [23] for the related technique like sums of squares, etc.), or we need to treat them using an appropriate technique from Convex Optimization? The results of Sect. 5 demonstrate a probably unbeatable superiority of optimization technique for the third-order polynomials. But what happens with polynomials of higher degree?

One of the difficult unsolved problems in our approach is related to dynamic adjustment of the Lipschitz constant for the highest derivative. This dynamic estimate should not be much bigger than the actual Lipschitz constant. On the other hand, it must ensure convexity of the auxiliary problem solved at each iteration of the tensor methods. This question is clearly crucial for the practical efficiency of the high-order schemes.

Simple comparison of the complexity bounds in Sects. 3 and 4 shows that we failed to develop an optimal tensor scheme. The missing factor in the complexity estimates is of the order of $O\left(\left(\frac{1}{\epsilon}\right)^{\frac{1}{p+1}-\frac{2}{3p+1}}\right) = O\left(\left(\frac{1}{\epsilon}\right)^{\frac{p-1}{(p+1)(3p+1)}}\right)$. For $p = 3$, this factor is of the order of $O\left(\left(\frac{1}{\epsilon}\right)^{\frac{1}{20}}\right)$. This means that from the viewpoint of practical efficiency, the cost of one iteration of the hypothetical optimal scheme must be of the same order as that of the accelerated tensor method (3.12). Any additional logarithmic factors in the complexity bound of this "optimal" method will definitely kill its tiny superiority in the convergence rate.

In the last years, we have seen an increasing interest to universal methods, which can adjust to the best possible Hölder condition instead of the Lipschitz one during the running optimization process (see [14,17,29]). Of course, it is very interesting to extend this philosophy onto the tensor minimization schemes. Another important extension could be the treatments of the constraints, either in functional form, or using the framework of composite minimization [28]. The main difficulty here is related to the complexity of the auxiliary optimization problems.

One of the main restrictions for practical implementation of our results is the necessity to know the Lipschitz constant of the corresponding derivative. If our estimate is too small, then the auxiliary problem (2.6) may loose convexity. Consequently, we will loose the fast convergence in the auxiliary process (5.8). However, this observation gives us a clue how to tune this constant: if we see that this process is too slow, this means that our estimate is too small. But of course it is very interesting to find a recipe with better theoretical justification.

Acknowledgements The author is very thankful to Geovani Grapiglia for interesting discussions of the results. The comments of two anonymous referees were extremely useful.

References

1. Agarwal, N., Hazan, E.: Lower Bounds for Higher-Order Convex Optimization (2017). arXiv:1710.10329v1 [math.OC]
2. Arjevani, Y., Shamir, O., Shiff, R.: Oracle Complexity of Second-Order Methods for Smooth Convex Optimization (2017). arXiv:1705.07260 [math.OC]
3. Baes, M.: Estimate sequence methods: extensions and approximations. Optim. Online (2009)
4. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuety: first-order methods revisited and applications. Math. Oper. Res. **42**, 330–348 (2017)
5. Bian, W., Chen, X., Ye, Y.: Complexity analysis of interior-point algorithms for non-Lipschitz and non-convex minimization. Math. Program. **139**, 301–327 (2015)
6. Birgin, E.G., Gardenghi, J.L., Martines, J.M., Santos, S.A.: Remark on Algorithm 566: Modern Fortran Routines for Testing Unconsrained Optimization Software with Derivatives up to Third-Order. Technical report, Department of Computer Sciences, University of Sao Paolo, Brazil (2018)
7. Birgin, E.G., Gardenghi, J.L., Martines, J.M., Santos, S.A.: On the Use of Third-Order Models with Fourth-Order Regularization for Unconstrained Optimization. Technical report, Department of Computer Sciences, University of Sao Paolo, Brazil (2018)
8. Birgin, E.G., Gardenghi, J.L., Martines, J.M., Santos, S.A., Toint, PhL: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularization models. Math. Program. **163**, 359–368 (2017)
9. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. Archiv (2017). arXiv:1710.11606
10. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points II. Archiv (2017). arXiv:1711.00841
11. Cartis, C., Gould, N.I.M., Toint, PhL: Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. Math. Program. **130**(2), 295–319 (2012)
12. Cartis, C., Gould, N.I.M., Toint, PhL: Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function evaluation complexity. Math. Program. **127**(2), 245–295 (2011)
13. Cartis, C., Gould, N.I.M., Toint, PhL: Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. Optim. Methods Softw. **27**(2), 197–219 (2012)
14. Cartis, C., Gould, N.I.M., Toint, PhL: Universal regularization methods–varying the power, the smoothness and the accuracy. SIAM. J. Optim. **29**(1), 595–615 (2019)
15. Conn, A.R., Gould, N.I.M., Toint, PhL: Trust Region Methods. MOS-SIAM Series on Optimization, New York (2000)
16. Gould, N.I.M., Orban, D., Toint, PhL: GALAHAD, a library of thread-safe Fortran 90 packages for large-scale nonlinear optimization. ACM Trans. Math. Softw. **29**(4), 353–372 (2003)
17. Grapiglia, G.N., Nesterov, Yu.: Regularized Newton methods for minimizing functions with Hölder continuous Hessians. SIOPT **27**(1), 478–506 (2017)
18. Grapiglia, G.N., Yuan, J., Yuan, Y.: On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization. Math. Program. **152**, 491–520 (2015)
19. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. Applied Mathematics, vol. 105, 2nd edn. SIAM, Philadelphia (2008)
20. Gundersen, G., Steihaug, T.: On large-scale unconstrained optimization problems and higher order methods. Optim. Methods. Softw. **25**(3), 337–358 (2010)
21. Lu, H., Freund, R., Nesterov, Yu.: Relatively smooth convex optimization by first-order methods, and applications. SIOPT **28**(1), 333–354 (2018)

22. Hoffmann, K.H., Kornstaedt, H.J.: Higher-order necessary conditions in abstract mathematical programming. JOTA **26**, 533–568 (1978)
23. Lasserre, J.B.: Moments, Positive Polynomials and Their Applications. Imperial College Press, London (2010)
24. Monteiro, R.D.C., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIOPT **23**(2), 1092–1125 (2013)
25. Nesterov, Yu.: Introductory Lectures on Convex Optimization. Kluwer, Boston (2004)
26. Nesterov, Yu.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
27. Nesterov, Yu.: Accelerating the cubic regularization of Newton's method on convex problems. Math. Program. **112**(1), 159–181 (2008)
28. Nesterov, Yu.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
29. Nesterov, Yu.: Universal gradient methods for convex optimization problems. Math. Program. **152**, 381–404 (2015)
30. Nesterov, Yu., Nemirovskii, A.: Interior Point Polynomial Methods in Convex Programming: Theory and Applications. SIAM, Philadelphia (1994)
31. Nesterov, Yu., Polyak, B.: Cubic regularization of Newton's method and its global performance. Math. Program. **108**(1), 177–205 (2006)
32. Schnabel, R.B., Chow, T.T.: Tensor methods for unconstrained optimization using second derivatives. SIAM J. Optim. **1**(3), 293–315 (1991)