INEXACT BASIC TENSOR METHODS FOR SOME CLASSES OF CONVEX OPTIMIZATION PROBLEMS

Yurii Nesterov

REPRINT | 3229

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

OPTIMIZATION METHODS & SOFTWARE
https://doi.org/10.1080/10556788.2020.1854252

Taylor & Francis
Taylor & Francis Group

Inexact basic tensor methods for some classes of convex optimization problems

Yurii Nesterov

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

ABSTRACT

In this paper, we analyse the Basic Tensor Methods, which use approximate solutions of the auxiliary problems. The quality of this solution is described by the residual in the function value, which must be proportional to $\epsilon^{\frac{p+1}{p}}$, where $p \geq 1$ is the order of the method and ϵ is the desired accuracy in the main optimization problem. We analyse in details the auxiliary schemes for the third- and second-order tensor methods. The auxiliary problems for the third-order scheme can be solved very efficiently by a linearly convergent gradient-type method with a preconditioner. The most expensive operation in this process is a preliminary factorization of the Hessian of the objective function. For solving the auxiliary problem for the second order scheme, we suggest two variants of the Fast Gradient Methods with restart, which converge as $O(\frac{1}{k^6})$, where k is the iteration counter. Finally, we present the results of the preliminary computational experiments.

ARTICLE HISTORY
Received 15 June 2020
Accepted 18 November 2020

KEYWORDS
High-order methods; tensor methods; complexity bounds; convex optimization

2010 MATHEMATICS SUBJECT CLASSIFICATION
90C25

1. Introduction

Motivation. Development of the theory of Tensor Methods in the last years created an additional motivation for the research on the efficient procedures for solving the corresponding auxiliary problems. Indeed, without this technique, all results on the faster convergence of the high-order methods remain only theoretical achievements. Starting from the paper [2], in several articles [3–6,9] the authors analysed the possibility of using points satisfying an approximate first-order optimality condition. However, it is not easy to analyse the complexity of computing such points by the auxiliary optimization schemes. This is the reason why we use in this paper a more traditional measure of inaccuracy, defined by the residual in the function value.

Indeed, this measure is standard in the theory of Convex Optimization. There exist a family of Fast Gradient Methods, which allow to solve the auxiliary problems much faster and with worst-case complexity guarantees. In this paper in Section 2 we start from deriving the natural conditions on the accuracy of the solution of the auxiliary problems, acceptable for the Basic Tensor Method. This accuracy is naturally related to the desired

CONTACT Yurii Nesterov yurii.nesterov@uclouvain.be

accuracy of the main optimization problem. After that, in Section 3 we analyse the complexity of the auxiliary problem for the third-order Tensor Method. We show that the auxiliary problem is this case can be solved very efficiently by a simple gradient method based on *relative smoothness condition* [1,8]. The most expensive operation in its implementation is the matrix factorization, which has to be done only one, in the beginning of the auxiliary minimization process. Moreover, we show that this scheme admits a reliable stopping criterion, which properly describes the quality of the approximate solution. Thus, we show that the computational cost of implementing the third-order methods is essentially the same as that of the second-order schemes based on matrix factorization.

In the next Section 4 we analyse the complexity of the auxiliary problem in the regularized second-order scheme. We show that the objective function in these problems, which is the sum of a quadratic function and the cubic term, can be minimized very efficiently by the Fast Gradient Methods with restarts. The complexity of this auxiliary optimization problem is $O(\frac{1}{\delta^{1/6}})$, where δ is the target accuracy. Thus, this approach have good chance to compete with the existing technique for the second-order method. Our approach may look similar to the developments in Section 6 of [16]. However, our scheme has no hidden parameters and it can be implemented directly.

Further, in Section 5, we describe another variant of the Fast Gradient Method with flexible restart, which can be applied to the auxiliary problem for the second-order methods. It has the same worst-case complexity bound $O(\frac{1}{\delta^{1/6}})$. Finally, in Section 6 we present the preliminary computational results, which confirm our theoretical conclusions.

Notation and generalities. In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space, and by $\mathbb{E}^*$ its dual space, composed by all linear functions on $\mathbb{E}$. The value of the linear function $s \in \mathbb{E}^*$ at point $x \in \mathbb{E}$ is denoted by $\langle s, x \rangle$. The most important example of linear function is the *gradient* $\nabla f(x)$ of the differentiable function $f(\cdot)$ at point $x \in \mathbb{E}$. The Hessian $\nabla^2 f(x)$ can be seen as a self-adjoint linear operator from $\mathbb{E}$ to $\mathbb{E}^*$.

Let us fix a positive definite linear operator $B : \mathbb{E} \to \mathbb{E}^*$. Then we can introduce in the primal and dual spaces the conjugate Euclidean norms:

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

In this paper, we work only with Euclidean norms defined by the above relations.

Recall that function $F(\cdot)$ is called *uniformly convex* of degree $q \geq 2$ if for any x and $y \in \mathrm{dom}\, F$ we have

$$F(y) \geq F(x) + \langle g_x, y - x \rangle + \frac{\sigma_q}{q} \|x - y\|^q, \tag{1}$$

where σ_q is a positive parameter and g_x is an arbitrary vector from subdifferential $\partial F(x)$. Minimizing both parts of inequality (1) in $y \in \mathrm{dom}\, F$, we get

$$\begin{aligned} \min_{y \in \mathrm{dom}\, F} F(y) &\geq F(x) + \min_{r \geq 0} \left\{ -r \|g_x\|_* + \frac{\sigma_q}{q} r^q \right\} \\ &= F(x) - \frac{q-1}{q\, \sigma_q^{1/(q-1)}} \|g_x\|_*^{\frac{q}{q-1}}. \end{aligned} \tag{2}$$

For p-times continuously differentiable function $f(\cdot)$, $p \geq 1$, with open convex domain $\mathrm{dom} f \subseteq \mathbb{E}$ we can introduce its pth directional derivative at point $x \in \mathrm{dom} f$ along

directions $h_1, \dots, h_p$ and denote it by

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

If $h_i = h$ for all $i = 1, \dots, p$, we use a shorter notation $D^p f(x)[h]^p$. The norm of this derivative is defined in the usual way:

$$\begin{aligned} \|D^p f(x)\| &= \max_{h_1,\dots,h_p} \left\{ |D^p f(x)[h_1, \dots, h_p]| : \|h_i\| \le 1, i = 1, \dots, p \right\} \\ &= \max_h \left\{ |D^p f(x)[h]^p| : \|h\| \le 1 \right\}. \end{aligned}$$

Similarly, the Lipschitz condition for the pth derivative has the following sense:

$$\begin{aligned} \|D^p f(x) - D^p f(y)\| &\stackrel{\text{def}}{=} \max_h \left\{ |D^p f(x)[h]^p - D^p f(y)[h]^p| : \|h\| \le 1 \right\} \\ &\le L_p(f)\|x - y\|, \quad x, y \in \operatorname{dom} f. \end{aligned}$$

2. Inexact basic tensor methods

In this paper, we consider a convex optimization problem in the following *composite form*:

$$\min_{x \in \operatorname{dom} \Psi} \{F(x) = f(x) + \Psi(x)\}, \tag{3}$$

where $f(\cdot)$ is a smooth convex function and $\Psi(\cdot)$ is a closed convex function such that

$$\operatorname{dom} \Psi \subseteq \operatorname{int}(\operatorname{dom} f).$$

Let us assume that the problem (3) is solvable and denote by $x^* \in \operatorname{dom} \Psi$ one of its optimal solutions, $F_* \stackrel{\text{def}}{=} F(x^*)$.

We assume that $f(\cdot)$ is p-times continuously differentiable ($p \ge 1$) and its p-th derivative satisfies Lipschitz condition:

$$\|D^p f(y) - D^p f(x)\| \le L_p(f)\|x - y\|, \quad x, y \in \operatorname{dom} \Psi. \tag{4}$$

It is well known (e.g. [15]) that this condition implies the following bound:

$$\begin{aligned} |f(y) - f_{\bar{x},p}(y)| &\le \frac{H}{(p+1)!}\|y - \bar{x}\|^{p+1}, \quad \bar{x}, y \in \operatorname{dom} f, \\ f_{\bar{x},p}(y) &\stackrel{\text{def}}{=} \sum_{k=0}^{p} \frac{1}{k!} D^k f(\bar{x})[y - \bar{x}]^k, \end{aligned} \tag{5}$$

where $H \ge L_p(f)$. On the other hand, if

$$H \ge p L_p(f), \tag{6}$$

then function

$$\varphi_{\bar{x},p,H}(y) \stackrel{\text{def}}{=} f_{\bar{x},p}(y) + \frac{H}{(p+1)!}\|y - \bar{x}\|^{p+1}$$

is convex (see Theorem 1 in [12]). Therefore, for generating test points in the minimization methods for problem (3) we can use solution of the following auxiliary problem:

$$\min_{x \in \mathrm{dom}\, \Psi} \left\{ F_{\bar{x},p,H}(x) \stackrel{\mathrm{def}}{=} \varphi_{\bar{x},p,H}(x) + \Psi(x) \right\}. \tag{7}$$

Our main assumption is as follows.

Assumption 2.1: Function $\Psi(\cdot)$ is simple enough for having problem (7) tractable.

However, even with this assumption, usually we cannot compute an exact solution of problem (7) in a closed form (unless, may be, for $p = 1$). This is the reason why we need to describe somehow the quality of the approximate solutions.

Definition 2.1: Let $\delta > 0$ be a measure of inaccuracy in problem (7) and $\bar{x} \in \mathrm{dom}\, \Psi$. Denote by $T = T_{\delta,p,H}(\bar{x})$ any point in $\mathrm{dom}\, \Psi$ satisfying the following inequality:

$$F_{\bar{x},p,H}(T) \leq \min_{x \in \mathrm{dom}\, \Psi} F_{\bar{x},p,H}(x) + \delta. \tag{8}$$

We refer to the point $T_{\delta,p,H}(\bar{x})$ as to *result of Inexact Tensor Step* of degree p from the point $\bar{x}$.

Let us choose a starting point $x_0 \in \mathrm{dom}\, \Psi$. We assume that the level sets of the objective function in problem (3) are compact:

$$R_F(x_0) \stackrel{\mathrm{def}}{=} \max_{x \in \mathrm{dom}\, \Psi} \left\{ \|x - x^*\| : F(x) \leq F(x_0) \right\} < +\infty. \tag{9}$$

Denote

$$C_{p,H}(x_0) = \frac{1}{p!}(L_p(f) + H) R_F^{p+1}(x_0). \tag{10}$$

Our Basic Tensor Method consists of the Preliminary Step and the Iteration Process. The goal of the Preliminary Step is to compute an appropriate starting point for the Iteration Process.

Lemma 2.1: *Let* $\delta \leq \frac{p}{p+1} C_{p,H}(x_0)$. *Then*

$$F_{x_0,p,H}(T_{\delta,p,H}(x_0)) - F^* \leq C_{p,H}(x_0). \tag{11}$$

Proof: Indeed, for $T = T_{\delta,p,H}(x_0)$ we have

$$\begin{aligned}
F_{x_0,p,H}(T) &\stackrel{(8)}{\leq} \delta + \min_{x \in \mathrm{dom}\, \Psi} \left\{ f_{x_0,p}(x) + \Psi(x) + \frac{H}{(p+1)!} \|x - x_0\|^{p+1} \right\} \\
&\stackrel{(5)}{\leq} \delta + \min_{x \in \mathrm{dom}\, \Psi} \left\{ F(x) + \frac{L_p(f) + H}{(p+1)!} \|x - x_0\|^{p+1} \right\} \\
&\leq \delta + F_* + \frac{L_p(f) + H}{(p+1)!} \|x^* - x_0\|^{p+1} \\
&\stackrel{(9)}{\leq} F_* + \delta + \frac{1}{p+1} C_{p,H}(x_0) \leq F_* + C_{p,H}(x_0).
\end{aligned}$$

■

Thus, after one Inexact Tensor Step, the residuals in the function value do not depend anymore on the size of derivatives of degree smaller than $p+1$.

For analysing the Iteration Process, we need the following result.

Lemma 2.2: *Let* $\bar{x} \in \operatorname{dom} \Psi$ *and*

$$F(\bar{x}) \leq \min\{F(x_0), F_* + C_{p,H}(x_0)\}. \tag{12}$$

Then for $T = T_{\delta,p,H}(\bar{x})$ *we have*

$$F_{\bar{x},p,H}(T) \leq F(\bar{x}) - \frac{p}{p+1}\left[\frac{F(\bar{x}) - F_*}{C_{p,H}(x_0)}\right]^{\frac{1}{p}} (F(\bar{x}) - F_*) + \delta. \tag{13}$$

Proof: Indeed, for $T = T_{\delta,p,H}(\bar{x})$ we have

$$\begin{aligned}
F_{\bar{x},p,H}(T) &\overset{(8)}{\leq} \delta + \min_{x \in \operatorname{dom}\Psi}\left\{f_{\bar{x},p}(x) + \Psi(x) + \frac{H}{(p+1)!}\|x - \bar{x}\|^{p+1}\right\} \\
&\overset{(5)}{\leq} \delta + \min_{0 \leq \alpha \leq 1}\left\{F(x) + \frac{L_p(f) + H}{(p+1)!}\|x - \bar{x}\|^{p+1} : x = \alpha x^* + (1-\alpha)\bar{x}\right\} \\
&\overset{(9)}{\leq} \delta + \min_{0 \leq \alpha \leq 1}\left\{\alpha F_* + (1-\alpha)F(\bar{x}) + \frac{L_p(f) + H}{(p+1)!}\alpha^{p+1}R^{p+1}(x_0)\right\} \\
&= \delta + \min_{0 \leq \alpha \leq 1}\left\{\alpha F_* + (1-\alpha)F(\bar{x}) + \frac{1}{p+1}\alpha^{p+1}C_{p,H}(x_0)\right\}.
\end{aligned}$$

The optimal solution in the latter optimization problem is $\alpha^* = [\frac{F(\bar{x}) - F_*}{C_{p,H}(x_0)}]^{\frac{1}{p}} \leq 1$. It remains to substitute it in the objective function. ■

Now we are ready to analyse the following Inexact Basic Tensor Method.

Initialization. Choose positive parameters δ and H.
Preliminary Step. Compute $T_0 = T_{\bar{\delta},p,H}(x_0)$ with $\bar{\delta} \leq \frac{p}{p+1}C_{p,H}(x_0)$. Set $x_1 = \arg\min_x\{F(x) : x \in \{x_0, T_0\}\}$.
Iteration $k \geq 1$**.** Compute $x_{k+1} = T_{\delta,p,H}(x_k)$.

(14)

Denote by $\epsilon > 0$ the desired accuracy of the approximate solution of problem (3).

Theorem 2.1: *Let sequence* $\{x_k\}_{k=1}^{T+1}$ *be generated by Inexact Basic Tensor Method with*

$$\delta \leq \frac{p}{2(p+1)C_{p,H}^{1/p}(x_0)} \cdot \epsilon^{\frac{p+1}{p}} \tag{15}$$

and $H \geq pL_p(f)$. *Assume also that*

$$F(x_k) - F_* \geq \epsilon, \quad k = 1, \ldots, T+1. \tag{16}$$

Then for all $k = 1, \ldots, T$ *we have*

$$F(x_k) \leq \min\{F(x_0), F_* + C_{p,H}(x_0)\}, \tag{17}$$

$$F(x_{k+1}) \leq F(x_k) - \frac{p}{2(p+1)} \left[\frac{F(x_k) - F_*}{C_{p,H}(x_0)}\right]^{\frac{1}{p}} (F(x_k) - F_*)\,. \tag{18}$$

Moreover,

$$F(x_{T+1}) - F_* \leq \frac{C_{p,H}(x_0)}{\left(1 + \frac{T}{2(p+1)}\right)^p}. \tag{19}$$

Proof: Let us prove the relations (17) and (18) by induction. Since $H > L_p(f)$, we have

$$F(T_0) \leq F_{x_0,p,H}(T_0) \overset{(11)}{\leq} F_* + C_{p,H}(x_0).$$

Hence, in view of the choice of the point x_1, condition (17) is satisfied for $k = 0$.

Assume it is satisfied for some $k \geq 0$. Then, by Lemma 2.2 we have

$$\begin{aligned} F(x_{k+1}) &\overset{(5)}{\leq} F_{x_k,p,H}(x_{k+1}) \overset{(13)}{\leq} F(x_k) - \frac{p}{p+1}\left[\frac{F(x_k) - F_*}{C_{p,H}(x_0)}\right]^{\frac{1}{p}} (F(x_k) - F_*) + \delta \\ &\overset{(15)}{\leq} F(x_k) - \frac{p}{p+1}\left[\frac{F(x_k) - F_*}{C_{p,H}(x_0)}\right]^{\frac{1}{p}} (F(x_k) - F_*) + \frac{p}{2(p+1)C_{p,H}^{1/p}(x_0)} \cdot \epsilon^{\frac{p+1}{p}} \\ &\overset{(16)}{\leq} F(x_k) - \frac{p}{2(p+1)}\left[\frac{F(x_k) - F_*}{C_{p,H}(x_0)}\right]^{\frac{1}{p}} (F(x_k) - F_*)\,. \end{aligned}$$

Hence, $F(x_{k+1}) \leq F(x_k)$ and we see that inequality (17) is valid for x_{k+1}.

Thus, we have proved that inequalities (17) and (18) are valid for al $k = 1, \ldots, T$.

Denoting now $\xi_k = \frac{F(x_k) - F_*}{C_{p,H}(x_0)} \overset{(17)}{\leq} 117$, we can rewrite (18) as follows:

$$\xi_k - \xi_{k+1} \geq \frac{p}{2(p+1)} \xi_k^{\frac{p+1}{p}}, \quad k = 1, \ldots, T. \tag{20}$$

Since function $\frac{1}{(1+\tau)^{1/p}}$ is convex for $\tau > -1$, we have

$$\frac{1}{(1+\tau)^{1/p}} \geq 1 - \frac{\tau}{p}. \tag{21}$$

Hence

$$\frac{\xi_k^{1/p}}{\xi_{k+1}^{1/p}} = \frac{1}{\left(1 + \frac{\xi_{k+1}-\xi_k}{\xi_k}\right)^{1/p}} \geq 1 - \frac{\xi_{k+1} - \xi_k}{p\xi_k}.$$

Consequently,

$$\frac{1}{\xi_{k+1}^{1/p}} - \frac{1}{\xi_k^{1/p}} = \frac{1}{\xi_k^{1/p}}\left(\frac{\xi_k^{1/p}}{\xi_{k+1}^{1/p}} - 1\right) \geq \frac{1}{\xi_k^{1/p}} \cdot \frac{\xi_k - \xi_{k+1}}{p\xi_k} \overset{(20)}{\geq} \frac{1}{2(p+1)}.$$

Summing up these inequalities for $k = 1, \ldots, T$, we get

$$\frac{1}{\xi_{T+1}^{1/p}} \geq \frac{1}{\xi_1^{1/p}} + \frac{T}{2(p+1)} \overset{(17)}{\geq} 1 + \frac{T}{2(p+1)}.$$

And this is inequality (19). ■

Corollary 2.1: *Condition* (16) *cannot remain valid more than for*

$$2(p+1)\left[\frac{C_{p,H}(x_0)}{\epsilon}\right]^{\frac{1}{p}} \tag{22}$$

iterations of method (14).

Surprisingly enough, condition (15) shows that the high-order methods require *less accurate* solutions of the auxiliary problem (7). Thus, for the first-order methods ($p = 1$) we need

$$\delta = O(\epsilon^2).$$

The second-order methods require

$$\delta = O(\epsilon^{3/2}),$$

and for the third-order schemes it is enough to have

$$\delta = O(\epsilon^{4/3}).$$

In the next two sections, for the last two cases, we are going to discuss the iteration complexity of the specific auxiliary methods computing the points $T_{\delta,p,H}(x_k)$.

3. Inexact third-order method

In this section, we consider a third-order method for solving the following problem

$$\min_{x \in \mathrm{dom}\,\Psi} \{F(x) \overset{\mathrm{def}}{=} f(x) + \Psi(x)\}, \tag{23}$$

where function $f(\cdot)$ is convex and its third derivative satisfies Lipschitz condition

$$\|D^3 f(x) - D^3 f(y)\| \leq L_3(f)\|x - y\|, \quad x, y \in \mathbb{E}. \tag{24}$$

Function $\Psi(\cdot)$ in this problem is a simple closed convex function with $\mathrm{dom}\,\Psi \subseteq \mathrm{dom}\, f$. We assume that the problem (23) is solvable and denote by x^* one of its optimal solutions with $F_* \overset{\mathrm{def}}{=} F(x^*)$.

As it was shown in Section 2, problem (23) can be solved by Inexact Tensor Method (14) provided that we are able to compute approximate solutions of the problem (7) satisfying the condition (8). In our case,

$$\varphi_{\bar{x},3,H}(y) = f(\bar{x}) + \langle \nabla f(\bar{x}), y - x\rangle + \frac{1}{2}\langle \nabla^2 f(\bar{x})(y-\bar{x}), y-\bar{x}\rangle + \frac{1}{6}D^3 f(\bar{x})[y-\bar{x}]^3 + \frac{H}{24}\|y-\bar{x}\|^4.$$

As it was suggested in [12], we are going to solve this problem using the framework of *relative smoothness* (see [1,8]). However, for our goals we need to provide its complexity analysis with more details.

Let us consider a minimization problem, which is more general than the problem (7):

$$\Phi_* = \min_{x\in \mathrm{dom}\,\Psi} \left\{\Phi(x) \stackrel{\mathrm{def}}{=} \varphi(x) + \Psi(x)\right\} = \Phi(x^*), \quad x^* \in \mathrm{dom}\,\Psi, \tag{25}$$

where $\varphi(\cdot)$ is a continuously differentiable convex function, and $\Psi(\cdot)$ is a simple closed convex function with $\mathrm{dom}\,\Psi \subseteq \mathrm{dom}\,\varphi$. In our framework, we also need a *simple scaling function* $d(\cdot)$ with $\mathrm{dom}\,d \supseteq \mathrm{dom}\,\Psi$, which is strictly convex and continuously differentiable. Then, we can define the *Bregman distance* between two points x and y from $\mathrm{dom}\,d$:

$$\beta_d(x,y) = d(y) - d(x) - \langle \nabla d(x), y - x\rangle \geq 0.$$

We assume that the smooth part of the objective function in problem (25) satisfies the following *relative smoothness conditions*:

$$\begin{aligned}\mu_\varphi \beta_d(x,y) \leq \beta_\varphi(x,y) &= \varphi(y) - \varphi(x) - \langle \nabla\varphi(x), y - x\rangle \\ &\leq L_\varphi \beta_d(x,y), \quad x,y \in \mathrm{dom}\,\Psi,\end{aligned} \tag{26}$$

where $L_\varphi \geq \mu_\varphi > 0$. Denote $\gamma_\varphi = \frac{\mu_\varphi}{L_\varphi} \leq 1$.

The right-hand side of inequality (26) suggests the following minimization scheme.

Choose $x_0 \in \mathrm{dom}\,\Psi$**. For** $k \geq 0$ **iterate:**

$$x_{k+1} = \arg\min_{x\in\mathrm{dom}\,\Psi} \left\{\varphi(x_k) + \langle \nabla\varphi(x_k), x - x_k\rangle + L_\varphi \beta_d(x_k, x) + \Psi(x)\right\}. \tag{27}$$

Note that the next point in this method satisfies the following variational principle: for all $x \in \mathrm{dom}\,\Psi$ we have

$$\langle \nabla\varphi(x_k) + L_\varphi(\nabla d(x_{k+1}) - \nabla d(x_k)),\ x - x_{k+1}\rangle + \Psi(y) \geq \Psi(x_{k+1}). \tag{28}$$

For analysing performance of method (27), we need to introduce the following aggregated objects.

- The average value of the objective function:

$$\tilde{\Phi}_N = \frac{\gamma_\varphi}{1 - (1-\gamma_\varphi)^N} \sum_{k=1}^{N} (1-\gamma_\varphi)^{N-k} \Phi(x_k).$$

- The aggregated model of the objective function:

$$\ell_N(x) = \frac{\gamma_\varphi}{1-(1-\gamma_\varphi)^N} \sum_{k=0}^{N-1} (1-\gamma_\varphi)^{N-k-1} [\varphi(x_k) + \langle \nabla\varphi(x_k), x - x_k\rangle + \mu_\varphi \beta_d(x_k, x)] + \Psi(x).$$

Note that in view of the first inequality in (26), we have

$$\ell_N(x) \le \Phi(x), \quad x \in \operatorname{dom}\Psi. \tag{29}$$

In the above definitions, we extend our notation onto the case $\gamma_f = 0$ in a continuous way:

$$\lim_{\gamma_\varphi \downarrow 0} \frac{\gamma_\varphi}{1-(1-\gamma_\varphi)^N} = \frac{1}{N}.$$

Let us prove the following result.

Lemma 3.1: *For any $N \ge 1$ we have:*

$$\frac{1}{L_\varphi}\tilde{\Phi}_N + \beta_d(x_N, x) \le (1-\gamma_\varphi)^N \beta_d(x_0, x) + \frac{1}{L_\varphi}\ell_N(x), \quad x \in \operatorname{dom}\Psi. \tag{30}$$

Proof: Note that for any $k \ge 0$ we have:

$$\begin{aligned}
\beta_d(x_{k+1}, x) - \beta_d(x_k, x) &= d(x) - d(x_{k+1}) - \langle \nabla d(x_{k+1}), x - x_{k+1}\rangle \\
&\quad - d(x) + d(x_k) + \langle \nabla d(x_k), x - x_k\rangle \\
&= \langle \nabla d(x_k) - \nabla d(x_{k+1}), x - x_{k+1}\rangle \\
&\quad - d(x_{k+1}) + d(x_k) + \langle \nabla d(x_k), x_{k+1} - x_k\rangle \\
&= \langle \nabla d(x_k) - \nabla d(x_{k+1}), x - x_{k+1}\rangle - \beta_d(x_k, x_{k+1}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\beta_d(x_{k+1}, x) &\overset{(28)}{\le} \beta_d(x_k, x) + \frac{1}{L_\varphi}[\langle \nabla\varphi(x_k), x - x_{k+1}\rangle + \Psi(x) - \Psi(x_{k+1})] - \beta_d(x_k, x_{k+1}) \\
&= \beta_d(x_k, x) + \frac{1}{L_\varphi}[\varphi(x_k) + \langle \nabla\varphi(x_k),\ x - x_k\rangle + \Psi(x)] \\
&\quad - \frac{1}{L_\varphi}[\varphi(x_k) + \langle \nabla\varphi(x_k), x_{k+1} - x_k\rangle + L_\varphi \beta_d(x_k, x_{k+1}) + \Psi(x_{k+1})] \\
&\overset{(26)}{\le} \beta_d(x_k, x) + \frac{1}{L_\varphi}[\varphi(x_k) + \langle \nabla\varphi(x_k), x - x_k\rangle + \Psi(x)] - \frac{1}{L_\varphi}\Phi(x_{k+1}) \\
&= (1-\gamma_\varphi)\beta_d(x_k, x) - \frac{1}{L_\varphi}\Phi(x_{k+1}) \\
&\quad + \frac{1}{L_\varphi}[\varphi(x_k) + \langle \nabla\varphi(x_k), x - x_k\rangle + \mu_\varphi \beta_d(x_k, x) + \Psi(x)].
\end{aligned}$$

Summing up these inequalities for $k = 0, \ldots, N-1$, we get the relation (30). ■

Lemma 3.1 has two important consequences. First of all, we can estimate the rate of convergence of method (27).

Corollary 3.1: *For any $N \geq 1$ we have*

$$\tilde{\Phi}_N - \Phi_* \leq (1 - \gamma_\varphi)^N L_\varphi \beta_d(x_0, x^*). \tag{31}$$

Proof: Indeed, it is enough to apply inequalities (30) and (29) to $x = x^*$. ■

Another important consequence of inequality (30) is the verifiable stopping criterion for method (27). Assume that we know an upper bound D_0 for the size of the solution:

$$\beta_d(x_0, x^*) \leq D_0. \tag{32}$$

Then,

$$\begin{aligned}
\frac{1}{L_\varphi}\tilde{\Phi}_N &\overset{(30)}{\leq} \min_x \left\{ (1-\gamma_\varphi)^N \beta_d(x_0, x) + \frac{1}{L_\varphi}\ell_N(x) : \ \beta_d(x_0, x) \leq D_0 \right\} \\
&\leq (1-\gamma_\varphi)^N D_0 + \frac{1}{L_\varphi} \min_x \{\ell_N(x) : \ \beta_d(x_0, x) \leq D_0\} \\
&\overset{(29)}{\leq} (1-\gamma_\varphi)^N D_0 + \frac{1}{L_\varphi}\Phi(x^*).
\end{aligned}$$

Thus,

$$\tilde{\Phi}_N - \Phi_* \leq \tilde{\Phi}_N - \min_x \{\ell_N(x) : \ \beta_d(x_0, x) \leq D_0\} \leq (1-\gamma_\varphi)^N L_\varphi D_0. \tag{33}$$

Note that the lower bound for the optimal value Φ_*, the estimate

$$\ell_N^* = \min_x \{\ell_N(x) : \ \beta_d(x_0, x) \leq D_0\}, \tag{34}$$

can be easily computed. In order to satisfy the stopping criterion (33) with accuracy $\delta > 0$, it is sufficient to ensure the inequality

$$(1-\gamma_\varphi)^N L_\varphi D_0 \leq e^{-\gamma_\varphi N} L_\varphi D_0 \leq \delta.$$

Thus, we need

$$\frac{L_\varphi}{\mu_\varphi} \ln \frac{L_\varphi D_0}{\delta} \tag{35}$$

iterations at most.

Let us show how this machinery works for Inexact Basic Tensor Method of degree three. For simplicity, we consider the case $\Psi(x) \equiv 0$, $x \in \mathbb{E}$. For computing the result of Inexact

Tensor Step from a point $\bar{x} \in \mathbb{E}$, we need to solve the auxiliary problem (25) with

$$\Phi(x) = \varphi(x) \stackrel{\text{def}}{=} \varphi_{\bar{x},3,H}(x).$$

In accordance to Section 5 of [12], a natural scaling function for this problem is as follows:

$$d_{\tau,\bar{x}}(x) = \frac{1}{2}\left(1 - \frac{1}{\tau}\right)\langle \nabla^2 f(\bar{x})(x - \bar{x}), x - \bar{x}\rangle + \frac{3\tau(\tau-1)L_3(f)}{24}\|x - \bar{x}\|^4, \tag{36}$$

where $\tau = \sqrt{\frac{H}{3L_3(f)}} > 1$. Then

$$\nabla^2 d_{\tau,\bar{x}}(x) \preceq \nabla^2 \varphi(x) \preceq \frac{\tau+1}{\tau-1}\nabla^2 d_{\tau,\bar{x}}(x). \quad x \in \mathbb{E}.$$

Let us choose $\tau = 2$ (this corresponds to $H = 12L_3(f)$). Then $\mu_\varphi = 1, L_\varphi = 3$, and $\gamma_\varphi = \frac{1}{3}$. Thus, with this choice, the method (27) has global linear rate of convergence dependent only on the absolute constant.

At the same time, since function $\rho_4(x) = \frac{1}{4}\|x\|^4$ is uniformly convex with constant $\sigma_4 = \frac{1}{4}$ (see, for example, Lemma 4 in [10]), we have

$$\|\nabla f(\bar{x})\|_* \|\bar{x} - \bar{x}^*\| \geq \langle \nabla f(\bar{x}), \bar{x} - \bar{x}^*\rangle = \langle \nabla \varphi(\bar{x}), \bar{x} - \bar{x}^*\rangle \geq \sigma_4 \|\bar{x} - \bar{x}^*\|^4,$$

where $\bar{x}^* = \arg\min_{x \in \mathbb{E}} \varphi_{\bar{x},3,H}(x)$. This means that we can choose

$$D_0 = \left(4\|\nabla f(\bar{x})\|_*\right)^{\frac{1}{3}} \tag{37}$$

for computing by (34) the lower estimate ℓ_N^* of the optimal function value. It can be used in the stopping criterion of the Inexact Tensor Method:

$$\tilde{\Phi}_N - \ell_N^* \leq \delta. \tag{38}$$

In accordance to (33), this inequality will be satisfied in $3 \ln \frac{L_\varphi D_0}{\delta}$ iterations at most.

It remains to discuss the complexity of one iteration of the process (27) with scaling function (36). Without loss of generality, we can assume that $\bar{x} = 0$. And let $\mathbb{E} \equiv \mathbb{R}^n$ with $B = I_n$, the identity matrix. In this case, $\nabla^2 f(0)$ is a symmetric $n \times n$-matrix, which can be factorized as follows:

$$\nabla^2 f(0) = UTU^{\mathrm{T}},$$

where $U \in \mathbb{R}^{n \times n}$, $UU^{\mathrm{T}} = I_n$, and $T \in \mathbb{R}^{n \times n}$ is a symmetric tri-diagonal matrix. Introducing new variables $y = U^{\mathrm{T}}x$, we can rewrite our objective function as follows:

$$\begin{aligned}\varphi(x) &= \langle \nabla f(0), x\rangle + \frac{1}{2}\langle \nabla^2 f(0), x, x\rangle + \frac{1}{6}D^3 f(0)[x]^3 + \frac{H}{24}\|x\|_{(2)}^4 \\ &= \langle g_y, y\rangle + \frac{1}{2}\langle Ty, y\rangle + \frac{1}{6}D^3 f(0)[Uy]^3 + \frac{H}{24}\|y\|_{(2)}^4 \stackrel{\text{def}}{=} \xi(y),\end{aligned}$$

where $g_y = U^{\mathrm{T}}\nabla f(0)$. Similarly, our scaling function (36) can be written in the following form:

$$\tilde{d}_\tau(y) = \frac{1}{2}\left(1 - \frac{1}{\tau}\right)\langle Ty, y\rangle + \frac{3\tau(\tau-1)L_3(f)}{24}\|y\|^4.$$

This transformation can be done by the standard Linear Algebra technique in $O(n^3)$ operations. Now, we can apply method (27) with $\Psi(\cdot) \equiv 0$ for minimizing function $\xi(y)$ using the scaling function $\tilde{d}_\tau(\cdot)$. At each iteration of this scheme, the most expensive operations are as follows.

- Computation of the gradient

$$\nabla\xi(y) = g_y + Ty + \frac{1}{2}U^T D^3 f(0)[Uy]^2 + \frac{H}{6}\|y\|^2 y.$$

  This computation needs $O(n^2)$ operations plus the operations necessary for computing the vector $D^3 f(0)[h]^2 \in \mathbb{R}^n$ with $h \in \mathbb{R}^n$. The second computation usually is not very expensive. Indeed, in many situations it can be arranged by the fast backward differentiation which complexity is proportional to the complexity of computing the function value $f(x)$ (see, for example, [7]).
- Computation of the next point by solving the auxiliary optimization problem

$$\min_{y\in\mathbb{R}^n}\{\langle\nabla\xi(y_k), y - y_k\rangle + L_\varphi\beta_{\tilde{d}_\tau}(y_k, y)\},$$

  where $\tilde{d}_\tau(y) = \frac{a}{2}\langle Ty, y\rangle + \frac{b}{4}\|y\|^4_{(2)}$ with certain positive constants a and b. In view of definition of the Bregmann distance, this problem is as follows:

$$\begin{aligned}&\min_{y\in\mathbb{R}^n}\left\{\langle g_k, y\rangle + L_\varphi\left(\frac{a}{2}\langle Ty, y\rangle + \frac{b}{4}\|y\|^4_{(2)}\right)\right\}\\ &\quad = \min_{y\in\mathbb{R}^n}\max_{r\geq 0}\left\{\langle g_k, y\rangle + L_\varphi\left(\frac{a}{2}\langle Ty, y\rangle + \frac{b}{2}r\|y\|^2_{(2)}\right) - L_\varphi\frac{b}{4}r^2\right\},\end{aligned}$$

  where $g_k = \nabla\xi(y_k) - L_\varphi(aTy_k + b\|y_k\|^2_{(2)}y_k)$. Exchanging in this presentation minimum and maximum, we come to the following univariate dual problem:

$$\max_{r\geq 0}\left\{-L_\varphi\frac{b}{4}r^2 - \frac{1}{2L_\varphi}\langle[aT + brI_n]^{-1}g_k, g_k\rangle\right\},$$

  which can be easily solved by any method for univariate convex minimization. Note that the computational cost of the function values and the derivatives in this problem is proportional to n.

Thus, we have seen that the computational cost of one iteration of the Inexact Tensor Method is basically the same as that of the standard second-order schemes.

4. Inexact second-order method

Let us study now the efficient implementations of the second-order methods for solving the following minimization problem:

$$\min_{x\in\mathbb{E}} f(x), \tag{39}$$

where function $f(\cdot)$ is convex and its second derivative satisfies Lipschitz condition

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L_2(f)\|x-y\|, \quad x,y\in\mathbb{E}. \tag{40}$$

As usual, we assume that the problem (39) is solvable and x^* is one of its optimal solutions with $f_* = f(x^*)$.

Inexact Basic Tensor Method (14) becomes now an Inexact Newton Method with Cubic Regularization. The efficiency of an exact version of this scheme was studied first in [14]. In its inexact variant, we need to compute an approximate solution of problem (7) with

$$F_{\bar{x},2,H}(y) = f(\bar{x}) + \langle \nabla f(\bar{x}), y-x\rangle + \frac{1}{2}\langle \nabla^2 f(\bar{x})(y-\bar{x}), y-\bar{x}\rangle + \frac{H}{6}\|y-\bar{x}\|^3. \tag{41}$$

Let us show how this can be done by the *first-order methods*.

Let us consider the auxiliary minimization problem (7) in a more general form:

$$\min_{x\in\mathrm{dom}\,\psi} \left\{\Phi(x) \stackrel{\mathrm{def}}{=} \varphi(x) + \psi(x)\right\}, \tag{42}$$

where $\varphi(\cdot)$ is a continuously differentiable convex function, and $\psi(\cdot)$ is a simple closed convex function with $\mathrm{dom}\,\psi \subseteq \mathrm{dom}\,\varphi$. The role of function $\psi(\cdot)$ here is different from the role of $\Psi(\cdot)$ in (7). Since in (39) $\Psi(\cdot)\equiv 0$, we use function $\psi(\cdot)$ to model the regularization term in the objective function (41).

Let us assume that the gradient of function $\varphi(\cdot)$ is Lipschitz continuous:

$$\|\nabla\varphi(x) - \nabla\varphi(y)\|_* \le L_1(\varphi)\|x-y\|, \quad x,y\in\mathrm{dom}\,\psi. \tag{43}$$

For convex function, this condition is equivalent to the following inequality (see, for example, Section 2.1.1 in [13]):

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x-y\rangle \ge \frac{1}{L_1(\varphi)}\|\nabla\varphi(x) - \nabla\varphi(y)\|_*^2, \quad x,y\in\mathrm{dom}\,\psi. \tag{44}$$

On the other hand, we assume that function $\psi(\cdot)$ is subdifferentiable and *uniformly convex* of degree $p+1$ with $p\ge 1$:

$$\psi(y) \ge \psi(x) + \langle g_x, y-x\rangle + \frac{\sigma_{p+1}}{p+1}\|y-x\|^{p+1}, \quad x,y\in\mathrm{dom}\,\psi, \tag{45}$$

where $g_x\in\partial\psi(x)$ and the *parameter of uniform convexity* σ_p is positive.

Note that usually the assumptions on smoothness and uniform convexity are introduced for the whole objective function $\Phi(\cdot)$. However, we will see that the separation of these assumptions allows us to construct much faster algorithms. It seems that in the first time such a separation was studied in [11] for a strongly convex composite part $\psi(\cdot)$.

Let us estimate efficiency of Composite Gradient Method as applied to problem (42).

$$\boxed{\begin{array}{l}\textbf{Choose } x_0 \in \operatorname{dom}\psi\textbf{. For } k \geq 0 \textbf{ iterate:} \\ \\ x_{k+1} = \arg\min\limits_{x\in\operatorname{dom}\psi}\left\{\varphi(x_k) + \langle\nabla\varphi(x_k), x - x_k\rangle + \frac{1}{2}H\|x - x_k\|^2 + \psi(x)\right\}.\end{array}} \tag{46}$$

Lemma 4.1: *Let $H \geq L_1(\varphi)$. Then*

$$\Phi(x_k) - \Phi(x_{k+1}) \geq \frac{1}{2H}\|\Phi'(x_{k+1})\|_*^2, \tag{47}$$

where $\Phi'(x_{k+1}) = \nabla\varphi(x_{k+1}) - \nabla\varphi(x_k) - HB(x_{k+1} - x_k) \in \partial\Phi(x_{k+1})$.

Proof: The first-order optimality condition for the auxiliary minimization problem in (46) is as follows:

$$\langle\nabla\varphi(x_k) + HB(x_{k+1} - x_k), x - x_{k+1}\rangle + \psi(x) \geq \psi(x_{k+1}), \quad \forall\, x \in \operatorname{dom}\psi. \tag{48}$$

This means that the vector $g_k = -(\nabla\varphi(x_k) + HB(x_{k+1} - x_k))$ belongs to the subdifferential $\partial\psi(x_{k+1})$. At the same time,

$$\begin{aligned}\Phi(x_{k+1}) &\overset{(5)}{\leq} \varphi(x_k) + \langle\nabla\varphi(x_k), x_{k+1} - x_k\rangle + \frac{H}{2}\|x_{k+1} - x_k\|^2 + \psi(x_{k+1}) \\ &\overset{(48)}{\leq} \varphi(x_k) + \langle\nabla\varphi(x_k), x_{k+1} - x_k\rangle + \frac{H}{2}\|x_{k+1} - x_k\|^2 \\ &\quad + \langle\nabla\varphi(x_k) + HB(x_{k+1} - x_k), x_k - x_{k+1}\rangle + \psi(x_k) \\ &= \Phi(x_k) - \frac{1}{2}H\|x_{k+1} - x_k\|^2.\end{aligned}$$

Let us define now $\Phi'(x_{k+1}) = \nabla\varphi(x_{k+1}) + g_k \in \partial\Phi(x_{k+1})$. Note that

$$\begin{aligned}\|\Phi'(x_{k+1})\|_*^2 &= \|\nabla\varphi(x_{k+1}) - \nabla\varphi(x_k) - HB(x_{k+1} - x_k))\|_*^2 \\ &= \|\nabla\varphi(x_{k+1}) - \nabla\varphi(x_k)\|_*^2 - 2H\langle\nabla\phi(x_{k+1}) \\ &\quad - \nabla\varphi(x_k), x_{k+1} - x_k\rangle + H^2\|x_{k+1} - x_k\|^2 \\ &\overset{(44)}{\leq} \left(1 - \frac{2H}{L_1(\varphi)}\right)\|\nabla\varphi(x_{k+1}) - \nabla\varphi(x_k)\|_*^2 \\ &\quad + H^2\|x_{k+1} - x_k\|^2 \leq H^2\|x_{k+1} - x_k\|^2.\end{aligned}$$

Substituting this inequality in the previous one, we get inequality (47). ■

Now we can prove the rate of convergence of the method (46).

Theorem 4.1: *Let $H \geq L_1(\varphi)$. Then*

$$\Phi(x_k) - \Phi_* \leq \left[\left(1 + \frac{1}{\alpha}\right) (B_{p,H}^{-\alpha} + (\Phi(x_0) - \Phi_*)^\alpha) \cdot \frac{1}{t} \right]^{\frac{1}{\alpha}}, \tag{49}$$

where $B_{p,H} \stackrel{\text{def}}{=} (\frac{1}{2H}\, \sigma_{p+1}^{\frac{2}{p+1}} (\frac{p+1}{p})^{1+\alpha})^{\frac{1}{\alpha}}$ *and* $\alpha = \frac{p-1}{p+1}$.

Proof: In view of inequality (47), we have:

$$\begin{aligned} \Phi(x_k) - \Phi(x_{k+1}) &\stackrel{(2)}{\geq} \frac{1}{2H} \left(\frac{p+1}{p} \sigma_{p+1}^{\frac{1}{p}} (\Phi(x_{k+1}) - \Phi_*) \right)^{\frac{2p}{p+1}} \\ &= \frac{1}{2H} \sigma_{p+1}^{\frac{2}{p+1}} \left(\frac{p+1}{p} \right)^{1+\alpha} (\Phi(x_{k+1}) - \Phi_*)^{1+\alpha}, \end{aligned}$$

where $\alpha = \frac{2p}{p+1} - 1 = \frac{p-1}{p+1}$. Denoting now $\xi_k = B_{p,H}(\Phi(x_k) - \Phi_*)$, we get

$$\xi_k - \xi_{k+1} \geq \xi_{k+1}^{1+\alpha}, \quad k \geq 0.$$

Hence, in view of Lemma A.1,

$$\begin{aligned} \Phi(x_k) - \Phi_* &\leq B_{p,H}^{-1} \left[\left(1 + \frac{1}{\alpha}\right) (1 + B_{p,H}^{\alpha} (\Phi(x_0) - \Phi_*)^\alpha) \cdot \frac{1}{t} \right]^{\frac{1}{\alpha}} \\ &= \left[\left(1 + \frac{1}{\alpha}\right) (B_{p,H}^{-\alpha} + (\Phi(x_0) - \Phi_*)^\alpha) \cdot \frac{1}{t} \right]^{\frac{1}{\alpha}}. \end{aligned}$$

And this is inequality (49). ■

Let us apply the result of Theorem 4.1 to the objective function (41). In this case,

$$\psi(y) = \frac{H}{6} \|y - \bar{x}\|^3.$$

Hence, $p = 2$ and $\alpha = \frac{p-1}{p+1} = \frac{1}{3}$. Note that for this choice of the composite term we have

$$\sigma_3 = \frac{H}{2} \left(\frac{1}{2}\right)^{p-1} = \frac{H}{4}$$

(see, for example, Lemma 4 in [10]). Therefore,

$$B_{2,H} = \left(\frac{1}{2H} \left(\frac{H}{4}\right)^{\frac{2}{3}} \left(\frac{4}{3}\right)^{4/3} \right)^3 = \frac{2}{81H}.$$

Thus, the rate of convergence of the Composite Gradient Method (46) is of the order

$$O\left(\frac{1}{k^3}\right).$$

This is much faster than we could expect from a non-accelerated gradient scheme. Recall that the usual rate of convergence of such a method, which does not take into account the properties of composite term, is of the order $O(\frac{1}{k})$ (see [11]).

Having in mind this encouraging observation, let us look at the performance of the Composite Fast Gradient Method as applied to problem (42). For the convenience of readers, let us present here a variant of this method with predefined number of iterations.

Input: Starting point $x_0 \in \operatorname{dom}\psi$ and number of steps $N \geq 1$.

Define $\Omega_0(x) = \frac{1}{2}\|x - x_0\|^2$ and set $A_0 = 0$, $v_0 = x_0$.

For $k = 0, \ldots, N-1$ **do**

(a) Choose $H_k > 0$ and find a_{k+1} from the equation $\frac{a_{k+1}^2}{a_{k+1}+A_k} = \frac{1}{H_k}$.

(b) Set $A_{k+1} = A_k + a_{k+1}$, $\tau_k = \frac{a_{k+1}}{A_{k+1}}$, and $y_k = (1-\tau_k)x_k + \tau_k v_k$.

(c) Update $\Omega_{k+1}(x) = \Omega_k(x) + a_{k+1}[\varphi(y_k) + \langle \nabla\varphi(y_k), x - y_k\rangle + \psi(x)]$.

(d) Compute $v_{k+1} = \arg\min_{x\in\operatorname{dom}\psi} \Omega_{k+1}(x)$ and $x_{k+1} = (1-\tau_k)x_k + \tau_k v_{k+1}$.

(e) Accept x_{k+1} if $\varphi(x_{k+1}) \leq \varphi(y_k) + \langle \nabla\varphi(y_k), x - y_k\rangle + \frac{H_k}{2}\|x_{k+1} - x_k\|^2$.

Output: Point $T(x_0, N) \stackrel{\text{def}}{=} x_N$.

(50)

In view of Rule (c) of method (50), it is easy to see that for all $x \in \operatorname{dom}\psi$ we have

$$\Omega_k(x) \leq \frac{1}{2}\|x - x_0\|^2 + A_k\Phi(x). \tag{51}$$

On the other hand, we can prove the following statement.

Lemma 4.2: *Let the acceptance condition be satisfied at all iterations of method* (50)*. Then for all* $k = 0, \ldots, N$ *we have*

$$A_k\Phi(x_k) \leq \Omega_k^* \stackrel{\text{def}}{=} \min_{x\in\operatorname{dom}\psi} \Omega_k(x). \tag{52}$$

Proof: Note that all functions $\Omega_k(\cdot)$ formed by this method are strongly convex with convexity parameter one. Since $A_0 = 0$, condition is trivial for $k = 0$. Let us assume that it is satisfied for some $k \geq 0$. Then

$$\begin{aligned}
\Omega_{k+1}^* &= \Omega_k(v_{k+1}) + a_{k+1}[\varphi(y_k) + \langle \nabla\varphi(y_k), v_{k+1} - y_k\rangle + \psi(v_{k+1})] \\
&\geq \Omega_k^* + \tfrac{1}{2}\|v_{k+1} - v_k\|^2 + a_{k+1}[\varphi(y_k) + \langle \nabla\varphi(y_k), v_{k+1} - y_k\rangle + \psi(v_{k+1})] \\
&\geq A_k\varphi(x_k) + a_{k+1}[\varphi(y_k) + \langle \nabla\varphi(y_k), v_{k+1} - y_k\rangle] + \tfrac{1}{2}\|v_{k+1} - v_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ A_k\psi(x_k) + a_{k+1}\psi(v_{k+1}) \\
&\geq A_{k+1}\varphi(y_k) + \langle \nabla\varphi(y_k), a_{k+1}(v_{k+1} - y_k) + A_k(x_k - y_k)\rangle + \tfrac{1}{2}\|v_{k+1} - v_k\|^2 \\
&\quad + A_{k+1}\psi(x_{k+1}) \\
&= A_{k+1}\left[\varphi(y_k) + \langle \nabla\varphi(y_k), x_{k+1} - y_k\rangle + \frac{H_k}{2}\|x_{k+1} - y_k\|^2\right] + A_{k+1}\psi(x_{k+1}) \\
&\overset{\text{(e)}}{\geq} A_{k+1}\Phi(x_{k+1}).
\end{aligned}
$$

■

Finally, we need to estimate the rate of growth of coefficients A_k.

Lemma 4.3: *Let all H_k in the method (50) satisfy condition*

$$H_k \leq \gamma L_1(\varphi), \quad k = 0, \ldots, N-1. \tag{53}$$

Then for all $k = 0, \ldots, N$ we have

$$A_k \geq \frac{k(k+2)}{4\gamma L_1(\varphi)}. \tag{54}$$

Proof: Indeed, for $k = 0$ the inequality (54) is valid. Assume that it is valid for some $k \geq 0$. Then, in view of the Rule (a) of the method, we have

$$a_{k+1} = \frac{1}{2H_k}\left[1 + \sqrt{1 + 4H_kA_k}\right] \overset{(53)}{\geq} \frac{1}{2\gamma L_1(\varphi)}\left[1 + \sqrt{1 + 4\gamma L_1(\varphi)A_k}\right] \geq \frac{k+2}{2\gamma L_1(\varphi)}.$$

Therefore,

$$A_{k+1} = A_k + a_{k+1} \geq \frac{k(k+2) + 2(k+2)}{4\gamma L_1(\varphi)} \geq \frac{(k+1)(k+3)}{4\gamma L_1(\varphi)}.$$

■

In method (50), it is possible to introduce an internal search procedure for updating the parameters H_k, which estimate the actual Lipschitz constant $L_1(\varphi)$ in (43). For practical efficiency of this method, it is important to keep H_k as small as possible. However, for the sake of simplicity, in what follows we assume that the constant $L_1(\varphi)$ is known and we take

$$H_k = L_1(\varphi), \quad k = 0, \ldots, N-1. \tag{55}$$

In this case, in view of inequality (5) with $p = 1$, the acceptance condition is satisfied. Moreover, we have

$$A_k \geq \frac{k(k+2)}{4L_1(\varphi)}. \tag{56}$$

Consequently, the estimates (51) and (52) give us the following rate of convergence:

$$\Phi(x_k) - \Phi_* \leq \frac{2L_1(\varphi)\|x_0 - x^*\|^2}{k(k+2)}, \quad k = 1, \ldots, N. \tag{57}$$

We can summarize our observations in the following statement.

Theorem 4.2: *Let $T = T(x_0, N)$ be the output of the process (50) with parameters H_k defined by (55). Then*

$$\Phi(T) - \Phi_* \leq \frac{2L_1(\varphi)}{N(N+2)} \left[\frac{p+1}{\sigma_{p+1}} (\Phi(x_0) - \Phi_*) \right]^{\frac{2}{p+1}}. \tag{58}$$

Proof: Indeed, since the composite term of the objective function $\Phi(\cdot)$ in problem (42) is uniformly convex, we have

$$\frac{\sigma_{p+1}}{p+1} \|x_0 - x^*\|^{p+1} \leq \Phi(x_0) - \Phi_*.$$

It remains to use inequality (57). ■

Let us consider now the following upper-level process for solving the problem (7). Recall that $\alpha = \frac{p-1}{p+1}$.

Input: Point $x_0 \in \operatorname{dom} \psi$ and parameter $\varkappa > 1$.
For $t \geq 0$ **iterate:** 1. Define $N_t = \left\lceil \sqrt{1 + \varkappa^{\alpha t}} - 1 \right\rceil$ and compute $\hat{x}_{t+1} = T(x_t, N_t)$. 2. Choose $x_{k+1} = \arg\min_x \{\Phi(x) : \ x \in \{x_0, \ldots, x_t, \hat{x}_{t+1}\}\}$.

(59)

Performance of this method is described by the following statement.

Lemma 4.4: *Let $t_0 \geq 0$ be the first integer such that*

$$\varkappa^{t_0} \geq \frac{1}{\Phi(x_0) - \Phi_*} \left[2L_1(\varphi)\varkappa \left(\frac{p+1}{\sigma_{p+1}} \right)^{\frac{2}{p+1}} \right]^{\frac{1}{\alpha}}. \tag{60}$$

Then for all $t \geq t_0$ we have

$$\Phi(x_t) - \Phi_* \leq \left(\frac{1}{\varkappa} \right)^{t-t_0} (\Phi(x_0) - \Phi_*). \tag{61}$$

Proof: In view of Rule 2 in method (59), for $t = t_0$, inequality (61) is trivial. Assume it is valid for some $t \geq t_0$. Note that the choice of N_t in method (59) ensures the following inequality:

$$N_t(N_t + 2) \geq \varkappa^{\alpha t}. \tag{62}$$

Therefore,

$$\Phi(x_{t+1}) - \Phi_* \overset{(58)}{\leq} \frac{2L_1(\varphi)}{N_t(N_t+2)} \left[\frac{p+1}{\sigma_{p+1}} (\Phi(x_t) - \Phi_*) \right]^{\frac{2}{p+1}}$$

$$
\begin{aligned}
&\overset{(62)}{\leq} \frac{2L_1(\varphi)}{\varkappa^{\alpha t}}\left[\frac{p+1}{\sigma_{p+1}}(\Phi(x_t)-\Phi_*)\right]^{\frac{2}{p+1}}\\
&\overset{(61)}{\leq} \frac{2L_1(\varphi)}{\varkappa^{\alpha t}}\left[\frac{p+1}{\sigma_{p+1}}\left(\frac{1}{\varkappa}\right)^{t-t_0}(\Phi(x_0)-\Phi_*)\right]^{\frac{2}{p+1}}\\
&= \frac{2L_1(\varphi)\varkappa^{-\alpha t_0}}{\varkappa^{t-t_0}}\left[\frac{p+1}{\sigma_{p+1}}(\Phi(x_0)-\Phi_*)\right]^{\frac{2}{p+1}}\\
&\leq \frac{2L_1(\varphi)}{\varkappa^{t-t_0}}\left[\frac{p+1}{\sigma_{p+1}}(\Phi(x_0)-\Phi_*)\right]^{\frac{2}{p+1}}(\Phi(x_0)-\Phi_*)^{\alpha}\\
&\quad\times\left[\frac{1}{2L_1(\varphi)\varkappa}\left(\frac{\sigma_{p+1}}{p+1}\right)^{\frac{2}{p+1}}\right]\\
&= \left(\frac{1}{\varkappa}\right)^{t+1-t_0}(\Phi(x_0)-\Phi_*).
\end{aligned}
$$

■

After K iterations of the procedure (59), let us estimate the total number M_K of the low-level steps, which is equal to the number of calls of oracle of problem (42).

Consider first the situation $t_0 > 0$. Note that at each iteration $t \geq 0$ we have

$$
N_t \leq \sqrt{1+\varkappa^{\alpha t}} = \varkappa^{\frac{\alpha}{2}t}\sqrt{1+\varkappa^{-\alpha t}} \leq \varkappa^{\frac{\alpha}{2}t}(1+\frac{1}{2}\varkappa^{-\alpha t}) = \varkappa^{\frac{\alpha}{2}t}+\frac{1}{2}\varkappa^{-\frac{\alpha}{2}t}.
$$

Therefore,

$$
\begin{aligned}
M_K &= \sum_{t=0}^{K} N_t \leq \frac{1}{2}\sum_{t=0}^{K}\varkappa^{-\frac{\alpha}{2}t}+\sum_{t=0}^{K}\varkappa^{\frac{\alpha}{2}t} \leq \frac{1}{2(1-\varkappa^{-\frac{\alpha}{2}})}+\frac{\varkappa^{\frac{\alpha}{2}(K+1)}-1}{\varkappa^{\frac{\alpha}{2}}-1}\\
&= \frac{\varkappa^{\frac{\alpha}{2}}-2}{2(\varkappa^{\frac{\alpha}{2}}-1)}+\frac{\varkappa^{\frac{\alpha}{2}(K+1)}}{\varkappa^{\frac{\alpha}{2}}-1}
\end{aligned}
$$

If $\epsilon \leq \Phi(x_K)-\Phi_* \overset{(61)}{\leq} (\frac{1}{\varkappa})^{K-t_0}(\Phi(x_0)-\Phi_*)$61, then $\varkappa^K \leq \frac{1}{\epsilon}(\Phi(x_0)-\Phi_*)\varkappa^{t_0}$. Hence,

$$
M_K \leq \frac{\varkappa^{\frac{\alpha}{2}}-2}{2(\varkappa^{\frac{\alpha}{2}}-1)}+\frac{\varkappa^{\frac{\alpha}{2}(t_0+1)}}{\varkappa^{\frac{\alpha}{2}}-1}\left[\frac{1}{\epsilon}(\Phi(x_0)-\Phi_*)\right]^{\frac{\alpha}{2}}.
$$

On the other hand, since $t_0 > 0$, we have

$$
\varkappa^{t_0-1} \leq \frac{1}{\Phi(x_0)-\Phi_*}\left[2L_1(\varphi)\varkappa\left(\frac{p+1}{\sigma_{p+1}}\right)^{\frac{2}{p+1}}\right]^{\frac{1}{\alpha}}.
$$

Hence,

$$
M_K \leq \frac{\varkappa^{\frac{\alpha}{2}}-2}{2(\varkappa^{\frac{\alpha}{2}}-1)}+\left[\frac{1}{\epsilon}\right]^{\frac{\alpha}{2}}\frac{\varkappa^{\alpha}}{\varkappa^{\frac{\alpha}{2}}-1}\left[2L_1(\varphi)\varkappa\left(\frac{p+1}{\sigma_{p+1}}\right)^{\frac{2}{p+1}}\right]^{\frac{1}{2}}.
$$

For the second case, we have $t_0 = 0$ if and only if

$$1 \geq \frac{1}{\Phi(x_0) - \Phi_*} \left[2L_1(\varphi) \varkappa \left(\frac{p+1}{\sigma_{p+1}} \right)^{\frac{2}{p+1}} \right]^{\frac{1}{\alpha}} \Leftrightarrow \Phi(x_0) - \Phi_*$$
$$\geq \left[2L_1(\varphi) \varkappa \left(\frac{p+1}{\sigma_{p+1}} \right)^{\frac{2}{p+1}} \right]^{\frac{1}{\alpha}}.$$

In this case, if $\epsilon \leq \Phi(x_K) - \Phi_* \leq (\frac{1}{\varkappa})^K (\Phi(x_0) - \Phi_*)$, then $\varkappa^K \leq \frac{1}{\epsilon}(\Phi(x_0) - \Phi_*)$. Hence,

$$M_K \leq \frac{\varkappa^{\frac{\alpha}{2}} - 2}{2(\varkappa^{\frac{\alpha}{2}} - 1)} + \left[\frac{1}{\epsilon} \right]^{\frac{\alpha}{2}} \frac{\varkappa^{\frac{\alpha}{2}}}{\varkappa^{\frac{\alpha}{2}} - 1} \cdot (\Phi(x_0) - \Phi_*)^{\frac{\alpha}{2}}$$
$$\leq \frac{\varkappa^{\frac{\alpha}{2}} - 2}{2(\varkappa^{\frac{\alpha}{2}} - 1)} + \left[\frac{1}{\epsilon} \right]^{\frac{\alpha}{2}} \frac{\varkappa^{\alpha}}{\varkappa^{\frac{\alpha}{2}} - 1} \cdot (\Phi(x_0) - \Phi_*)^{\frac{\alpha}{2}}.$$

Thus, we have proved the following theorem.

Theorem 4.3: *If the point x_K, $K \geq 1$, in the process (59) satisfies inequality*

$$\Phi(x_K) - \Phi_* \geq \epsilon,$$

then the total number M_K of the lower-level steps in this method cannot exceed the following bound:

$$M_K \leq \frac{\varkappa^{\frac{\alpha}{2}} - 2}{2(\varkappa^{\frac{\alpha}{2}} - 1)} + \left[\frac{1}{\epsilon} \right]^{\frac{\alpha}{2}} \frac{\varkappa^{\alpha}}{\varkappa^{\frac{\alpha}{2}} - 1} \cdot \max \left\{ \Phi(x_0) - \Phi_*, \left[2L_1(\varphi) \varkappa \left(\frac{p+1}{\sigma_{p+1}} \right)^{\frac{2}{p+1}} \right]^{\frac{1}{\alpha}} \right\}^{\frac{\alpha}{2}}. \tag{63}$$

The optimal value of parameter $\varkappa$ is defined by the equation

$$\varkappa^{\alpha} = 4. \tag{64}$$

The optimal value (64) can be obtained by minimizing the factor for the maximum in the estimate (63). Thus, in view of the rules of the process (59), the choice

$$N_t = 2^t, \quad t \geq 0, \tag{65}$$

is optimal. Indeed, since

$$\sqrt{1 + 4^t} - 1 < 2^t < \sqrt{1 + 4^t}, \quad t \geq 0,$$

then $\lceil \sqrt{1 + 4^t} - 1 \rceil = 2^t$. The choice (65) corresponds to $\varkappa^{\frac{\alpha}{2}} = 2$, and the estimate for the total number of internal steps can be written now in the following form:

$$M_K \leq 4 \left[\frac{1}{\epsilon} \right]^{\frac{\alpha}{2}} \max \left\{ \Phi(x_0) - \Phi_*, 2 \left[2L_1(\varphi) \left(\frac{p+1}{\sigma_{p+1}} \right)^{\frac{2}{p+1}} \right]^{\frac{1}{\alpha}} \right\}^{\frac{\alpha}{2}}. \tag{66}$$

Let us point out the consequences of the estimate (63) for our main problem of interest, the problem (42) with $\Phi(y) = F_{\bar{x},2,H}(y)$ (see (41)). As we have already seen, in this case

$$p = 2, \quad \alpha = \frac{1}{3}, \quad \sigma_3 = \frac{H}{4}.$$

Besides that, we have one more parameter in the complexity bound (63):

$$L_1(\varphi) = \lambda_{\max}(\nabla^2 f(\bar{x})).$$

With respect to these parameters, the bound (63) on the *analytical complexity* of problem (42) is as follows:

$$\begin{aligned} M_K &\le 4\left[\frac{1}{\epsilon}\right]^{\frac{1}{6}} \max\left\{\Phi(\bar{x}) - \Phi_*, 2\left[2L_1(\varphi)\left(\frac{12}{H}\right)^{\frac{2}{3}}\right]^3\right\}^{\frac{1}{6}} \\ &= 4\left[\frac{1}{\epsilon}\right]^{\frac{1}{6}} \max\left\{\Phi(\bar{x}) - \Phi_*, \frac{9 \cdot 2^8 L_1^3(\varphi)}{H^2}\right\}^{\frac{1}{6}}. \end{aligned} \tag{67}$$

Choosing now $\epsilon = \Phi(x_K) - \Phi_*$, we get the following relation:

$$\Phi(x_K) - \Phi_* \le \left(\frac{4}{M_K}\right)^6 \max\left\{\Phi(\bar{x}) - \Phi_*, \frac{9 \cdot 2^8 L_1^3(\varphi)}{H^2}\right\}. \tag{68}$$

To the best of our knowledge, this is the fastest sublinear rate of convergence known so far in Convex Optimization. In certain situations, it looks even more attractive than the linear rate. Recall that this complexity result corresponds to a *first-order scheme* (59) as applied to the problem (42).

5. Flexible strategy for computing Newton step

Despite to the attractive complexity bound (67), the computational strategy (59) is very rigid and cannot adjust to the favourable properties of a particular optimization problem. In this section, we present a more flexible upper-level process, which is based on some variant of the Fast Gradient Method (50). It is applied to the auxiliary minimization problem in the following form:

$$\min_{x \in \operatorname{dom}\psi} \left\{\Phi(x) \stackrel{\text{def}}{=} \varphi(x) + \psi(x)\right\}, \quad \psi(x) \stackrel{\text{def}}{=} \frac{H}{6}\|x - \bar{x}\|^3, \tag{69}$$

where $\varphi(\cdot)$ is a convex function with Lipschitz continuous gradient. The constant $H > 0$ is supposed to be known.

Method $\mathcal{B}(\hat{x}, H, L)$

Define $\Omega_0(x) = \frac{1}{2}\|x - \hat{x}\|^2$. Set $A_0 = 0$, $x_0 = \hat{x}$, $v_0 = \hat{x}$, and $L_0 = L$.

k**th iteration** ($k \geq 0$)**.**

1. Find the minimal $i_k \geq 0$, defining the following objects:
- bound $H_k = 2^{i_k} L_k$,
- root $a_{k+1} > 0$ of the equation $\frac{(a_{k+1})^2}{A_k + a_{k+1}} = \frac{1}{H_k}$,
- parameters $A_{k+1} = A_k + a_{k+1}$ and $\tau_k = \frac{a_{k+1}}{A_{k+1}}$,
- points $y_k = (1 - \tau_k)x_k + \tau_k v_k$ and $x_{k+1} = (1 - \tau_k)x_k + \tau_k v_{k+1}$
  with $v_{k+1} = \arg\min_{x \in \mathbb{E}} \left\{\Omega_k(x) + a_{k+1}[\varphi(y_k) + \langle \nabla\varphi(y_k), x - y_k\rangle + \psi(x)]\right\}$,

which ensure $\varphi(x_{k+1}) \leq \varphi(y_k) + \langle \nabla\varphi(y_k), x - y_k\rangle + \frac{H_k}{2}\|x_{k+1} - x_k\|^2$. (70)

2. Update $L_{k+1} = \frac{1}{2}H_k$ and

$$\Omega_{k+1}(x) = \Omega_k(x) + a_{k+1}[\varphi(y_k) + \langle \nabla\varphi(y_k), x - y_k\rangle + \psi(x)].$$

Stopping criterion: $F(\hat{x}) - F(x_{k'}) \geq \frac{1}{A_{k'}^3}\left(\frac{12}{H}\right)^2$ for some $k' \geq 1$.

Output: Point $\mathcal{B}_{\hat{x},H,L} \stackrel{\text{def}}{=} x_{k'}$, constant $L_{\hat{x},H,L} = L_{k'}$, and linear function
$\ell_{\hat{x},H,L}(x) = \frac{1}{A_{k'}} \sum_{i=0}^{k'-1} a_{i+1}[\varphi(y_i) + \langle \nabla\varphi(y_i), x - y_i\rangle]$.

We are going to use this scheme at the lower level of the following procedure.

Input: Point $u_0 \in \mathbb{E}$ and parameters $\hat{L} \leq L_1(\varphi)$, $R \geq \|u_0 - x^*\|$.

Set $L_0 = \hat{L}$. and $r_0 = R$.

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

1. Compute $u_{t+1} = \mathcal{B}_{u_t,L_t,H}$.

2. Compute $\ell^*_{t+1} = \min_{x \in \mathbb{E}}\{\ell_{u_t,L_t,H}(x) + \psi(x) : \|x - u_t\| \leq r_t\}$.

3. Check stopping criterion $\delta_{t+1} \stackrel{\text{def}}{=} F(u_{t+1}) - \ell^*_{t+1} \leq \delta$.

4. If it is not satisfied, set $r_{t+1} = \left(\frac{12}{H}\delta_{t+1}\right)^{\frac{1}{3}}$ and $L_{t+1} = L_{u_t,L_t,H}$. (71)

The rate of convergence of this procedure is given by the following statement.

Lemma 5.1: *For all $t \geq 0$ we have*

$$\|u_t - x^*\| \leq r_t. \tag{72}$$

Therefore $F(x^*) \geq \ell_t^*$, $t \geq 1$. Moreover, for all $t \geq 0$ we have

$$\delta_t \leq \left(\frac{1}{2}\right)^t \left(\frac{H}{12}\right)^{\frac{2}{3}} (F(u_0) - F_*)^{\frac{1}{3}} R^2, \tag{73}$$

and

$$F(u_{t+1}) - F_* \leq \frac{1}{2}(F(u_t) - F_*), \quad t \geq 0. \tag{74}$$

Proof: For $t = 0$, inequality (72) is valid in view of the initial choice of the parameter R. Assume it is valid for some $t \geq 0$. Since

$$\ell_{u_t,L_t,H}(x) \leq \varphi(x), \quad x \in \mathbb{E},$$

we have $F(x^*) \geq \ell^*_{u_t,L_t,H}$. Therefore, since function $\psi(\cdot)$ is uniformly convex, we have

$$\frac{H}{12}\|u_{t+1} - x^*\|^3 \overset{(45)}{\leq} F(u_{t+1}) - F(x^*) \leq \delta_{t+1} = F(u_{t+1}) - \ell^*_{u_t,L_t,H} = \frac{H}{12} r_{t+1}^3.$$

Thus, inequality (72) is valid for all $t \geq 0$.

Note that using the same arguments as in Lemma 4.2, it is possible to prove that at the last step $k'(t)$ of the procedure (70) used at t-th iteration of the method (71), we have

$$\begin{aligned} F(u_{t+1}) &\leq \min_{x \in \mathbb{E}} \left\{ \frac{1}{2A_{k'(t)}} \|x - u_t\|^2 + \ell_{u_t,L_t,H}(x) + \psi(x) \right\} \\ &\leq \min_{\|x - u_t\| \leq r_t} \left\{ \frac{1}{2A_{k'(t)}} \|x - u_t\|^2 + \ell_{u_t,L_t,H}(x) + \psi(x) \right\} \\ &\leq \frac{r_t^2}{2A_{k'(t)}} + \ell^*_{u_t,L_t,H}. \end{aligned}$$

For $t \geq 1$ this means that

$$\delta_{t+1} \leq \frac{r_t^2}{2A_{k'(t)}} = \left(\frac{12}{H}\delta_t\right)^{\frac{2}{3}} \frac{1}{2A_{k'(t)}} \leq \left(\frac{12}{H}\delta_t\right)^{\frac{2}{3}} \cdot \frac{1}{2} \left(\frac{H}{12}\right)^{\frac{2}{3}} (F(u_t) - F(u_{t+1}))^{\frac{1}{3}},$$

where the last inequality is just the termination criterion of method (70). Thus, for $t \geq 1$, we come to the following inequality:

$$\delta_{t+1} \leq \frac{1}{2}\delta_t^{\frac{2}{3}} (F(u_t) - F_*)^{\frac{1}{3}} \leq \frac{1}{2}\delta_t.$$

For $t = 0$, by the stopping criterion in (70), we have

$$\delta_1 \leq \frac{R^2}{2A_{k'(0)}} \leq \frac{R^2}{2} \left(\frac{H}{12}\right)^{\frac{2}{3}} (F(u_0) - F_*)^{\frac{1}{3}}.$$

Hence, inequality (73) is proved for all $t \geq 0$.

Similarly, for the function values with $t \geq 0$, we have

$$\begin{aligned} F(u_{t+1}) - F_* &\leq \frac{\|u_t - x^*\|^2}{2A_{k'(t)}} \overset{(45)}{\leq} \frac{1}{2A_{k'(t)}} \left(\frac{12}{H}(F(u_t) - F_*)\right)^{\frac{2}{3}} \\ &\leq \frac{1}{2}\left(\frac{12}{H}(F(u_t) - F_*)\right)^{\frac{2}{3}} \left(\frac{H}{12}\right)^{\frac{2}{3}} (F(u_t) - F(u_{t+1}))^{\frac{1}{3}} \\ &\leq \frac{1}{2}(F(u_t) - F_*). \end{aligned}$$

And this is inequality (74). ■

Let us estimate now the total amount of iterations of the lower-level procedure (70) in the method (71). For this, we need the following lemma.

Lemma 5.2: *Any iteration of the method* (70) *satisfying the inequality*

$$F(\hat{x}) - F_* \geq \frac{2}{A_k^3}\left(\frac{12}{H}\right)^2 \tag{75}$$

satisfies also the stopping criterion of this scheme.

Proof: Indeed, if inequality (75) is valid, then

$$\begin{aligned} F(x_k) - F_* &\leq \frac{\|\hat{x} - x^*\|^2}{2A_k} \overset{(45)}{\leq} \frac{1}{2A_k}\left(\frac{12}{H}(F(\hat{x}) - F_*)\right)^{\frac{2}{3}} \\ &\overset{(75)}{\leq} \frac{1}{2}\left(\frac{12}{H}(F(\hat{x}) - F_*)\right)^{\frac{2}{3}} \left(\frac{H}{12}\right)^{\frac{2}{3}} \left(\frac{1}{2}(F(\hat{x}) - F_*)\right)^{\frac{1}{3}} \\ &= \left(\frac{1}{2}\right)^{\frac{4}{3}} (F(\hat{x}) - F_*). \end{aligned}$$

Therefore,

$$F(\hat{x}) - F(x_k) \geq \left(1 - \left(\frac{1}{2}\right)^{\frac{4}{3}}\right)(F(\hat{x}) - F_*) \overset{(75)}{\geq} \left(2 - \left(\frac{1}{2}\right)^{\frac{1}{3}}\right) \cdot \frac{1}{A_k^3}\left(\frac{12}{H}\right)^2 > \frac{1}{A_k^3}\left(\frac{12}{H}\right)^2.$$

And this is the stopping criterion in the method (70). ■

Now we can bound from above the number of iterations $k'(t)$. Indeed, in the method (70) we can guarantee that

$$L_k \leq L_1(\varphi), \quad H_k \leq 2L_1(\varphi), \quad k \geq 0.$$

Therefore, in view of Lemma 4.3, we can guarantee the following growth of the coefficients A_k:

$$A_k \geq \frac{k(k+2)}{8L_1(\varphi)}, \quad k \geq 0.$$

In accordance to the result of Lemma 5.2, this gives us the following bound on the number of steps of method (70):

$$k'(t) \le C_1 \left(\frac{2}{F(u_t) - F_*} \right)^{\frac{1}{6}}, \quad t \ge 0, \tag{76}$$

where $C_1 = \sqrt{8L_1(\varphi)}(\frac{12}{H})^{\frac{1}{3}}$. Let us give an upper bound for the number of the lower-level iterations, which are necessary for solving problem (69) up to accuracy $\delta > 0$.

Lemma 5.3: *Let for some $N \ge 1$ we have $F(u_N) - F_* \ge \delta$. Then*

$$M_N \stackrel{\text{def}}{=} \sum_{t=0}^{N} k'(t) \le \frac{2^{1/3} C_1}{2^{1/6} - 1} \left[\frac{1}{\delta} \right]^{\frac{1}{6}}.$$

Proof: Indeed, in view of inequality (74), we have

$$F(u_t) - F_* \ge 2^{N-t}\delta, \quad 0 \le t \le N.$$

Therefore,

$$M_N \stackrel{(76)}{\le} \sum_{t=0}^{N} C_1 \left(\frac{2}{F(u_t) - F_*} \right)^{\frac{1}{6}} \le C_1 \left[\frac{1}{\delta} \right]^{\frac{1}{6}} \sum_{t=0}^{N} \left(\frac{1}{2} \right)^{\frac{N-t-1}{6}} \le C_1 \left[\frac{1}{\delta} \right]^{\frac{1}{6}} \cdot \frac{2^{\frac{1}{6}}}{1 - \left(\frac{1}{2}\right)^{\frac{1}{6}}}.$$

■

Thus, our scheme has the same worst-case complexity bound as (59). However, it is much more flexible. This will be confirmed by preliminary computational experiments presented in the next section.

6. Preliminary computational experiments

In this section we discuss the results of the numerical experiments with Inexact Cubic Newton Method, which uses at each iteration an approximate solution of the problem (69) with $\varphi(x) = f_{\bar{x},2}(x)$ computed by the method (71). Our objective functions have the following structure:

$$f_\mu(x) = \mu \ln \left(\sum_{i=1}^{m} e^{(\langle a_i, x \rangle - b_i)/\mu} \right),$$

where μ is a positive parameter. For minimization problem

$$\min_{x \in \mathbb{R}^n} f_\mu(x) \tag{77}$$

we choose a starting point $x_0$close enough to the solution x^* of our problem:

$$\|x_0 - x^*\| = 1.$$

Coefficients of the vectors $a_i \in \mathbb{R}^n$ are randomly generated with uniform density in interval $[-1, 1]$.

Table 1. $\epsilon = 10^{-3}, \mu = 0.05$.

Dim	Iteration	NumFunc	FGM-Total	FGM-Average	Time (s)
100	11	19	1200	63.2	0.38
200	16	27	2780	102.9	2.92
500	20	35	2468	70.5	14.75
1000	19	34	2838	83.5	66.8

Table 2. $\epsilon = 10^{-4}, \mu = 0.05$.

Dim	Iteration	NumFunc	FGM-Total	FGM-Average	Time (s)
100	14	22	2743	124.7	0.83
200	22	34	8257	242.9	8.50
500	24	41	7337	178.9	43.39
1000	26	46	9429	204.9	220.91

Table 3. $\epsilon = 10^{-5}, \mu = 0.05$.

Dim	Iteration	NumFunc	FGM-Total	FGM-Average	Time (s)
100	17	25	6994	279.7	2.03
200	30	48	30450	634.4	31.05
500	30	49	19010	387.9	112.06
1000	31	50	22036	440.7	515.44

In our experiments, the accuracy parameter δ for the auxiliary minimization problem (69) was chosen in accordance to the theoretical recommendation (15) with $p = 2$, where the exact values of Lipschitz constant and the distance were replaced by some estimates, obtained during the minimization process. This was the only heuristic rule in our tests. All other elements were implemented exactly as they are described in the methods (14), (70), and (71).

Let us look at the results of our testing for different values of desired accuracy and dimension. In the first column we put dimension of the problem. The second column shows the number of iterations of the Inexact Cubic Newton Method. Next column displays the number of calls of oracle in the main minimization problem (77). Fourth column shows the total number of iterations of the method (70) and next column shows the average number of iterations for computing one Newton step. The last column show the total computational time.

As we can see from these tables, the Inexact Cubic Newton Method confirms its reputation of a fast method. The number of iterations and the number of calls of oracle is always small. The growth of the average number of the gradient steps is a little bit faster that the theoretical predictions. This gives us a good motivation to continue the research on the efficient termination criterions for the auxiliary processes.

Acknowledgments

The author is very thankful to three anonymous reviewers for providing comments significantly improved the presentation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This paper was prepared in the framework of Advanced Grant 788368 of European Research Council.

Notes on contributor

Yurii Nesterov Born: 1956, Moscow. Master degree 1977, Moscow State University. Doctor degree 1984. Professor at Center for Operations Research and Econometrics, UCLouvain, Belgium. Author of 5 monographs and more than 100 refereed papers in the leading optimization journals. Dantzig Prize, John von Neumann Theory Prize 2009, Charles Broyden prize 2010, Francqui Chair (Liege University 2011–2012), SIAM Outstanding paper award (2014), EURO Gold Medal 2016. Main direction is the development of efficient numerical methods for convex and nonconvex optimization problems supported by the global complexity analysis: general interior-point methods (theory of self-concordant functions), fast gradient methods (smoothing technique), global complexity analysis of second-order and tensor schemes (cubic regularization of the Newton's method).

ORCID

Yurii Nesterov http://orcid.org/0000-0002-0542-8757

References

[1] H.H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: first order methods revisited and applications*, Math. Oper. Res. 42 (2016), pp. 330–348.
[2] E.G. Birgin, J.L. Gardenghi, J.M. Martinez, S.A. Santos, and P. L. Toint, *Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*, Math. Program. 163 (2017), pp. 359–368.
[3] C. Cartis, N.I.M. Gould, and P.L. Toint, *Universal regularized methods – varying the power, the smoothness, and the accuracy*, SIAM. J. Optim. 29 (2019), pp. 595–615.
[4] G. Grapiglia and Y. Nesterov, *On inexact solution of auxiliary problems in tensor methods*, (2019). Available at arXiv 1907.13023 [math. OC].
[5] G.N. Grapiglia and Y. Nesterov, *Tensor methods for minimizing functions with Hölder continuous higher-order derivatives*, (2019). Available at arXiv:1904.12559 [math. OC].
[6] G.N. Grapiglia and Y. Nesterov, *Tensor methods for finding approximate stationary points of convex functions*, (2019). Available at arXiv: 1907.07053 [math. OC].
[7] A. Griewank and A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, *Other Titles in Applied Mathematics, Vol. 105*, 2nd ed., SIAM, 2008.
[8] H. Lu, R.M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM. J. Optim. 28 (2018), pp. 333–354.
[9] J.M. Martinez, *On high-order model regularization for constrained optimization*, SIAM. J. Optim. 27(4) (2017), pp. 2447–2458.
[10] Y. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program. 112 (2008), pp. 159–181.
[11] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program. 140(1) (2013), pp. 125–161.
[12] Y. Nesterov, *Implementable tensor methods in convex optimization*, CORE Discussion Paper 2018/05, 2018.
[13] Y. Nesterov, *Lecture notes on Convex Optimization*, Springer, Switzerland, 2018.

[14] Y. Nesterov and B.T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program. 108 (2006), pp. 177–205.
[15] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, *Computer Science and Applied Mathematics*, Academic Press, New York/London, 1970.
[16] C. Song and Y. Ma, *Towards Unified Acceleration of High-order Algorithms under Holder Continuity and Uniform Convexity*, (June 3, 2019). Available at arXiv 2714782.

Appendix

Lemma A.1: *Let the sequence of positive numbers* $\{\xi_t\}_{t\geq 0}$ *satisfies the following condition:*

$$\xi_t - \xi_{t+1} \geq \xi_{t+1}^{1+\alpha}, \quad t \geq 0, \tag{A1}$$

where $\alpha \in (0, 1]$. *Then for any* $t \geq 0$ *we have*

$$\xi_t \leq \frac{\xi_0}{\left(1 + \frac{\alpha t}{1+\alpha} \ln(1 + \xi_0^\alpha)\right)^{1/\alpha}} \leq \left[\left(1 + \frac{1}{\alpha}\right)(1 + \xi_0^\alpha) \cdot \frac{1}{t}\right]^{\frac{1}{\alpha}}. \tag{A2}$$

Proof: We are going to prove inequality

$$\xi_t \leq \frac{\xi_0}{(1 + at)^{1/\alpha}}, \quad t \geq 0, \tag{A3}$$

with certain $a > 0$. Clearly, this inequality is valid for $t = 0$. Assume that it is valid for some $t \geq 0$, but it is not valid for the next value. Then

$$\frac{\xi_0}{(1 + at)^{1/\alpha}} \overset{(A1)}{>} \frac{\xi_0}{(1 + a(t + 1))^{1/\alpha}} \left(1 + \frac{\xi_0^\alpha}{1 + a(t + 1)}\right).$$

Let us prove that, for a certain choice of a, this is impossible. Thus, we need to justify the following inequality:

$$\frac{\xi_0}{(1 + a(t + 1))^{1/\alpha}} \left(1 + \frac{\xi_0^\alpha}{1 + a(t + 1)}\right) \geq \frac{\xi_0}{(1 + at)^{1/\alpha}}.$$

It can be rewritten as follows:

$$1 + \frac{\xi_0^\alpha}{1 + a(t + 1)} \geq \left(1 + \frac{a}{1 + at}\right)^{1/\alpha}.$$

Denoting $\tau = \frac{1}{1+at}$, we get inequality

$$1 + \frac{\xi_0^\alpha}{\frac{1}{\tau} + a} = 1 + \frac{\tau \xi_0^\alpha}{1 + a\tau} \geq (1 + a\tau)^{1/\alpha}, \tag{A4}$$

which we want to ensure for all $\tau \in [0, 1]$. Since $\alpha \in (0, 1]$, the right-hand side of this inequality is convex in τ, and its left-hand side is concave in τ. For $\tau = 0$, inequality (A4) is trivial. So, we need to check only the case $\tau = 1$:

$$1 + \frac{\xi_0^\alpha}{1 + a} \geq (1 + a)^{1/\alpha}.$$

This is

$$1 + a + \xi_0^\alpha \geq (1 + a)^{\frac{1+\alpha}{\alpha}}.$$

Hence, it is sufficient to ensure

$$1 + a + \xi_0^\alpha \geq e^{\frac{1+\alpha}{\alpha} a}.$$

For this, it is enough to choose $a = \frac{\alpha}{1+\alpha}\ln(1+\xi_0^\alpha)$. Thus, we come to a contradiction, which proves the first part of inequality (A2). Its second part follows from the following relation:

$$\ln(1+\xi_0^\alpha) = -\ln\left(1 - \frac{\xi_0^\alpha}{1+\xi_0^\alpha}\right) \geq \frac{\xi_0^\alpha}{1+\xi_0^\alpha}.$$

■

Remark A.1: Note that the middle part of inequality (A2) has a correct limit as $\alpha \downarrow 0$. Indeed,

$$\lim_{\alpha\downarrow 0}\left(1 + \frac{\alpha t}{1+\alpha}\xi_0^\alpha \ln(1+\xi_0^\alpha)\right)^{1/\alpha} = \exp\left[\lim_{\alpha\downarrow 0}\frac{1}{\alpha}\ln\left(1 + \frac{\alpha t}{1+\alpha}\xi_0^\alpha \ln(1+\xi_0^\alpha)\right)\right] = 2^t.$$

This is compatible with the variant of inequality (A1) for $\alpha = 0$, which is $\xi_t \geq 2\xi_{t+1}$. □