Inexact tensor methods and their application to stochastic convex optimization

Artem Agafonov\textsuperscript{a,b}, Dmitry Kamzolov \textsuperscript{a,a,b}, Pavel Dvurechensky \textsuperscript{c}, Alexander Gasnikov\textsuperscript{a,d,e} and Martin Takáč \textsuperscript{b}

\textsuperscript{a}Moscow Institute of Physics and Technology, Dolgoprudny, Russia; \textsuperscript{b}Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE; \textsuperscript{c}Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany; \textsuperscript{d}Institute for Information Transmission Problems, Moscow, Russia; \textsuperscript{e}HSE University, Moscow, Russia

\textbf{ABSTRACT}

We propose general non-accelerated \cite{Agafonov2020} and accelerated tensor methods under inexact information on the derivatives of the objective, analyze their convergence rate. Further, we provide conditions for the inexactness in each derivative that is sufficient for each algorithm to achieve a desired accuracy. As a corollary, we propose stochastic tensor methods for convex optimization and obtain sufficient mini-batch sizes for each derivative.

\textbf{ARTICLE HISTORY}

Received 15 March 2022
Accepted 18 September 2023

\textbf{KEYWORDS}

High-order methods; tensor methods; inexact derivatives; convex optimization; stochastic optimization

1. Introduction

The idea of using the derivatives of the order \(p\) higher than two in optimization methods is known at least since 1970s \cite{Chebyshev1852}. In numerical analysis it was used much earlier, see the student work of P.L. Chebyshev \cite{Chebyshev1852} and more recent reviews \cite{Kamzolov2020, Dvurechensky2020}. Despite theoretical advantages, the practical application of such tensor methods was limited until the recent work \cite{Agafonov2022} since each iteration of such methods includes minimization of a polynomial with degree larger than 3, which may be non-convex even for convex optimization problems. As it was shown in \cite{Agafonov2022}, if the minimization problem is convex, then in each iteration of the tensor method it is sufficient to minimize a convex polynomial, and, for \(p = 3\), this can be done with approximately the same cost as the step of the cubic-regularized Newton’s method in the convex case \cite{Nesterov2006}. This leads to a method with faster convergence rate than that of the accelerated cubic-regularized Newton’s method, and with approximately the same cost of each iteration. These ideas were further developed to obtain accelerated versions of tensor methods \cite{Kamzolov2020, Dvurechensky2020} and very fast second-order methods \cite{Gasnikov2020, Gasnikov2020a, Gasnikov2020b}.

At the same time, many optimization problems in machine learning or image analysis have the form of minimization of a finite sum of functions and are solved by second-order methods \cite{Beck2017, Nesterov2006}. A more general setting, for which the finite-sum setting is a special
case, is the setting of general stochastic optimization, for which second-order methods are also developed in the literature [53] for non-convex problems. Tensor methods for such problems are also developed, but with the focus on non-convex problems [7,39]. Thus, motivated by the lack of results on tensor methods for stochastic convex optimization, in this paper, we study stochastic convex optimization problems and develop tensor methods for this setting. We do this by a more general framework of Inexact Tensor Methods, which use inexact values of higher-order derivatives. First, we analyse such methods under a particular condition on the inexactness in the derivatives, and, then, we show how to satisfy this condition in the setting of stochastic optimization and propose a stochastic tensor methods for convex optimization.

1.1. Problem statement

We consider convex optimization problems of the following form

\[ \min_{x \in \mathbb{R}^n} f(x) \]  

under inexact information on the derivatives of the objective \( f \) up to order \( p \geq 2 \). We are motivated by two particular cases of this problem in the stochastic optimization setting, which we refer to as online and offline settings. In the online setting we assume that \( f(x) \) is given as

\[ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x; \xi)], \]  

where the random variable \( \xi \) is sampled from a distribution \( \mathcal{D} \), and an optimization procedure has access only to stochastic realizations of the function \( f(x; \xi) \) via its higher-order derivatives up to the order \( p \geq 2 \). In the offline setting the function \( f \) has finite-sum form:

\[ f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x) \]  

with available derivatives of the order \( p \geq 2 \) for each function \( f_i \). Clearly, the offline setting can be considered as a particular case of the online setting if we set \( \xi \) to be uniformly distributed over \( i = 1, \ldots, m \). At the same time, we distinguish these two settings since the analysis of the proposed methods in the second setting can be made under weaker assumptions.

1.2. Related work

1.2.1. Second-order methods

Beyond first-order methods, the most developed methods are maybe second-order methods, among which the closest to our setting are cubic-regularized Newton’s methods originating from [47]. This line of works includes the development of accelerated methods [41–43], extensions of trust region methods [10–13,17], methods with inexact evaluations of gradients and/or Hessians [6,20,31,54] with application to stochastic optimization in the online and offline settings. We especially point to [31] which largely inspired our work. Stochastic second-order methods for convex optimization [32,37,50] and non-convex optimization [48,53,55,58] have been extensively studied in the recent literature. The difference with our setting is that these works consider a particular case of \( p = 2 \).
1.2.2. **Tensor methods**

To distinguish the general methods that use the derivatives up to the order \( p > 2 \), we refer to them as tensor methods. The idea of such methods was proposed quite long ago \([33]\), and accelerated tensor methods for convex optimization were proposed in \([4]\). Recent interest to this type of methods in convex optimization was probably motivated by the lower bounds obtained in \([2,3]\). In \([44]\), it was shown that appropriately regularized Taylor expansion of a convex function is also convex, which leads to implementability of such methods that minimize this regularized Taylor expansion. Accelerated tensor methods were also proposed in \([44]\), yet with a remaining gap between the upper and lower complexity bounds. In the same paper, the author also shows how tensor methods with \( p = 3 \) can be implemented by solving the auxiliary problem with the Bregman projected gradient method in the relative smoothness setting. Near-optimal, i.e. with optimal up to a logarithmic factor iteration complexity, tensor methods for convex optimization were recently proposed in a number of works \([9,24,26–30,34,45]\). These developments allowed to propose faster second-order methods via implementing third-order methods with inexact third derivative \([35,45,46]\), which lead to an improvement of the complexity bound from \( O(\varepsilon^{-1/3.5}) \) to \( O(\varepsilon^{-1/5}) \). Stochastic tensor methods were developed for non-convex smooth enough problems in \([7,39]\). We are not aware of an analysis of tensor methods for stochastic convex smooth enough problems.

Our analysis is based on inexact versions of tensor methods for convex optimization, which use inexact derivatives of the order up to \( p \geq 2 \), mainly inspired by \([44]\). First-order methods with inexact gradients are well-developed in the literature, see, for example, \([16,18,19,21,22,25,49,51]\) and references therein. Some results on inexact second-order methods are listed above. The paper \([45]\) proposes an analysis of third-order method for convex optimization with inexact third derivative. The general theory of inexact tensor methods of an arbitrary order \( p \) for convex problems still has to be developed and we make a step in this direction.

While there are numerous tensor methods available for the non-convex case under the assumption of inexact derivatives \([5,7,8,14,36,39,53,56]\), the same cannot be said for the convex case. To the best of our knowledge, we propose the first high-order methods with inexact derivatives for convex optimization problems. This is because there is a fundamental difference between tensor methods used for convex and non-convex problems.

The main challenge in constructing convex tensor methods is ensuring the convexity of the subproblem. This issue has been successfully addressed in papers by Yu. Nesterov \([44,46]\). In these works, it was shown that the subproblem’s convexity is essential for achieving an efficient solution. Moreover, for the convexity of the subproblem it is necessary to choose the regularization parameter \( \sigma \geq L_p \) \([44]\). Therefore, it is not possible to directly apply adaptive schemes, as in non-convex tensor methods. In the case of non-convex problems, the computational costs to solve the high-order subproblem are often neglected.

### 1.3. Our contribution

Motivated by stochastic optimization methods, we propose and analyse inexact tensor methods of the general order \( p \geq 2 \) for convex optimization problems. The idea of these algorithms is to use inexact values of the derivatives up to the order \( p \) and to construct
a regularized inexact Taylor expansion. More precisely, we assume that inexact derivative $G_{x,i}$ satisfies the following condition (see Condition 2.1 below):

$$\| G_{x,i} - \nabla^i f(x) \| y - x \| \leq \delta_i \| y - x \|^{i-1},$$

for all $y$ and some $\delta_i \geq 0$, and $i = 1, 2, \ldots, p$. Based on the $p$th order inexact Taylor polynomial we build Inexact Tensor Model and use it to construct Inexact Tensor Method and its accelerated version. For the Inexact Tensor Method, we obtain the following convergence rate result:

$$f(x_T) - f(x^*) \leq O \left( \sum_{i=1}^{p} \frac{\delta_i}{T^{i-1}} \max_{x \in L'(x_0)} \| x - x^* \|^i + \frac{L_p}{T^p} \max_{x \in L'(x_0)} \| x - x^* \|^{p+1} \right),$$

where $L'(x_0)$ is defined in (24), $L_p$ is the Lipschitz constant of the $p$th derivative, and $T$ is the iteration counter.

For the Accelerated Inexact Tensor Method we obtain the following convergence rate result:

$$f(x_T) - f(x^*) \leq O \left( \delta_1 \bar{R} + \sum_{i=2}^{p} \frac{\delta_i}{T^{i-1}} \| x^* - x_0 \|^i + \frac{L_p}{T^{p+1}} \| x^* - x_0 \|^{p+1} \right),$$

where $\bar{R}$ is defined in (62). Based on the above convergence rate estimates, we obtain conditions on the inexactnesses $\delta_i$, $i = 1, 2, \ldots, p$ of the derivatives that are sufficient for the proposed inexact methods to have the same convergence rate as their exact counterparts.

As a corollary, we propose stochastic non-accelerated and accelerated tensor methods for stochastic convex optimization problems in the online and offline settings. The idea of the algorithms is to sample in each iteration mini-batches of derivatives up to the order $p$ and use them to construct a regularized inexact Taylor expansion. Using Hoeffding Tensor concentration inequality, we estimate mini-batch sizes for each derivative that are sufficient for the whole method to achieve accuracy $\epsilon$ in the same number of iterations as the exact method requires. For the non-accelerated method these mini-batch sizes are $O(\epsilon^{-2(p+1-i)/p})$, $i = 1, 2, \ldots, p$ and for the accelerated method they are $O(\epsilon^{-2})$ for $i = 1$ and $O(\epsilon^{-2(p-1+i)/(p+1)})$, $i = 2, \ldots, p$. Interestingly, the higher is the derivative order, the smaller batch size turns out to be sufficient.

We also consider a particular case $p = 3$, for which, in the spirit of [44], we describe how to implement the resulting inexact tensor methods. The theoretical analysis is mainly inspired by inexact second-order methods from [31] and tensor methods from [44].

### 1.4. Paper organization

The remaining part of the paper is organized as follows. In Section 2, we introduce main notations and assumptions. Then, in Section 3 we present an inexact tensor model of the objective function, prove its convexity, and show that it majorizes the objective function. Section 4 is dedicated to the Inexact Tensor Method itself and its convergence rate analysis.

Next, we introduce and analyse the Accelerated Inexact Tensor Method in Section 5. In Section 6, we discuss some implementation details. Finally, in Section 7 we apply the general inexact methods to introduce Stochastic Tensor Methods.
2. Preliminaries

We denote the $i$th directional derivative of function $f$ at $x$ along directions $s_1, \ldots, s_i \in \mathbb{R}^n$ as

$$\nabla^i f(x)[s_1, \ldots, s_i].$$

For example, $\nabla f(x)[s_1] = \langle \nabla f(x), s_1 \rangle$ and $\nabla^2 f(x)[s_1, s_2] = \langle \nabla^2 f(x)s_1, s_2 \rangle$. If all directions are the same we write $\nabla^i f(x)[s]$. For a $p$th order tensor $T$, we denote by $\|T\|$ its tensor norm recursively induced by the Euclidean norm on the space of $p$th order tensors:

$$\|T\| = \max_{\|s_1\| = \ldots = \|s_p\| = 1} \{|T[s_1, \ldots, s_p]|\},$$

where we denote by $\|\cdot\|$ the standard Euclidean norm. Throughout the paper we make the following assumption.

**Assumption 2.1:** Function $f$ is convex, $p$ times differentiable on $\mathbb{R}^n$, and its $p$th derivative is Lipschitz continuous, i.e. for all $x, y \in \mathbb{R}^n$

$$\|\nabla^p f(x) - \nabla^p f(y)\| \leq L_p \|x - y\|.$$

Following the previous works, we construct tensor methods based on the $p$th order Taylor approximation of the function $f(x)$, which can be written as follows:

$$\Phi_{x,p}(y) \overset{\text{def}}{=} f(x) + \sum_{i=1}^{p} \frac{1}{i!} \nabla^i f(x)[y - x]^i, \quad y \in \mathbb{R}^n. \quad (4)$$

The full Taylor expansion of $f$ requires computing all the derivatives up to the order $p$, which can be expensive to calculate. Thus, it is natural to use some approximations $G_{x,i}$ for the derivatives $\nabla^i f(x)$, $i = 1, \ldots, p$ and to construct an inexact $p$th order Taylor expansion of the objective:

$$\phi_{x,p}(y) = f(x) + \sum_{i=1}^{p} \frac{1}{i!} G_{x,i}[y - x]^i, \quad (5)$$

where $G_{x,i}$ satisfies the following

**Condition 2.1:** Given the inexactness levels $\delta_i \geq 0$ for $i = 1, \ldots, p$, for all $x \in \mathbb{R}^n$ the approximate derivatives $G_{x,i}$ satisfy for all $y \in \mathbb{R}^n$ the following inequalities:

$$\|(G_{x,i} - \nabla^i f(x))[y - x]^{i-1}\| \leq \delta_i \|y - x\|^{i-1}, \quad i = 1, \ldots, p. \quad (6)$$

Below, we first analyse inexact tensor methods under different conditions on the accuracy of the approximations for the derivatives. Then, motivated by stochastic optimization problems (1), (2), we focus on stochastic approximations of the derivatives through sampling. More precisely, for $S_1, S_2, \ldots, S_p$ being sample sets, we construct the sampled approximations as

$$G_{x,i} = \frac{1}{|S_i|} \sum_{j \in S_i} \nabla^i f(x, \xi_j), \quad i = 1, \ldots, p. \quad (7)$$

In Section 7 we show how to choose the size of the sample sets $S_i$ so that our algorithms achieve a desired accuracy with high probability.
In our methods we use the following \( p \)-th order prox functions:

\[
d_p(x) = \frac{1}{p} \|x\|^p, \quad p \geq 2. \tag{8}
\]

Note that

\[
\nabla d_p(x) = \|x\|^{p-2}x,
\]

\[
\nabla^2 d_p(x) = (p-2)\|x\|^{p-4}xx^* + \|x\|^{p-2}I \geq \|x\|^{p-2}I,
\tag{9}
\]

where \( I \) is the identity matrix in \( \mathbb{R}^n \).

As it is shown, e.g. in [44], Assumption 2.1 allows to control the quality of the approximation of the objective \( f \) by its Taylor polynomial:

\[
|f(y) - \Phi_{x,p}(y)| \leq \frac{L_p}{(p+1)!} \|y-x\|^{p+1}, \quad x, y \in \mathbb{R}^n. \tag{10}
\]

If \( p \geq 2 \), the same can be done with the first and second derivatives:

\[
\|\nabla f(y) - \nabla \Phi_{x,p}(y)\| \leq \frac{L_p}{p!} \|y-x\|^{p}, \quad x, y \in \mathbb{R}^n, \tag{11}
\]

\[
\|\nabla^2 f(y) - \nabla^2 \Phi_{x,p}(y)\| \leq \frac{L_p}{(p-1)!} \|y-x\|^{p-1}, \quad x, y \in \mathbb{R}^n. \tag{12}
\]

### 3. Inexact tensor model

In this section, we analyse inexact Taylor approximation (5) under Condition 2.1. First, we obtain under Condition 2.1 approximation bounds in the spirit of (10), (11), (12). Then, based on these bounds, we construct a regularized inexact Taylor polynomial and show that it is a global upper bound for the objective function \( f \). The latter is the key to construct the proposed Inexact Tensor Method, which we analyse in the next section.

The following lemma gives a counterpart of (10) when the inexact derivatives are used and shows that we can bound the residual between function \( f \) and the \( p \)-th order inexact Taylor polynomial \( \phi_{x,p}(y) \) defined in (5).

**Lemma 3.1:** Assume that Condition 2.1 is satisfied. Then, for any \( x, y \in \mathbb{R}^n \), we have

\[
|f(y) - \phi_{x,p}(y)| \leq \sum_{i=1}^{p} \frac{\delta_i}{i!} \|y-x\|^i + \frac{L_p}{(p+1)!} \|y-x\|^{p+1}. \tag{13}
\]

**Proof:** For any \( x, y \in \mathbb{R}^n \):

\[
|f(y) - \phi_{x,p}(y)| \leq |f(y) - \Phi_{x,p}(y)| + |\Phi_{x,p}(y) - \phi_{x,p}(y)| \leq \frac{L_p}{(p+1)!} \|y-x\|^{p+1} + |\Phi_{x,p}(y) - \phi_{x,p}(y)|. \tag{10}
\]
Let us bound the second term in the right-hand side of the inequality above:

\[
|\Phi_{x,p}(y) - \phi_{x,p}(y)| \overset{(4),(5)}{=} \left| \sum_{i=1}^{p} \frac{1}{i!} (G_{x,i} - \nabla^i f(x))[y - x]^i \right| \\
\leq \sum_{i=1}^{p} \frac{1}{i!} \| (G_{x,i} - \nabla^i f(x))[y - x]^{i-1}\| \|y - x\|^i \overset{(6)}{=} \sum_{i=1}^{p} \frac{\delta_i}{i!} \|y - x\|^i.
\]

Combining both inequalities above, we finish the proof. ■

The following lemma gives a counterpart of (11), (12) for the inexact Taylor polynomial \( \phi_{x,p}(y) \) defined in (5).

**Lemma 3.2:** Assume that Condition 2.1 is satisfied. Then, for any \( x, y \in \mathbb{R}^n \), we have

\[
\|\nabla f(y) - \nabla \phi_{x,p}(y)\| \leq \sum_{i=1}^{p} \frac{\delta_i}{(i-1)!} \|y - x\|^{i-1} + \frac{L_p}{p!} \|y - x\|^p,
\]

(14)

\[
\|\nabla^2 f(y) - \nabla^2 \phi_{x,p}(y)\| \leq \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} \|y - x\|^{i-2} + \frac{L_p}{(p - 1)!} \|y - x\|^{p-1},
\]

(15)

where we use the standard convention \( 0! = 1 \).

**Proof:** First, we prove the bound for the first derivatives. For any \( x, y \in \mathbb{R}^n \):

\[
\|\nabla f(y) - \nabla \phi_{x,p}(y)\| \leq \|\nabla f(y) - \nabla \Phi_{x,p}(y)\| + \|\nabla \Phi_{x,p}(y) - \nabla \phi_{x,p}(y)\| \overset{(11)}{\leq} \frac{L_p}{p!} \|y - x\|^p + \|\nabla \Phi_{x,p}(y) - \nabla \phi_{x,p}(y)\|.
\]

Let us bound the second term in the right-hand side of the inequality above:

\[
\|\nabla \Phi_{x,p}(y) - \nabla \phi_{x,p}(y)\| \overset{(4),(5)}{=} \left| \sum_{i=1}^{p} \frac{1}{(i-1)!} (G_{x,i} - \nabla^i f(x))[y - x]^{i-1}\right| \\
\leq \sum_{i=1}^{p} \frac{1}{(i-1)!} \| (G_{x,i} - \nabla^i f(x))[y - x]^{i-1}\| \overset{(6)}{=} \sum_{i=1}^{p} \frac{\delta_i}{(i-1)!} \|y - x\|^{i-1}.
\]

Combining both inequalities above, we get (14).
Now, we obtain the bound for the second derivatives. For any \( x, y \in \mathbb{R}^n \), we have
\[
\| \nabla^2 f(y) - \nabla^2 \phi_{x,p}(y) \| \leq \| \nabla^2 f(y) - \nabla^2 \Phi_{x,p}(y) \| + \| \nabla^2 \Phi_{x,p}(y) - \nabla^2 \phi_{x,p}(y) \|
\]
(12) \( \frac{L_p}{(p-1)!} \| y - x \|^{p-1} + \| \nabla^2 \Phi_{x,p}(y) - \nabla^2 \phi_{x,p}(y) \| .
\]

Let us bound the second term in the right-hand side of the inequality above:
\[
\| \nabla^2 \Phi_{x,p}(y) - \nabla^2 \phi_{x,p}(y) \| \overset{(4),(5)}{=} \left\| \sum_{i=2}^{p} \frac{1}{(i-2)!} (G_{x,i} - \nabla^i f(x))[y - x]^{i-2} \right\|
\]
\[
\leq \sum_{i=2}^{p} \frac{1}{(i-2)!} \| (G_{x,i} - \nabla^i f(x))[y - x]^{i-2} \|
\]
\[
\leq \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} \| y - x \|^{i-2},
\]
where the last inequality is valid due to conditions (6) and the definition of the tensor norm. Indeed, for \( i = 2, \ldots, p \), we have
\[
\| (G_{x,i} - \nabla^i f(x))[y - x]^{i-2} \| = \max_{\| s_1 \| = \| s_2 \| = 1} |(G_{x,i} - \nabla^i f(x))[y - x]^{i-2}[s_1, s_2]|
\]
\[
= \max_{y \in \mathbb{R}^n, y \neq x} \frac{|(G_{x,i} - \nabla^i f(x))[y - x]^{i-2}[y - x, y - x]|}{\| y - x \|^2}
\]
\[
\leq \max_{y \in \mathbb{R}^n, y \neq x} \frac{\| (G_{x,i} - \nabla^i f(x))[y - x]^{i-1} \|}{\| y - x \|}
\]
(6) \( \leq \delta_i \| y - x \|^{i-2}
\]
Combining the above bounds, we obtain (15). \( \blacksquare \)

We finish this section by constructing a global upper bound for the objective \( f \) based on the inexact Taylor expansion \( \phi_{x,p} \) which is regularized by functions \( d_i, i = 2, \ldots, p \).

**Theorem 3.1:** Assume that Condition 2.1 is satisfied. Then, for any \( \gamma \in \mathbb{R}, x, y \in \mathbb{R}^n \), and \( \sigma \geq L_p \) the function
\[
\omega_{x,p}(y) = \phi_{x,p}(y) + \frac{\delta_1}{2\gamma} + \gamma \delta_1 d_2(y - x) + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} d_i(y - x) + \frac{\sigma}{(p-1)!} d_{p+1}(y - x)
\]
(16)
is convex and majorizes the function \( f \):
\[
f(y) \leq \omega_{x,p}(y), \quad x, y \in \mathbb{R}^n.
\]
(17)
Proof: For any \(x, y \in \mathbb{R}^n\) and \(h \in \mathbb{R}^n\):

\[
\langle (\nabla^2 f(y) - \nabla^2 \phi_{x,p}(y)) h, h \rangle \leq \|\nabla^2 f(y) - \nabla^2 \phi_{x,p}(y)\| \cdot \|h\|^2
\]

\[\overset{(15)}{=} \left( \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} \|y-x\|^{i-2} + \frac{L_p}{(p-1)!} \|y-x\|^{p-1} \right) \|h\|^2.
\]

Whence, using the convexity of \(f\), we have

\[
0 \leq \nabla^2 f(y) \leq \nabla^2 \phi_{x,p}(y) + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} \|y-x\|^{i-2} I + \frac{L_p}{(p-1)!} \|y-x\|^{p-1} I
\]

\[
\overset{(9)}{=} \nabla^2 \phi_{x,p}(y) + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} \nabla^2 d_i(y-x) I + \frac{\sigma}{(p-1)!} \nabla^2 d_{p+1}(y-x) I = \nabla^2 \tilde{\omega}_{x,p}(y),
\]

where we used that \(\sigma \geq L_p\) and defined

\[
\tilde{\omega}_{x,p}(y) = \phi_{x,p}(y) + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} d_i(y-x) + \frac{\sigma}{(p-1)!} d_{p+1}(y-x).
\]

Thus, we conclude that \(\tilde{\omega}_{x,p}(y)\) is convex. Further, \(\omega_{x,p}(y) = \tilde{\omega}_{x,p}(y) + \frac{\delta_1}{2\gamma} + \gamma \delta_1 d_2(y-x)\) is convex as a sum of three convex functions.

Finally, we prove inequality (17):

\[
f(y) \overset{(13)}{\leq} \phi_{x,p}(y) + \sum_{i=1}^{p} \frac{\delta_i}{i!} \|y-x\|^i + \frac{L_p}{(p+1)!} \|y-x\|^{p+1}
\]

\[= \phi_{x,p}(y) + \delta_1 \|y-x\| + \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} d_i(y-x) + \frac{L_p}{p!} d_{p+1}(y-x)
\]

\[\leq \phi_{x,p}(y) + \delta_1 \|y-x\| + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} d_i(y-x) + \frac{\sigma}{(p-1)!} d_{p+1}(y-x)
\]

\[\|x\| \leq \frac{1}{2\gamma} + \frac{\gamma}{2} \|x\|^2 \leq \phi_{x,p}(y) + \frac{\delta_1}{2\gamma} + \gamma \delta_1 d_2(y-x) + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} d_i(y-x) + \frac{\sigma}{(p-1)!} d_{p+1}(y-x)
\]

\[= \phi_{x,p}(y) + \delta_1 d_1(y-x) + \frac{\sigma}{(p-1)!} d_{p+1}(y-x) = \omega_{x,p}(y).
\]

Thus, we have constructed a regularized inexact Taylor polynomial \(\omega_{x,p}(y)\) defined in (16) as a model of the objective \(f(y)\). Theorem 3.1 claims that this model satisfies two main conditions:
• Model $\omega_{x,p}(y)$ is a global upper bound for the function $f$:

$$f(y) \leq \omega_{x,p}(y), \quad x, y \in \mathbb{R}^n. \quad (21)$$

• Model $\omega_{x,p}(y)$ is convex.

In the next section we use the model $\omega_{x,p}(y)$ to construct Inexact Tensor Method.

4. Inexact tensor method

Based on the inexact regularized Taylor polynomial $\omega_{x,p}(y)$ defined in the previous section, in this section, we present the Inexact Tensor Method (ITM). Each step of this algorithm uses minimization of our model $\omega_{x,p}(y)$ to make a step from a point $x$. To be more precise, we define an operator $S(x)$ as

$$S(x) = \arg\min_{y \in \mathbb{R}^n} \omega_{x,p}(y). \quad (22)$$

The resulting Inexact Tensor Method is listed as Algorithm 1.

**Algorithm 1 Inexact Tensor Method (ITM)**

1. **Input:** convex function $f$ such that $\nabla^p f$ is $L_p$-Lipschitz; $x_0$ is starting point; constant $\sigma \geq L_p$.
2. **for** $t \geq 0$ **do**
3. Call the inexact oracle to compute $G_{x_t,i}$ for $i = 1, \ldots, p$ such that Condition 2.1 is satisfied.
4. Make the step:

$$x_{t+1} = \arg\min_{y \in \mathbb{R}^n} \omega_{x_t,p}(y). \quad (23)$$
5. **end for**

In view of inequality (17), the process $S(x)$ guarantees that $f(S(x)) < f(x) + \frac{\delta_1}{2\gamma}$, i.e. Algorithm 1 is almost monotone, i.e. is monotone up to accuracy $\frac{\delta_1}{2\gamma}$. Indeed,

$$f(x_{t+1}) \overset{(17)}{\leq} \omega_{x_t,p}(x_{t+1}) \overset{(23)}{\leq} \omega_{x_t,p}(x_t) = f(x_t) + \frac{\delta_1}{2\gamma}.$$ 

Therefore, for any iteration $T \geq 1$, we have $f(x_T) \leq f(x_0) + \frac{T\delta_1}{2\gamma}$. We define

$$\mathcal{L}'(x_0) = \left\{ x \left| f(x) \leq f(x_0) + \frac{T\delta_1}{2\gamma} \right. \right\}.$$ 

**Remark 4.1:** The approximate monotonicity is caused by our construction of the model $\omega_{x,p}(y)$ which uses in (20) the inequality $\|y - x\| \leq \frac{1}{2\gamma} + \frac{\gamma}{2}\|y - x\|^2$ which holds for any
\( \gamma > 0 \). This is needed to obtain a smooth model which is easier to minimize. Otherwise, we can construct a non-smooth model

\[
\bar{\omega}_{x,p}(y) = \phi_{x,p}(y) + \delta_1 \|y - x\| + \sum_{i=1}^{p} \frac{\delta_i}{(i-2)!} d_i(y - x) + \frac{\sigma}{(p-1)!} d_{p+1}(y - x)
\]

and minimize it in each iteration of the method. Such constructed modification of the ITM becomes monotone. Indeed,

\[
f(x_{t+1}) \leq \bar{\omega}_{x_t,p}(x_{t+1}) \leq \bar{\omega}_{x_t,p}(x_t) = f(x_t).
\]

The next technical result is used to prove the convergence rate theorem for ITM.

**Lemma 4.1:** Assume that Condition 2.1 is satisfied. Then, for any \( \gamma \in \mathbb{R} \), \( x \in \mathbb{R}^n \), and \( \sigma \geq L_p \), we have

\[
f(S(x)) = \min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{\delta_1}{2\gamma} + 2\gamma \delta_1 d_2(y - x) + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} d_i(y - x) + \frac{L_p + p\sigma}{p!} d_{p+1}(y - x) \right\}. \tag{25}
\]

**Proof:**

\[
f(S(x)) \overset{(21)}{=} \omega_{x,p}(S(x)) \overset{(22),(16)}{=} \min_{y \in \mathbb{R}^n} \left\{ \phi_{x,p}(y) + \frac{\delta_1}{2\gamma} + \gamma \delta_1 d_2(y - x) + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} d_i(y - x) + \frac{L_p + p\sigma}{p!} d_{p+1}(y - x) \right\}
\]

\[
\overset{(13)}{\leq} \min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{\delta_1}{2\gamma} + \gamma \delta_1 d_2(y - x) + \frac{L_p + p\sigma}{p!} d_{p+1}(y - x) \right\}
\]

We are now in a position to estimate the convergence rate of the Inexact Tensor Method.
Theorem 4.1: Let $f$ be convex function and let Assumption 2.1 and Condition 2.1 hold. Let also $\sigma \geq L_p$. Then, for any $\gamma \in \mathbb{R}^n$, after $T + 1$ iterations of Algorithm 1, we have the following bound for the objective residual in problem (1):

$$f(x_{T+1}) - f(x^*)$$

$$\leq \frac{\delta_1}{\gamma} (T + 1) + \frac{\gamma \delta_1 D^2 (p + 1)^2}{T + p + 1} + \sum_{i=2}^{p} \frac{\delta_i D^i}{(i - 1)! (T + p + 1)^{i-1}}$$

$$+ \frac{(L_p + p\sigma) D^{p+1}}{(p + 1)!} \frac{(p + 1)^{p+1}}{(T + p + 1)^p}.$$  \hspace{1cm} (26)

By introducing $O(\cdot)$ notation and fixing $\gamma = O(T/D)$, we get

$$f(x_{T+1}) - f(x^*) \leq \sum_{i=1}^{p} O \left( \frac{\delta_i D^i}{T^{i-1}} \right) + O \left( \frac{L_p D^{p+1}}{T^p} \right),$$  \hspace{1cm} (27)

where

$$D = \max_{x \in \mathcal{L}'(x_0)} \|x - x_0\|.$$  \hspace{1cm} (28)

Proof: Applying Lemma 4.1 for any $t \geq 0$, we obtain

$$f(x_{t+1}) \leq \min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{\delta_1}{\gamma} + 2\gamma \delta_1 d_2 (y - x_t) + \sum_{i=2}^{p} \frac{i \delta_i}{(i - 1)!} d_i (y - x_t) \right\}$$

$$\leq \min_{\alpha_t \in [0,1]} \left\{ f(x_t + \alpha_t (x^* - x_t)) + \frac{\delta_1}{\gamma} + \gamma \delta_1 (\alpha_t D)^2 + \sum_{i=2}^{p} \frac{\delta_i}{(i - 1)!} (\alpha_t D)^i \right\}$$

$$\leq \min_{\alpha_t \in [0,1]} \left\{ (1 - \alpha_t) f(x_t) + \alpha_t f(x^*) + \frac{\delta_1}{\gamma} + \gamma \delta_1 (\alpha_t D)^2 + \sum_{i=2}^{p} \frac{\delta_i}{(i - 1)!} (\alpha_t D)^i \right\}$$

$$\leq \min_{\alpha_t \in [0,1]} \left\{ \frac{L_p + p\sigma}{p!} (\alpha_t D)^{p+1} \right\}.$$  \hspace{1cm} (28)

Therefore, subtracting $f(x^*)$ from both sides, we obtain, for any $\alpha_t \in [0,1]$

$$f(x_{t+1}) - f(x^*) \leq (1 - \alpha_t) (f(x_t) - f(x^*))$$

$$\leq \frac{\delta_1}{\gamma} + \gamma \delta_1 (\alpha_t D)^2 + \sum_{i=2}^{p} \frac{\delta_i}{(i - 1)!} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p + 1)!} (\alpha_t D)^{p+1}. \hspace{1cm} (29)$$
Let us define the sequence \( \{A_t\}_{t \geq 0} \) as follows:

\[
A_t = \begin{cases} 
1, & t = 0 \\
\prod_{i=1}^{t} (1 - \alpha_i), & t \geq 1.
\end{cases}
\]

Then, \( A_t = (1 - \alpha_t)A_{t-1} \). Also, we define \( \alpha_0 = 1 \). Then, dividing both sides of (29) by \( A_t \), we get

\[
\frac{1}{A_t} \left( f(x_{t+1}) - f(x^*) \right) \leq \frac{1}{A_t} (1 - \alpha_t)(f(x_t) - f(x^*)) 
\]

\[
+ \frac{1}{A_t} \left( \frac{\delta_1}{\gamma} + \gamma \delta_1 (\alpha_t D)^2 + \sum_{i=2}^{P} \frac{\delta_i}{(i-1)!} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p+1)!} (\alpha_t D)^{p+1} \right) 
\]

\[
= \frac{1}{A_{t-1}} (f(x_t) - f(x^*)) + \frac{1}{A_t} \left( \frac{\delta_1}{\gamma} + \gamma \delta_1 (\alpha_t D)^2 
\right.
\]

\[
+ \sum_{i=2}^{P} \frac{\delta_i}{(i-1)!} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p+1)!} (\alpha_t D)^{p+1} \right).
\]

Summing both sides of inequality (31) for \( t = 0, \ldots, T \), we obtain:

\[
\frac{1}{A_T} (f(x_{T+1}) - f(x^*)) \leq \frac{(1 - \alpha_0)}{A_0} \left( f(x_0) - f(x^*) \right) 
\]

\[
+ \left( \frac{\delta_1}{\gamma} \sum_{t=0}^{T} \frac{1}{A_t} + \gamma \delta_1 D^2 \sum_{t=0}^{T} \frac{\alpha_t^2}{A_t} + \sum_{i=2}^{P} \frac{\delta_i}{(i-1)!} \sum_{t=0}^{T} \frac{\alpha_t^i}{A_t} \right) 
\]

\[
+ \left( \frac{L_p + p\sigma}{(p+1)!} \sum_{t=0}^{T} \frac{\alpha_t^{p+1}}{A_t} \right)
\]

\[
= \frac{1}{A_0} \frac{\delta_1}{\gamma} \sum_{t=0}^{T} \frac{1}{A_t} + \gamma \delta_1 D^2 \sum_{t=0}^{T} \frac{\alpha_t^2}{A_t} + \sum_{i=2}^{P} \frac{\delta_i}{(i-1)!} \sum_{t=0}^{T} \frac{\alpha_t^i}{A_t} 
\]

\[
+ \left( \frac{L_p + p\sigma}{(p+1)!} \sum_{t=0}^{T} \frac{\alpha_t^{p+1}}{A_t} \right) \leq \frac{(1 - \alpha_0)}{A_0} \left( f(x_0) - f(x^*) \right).
\]

Let us take

\[
\alpha_t = \frac{p+1}{t+p+1}, \quad t \geq 1.
\]

Then, we have

\[
A_T = \prod_{t=1}^{T} (1 - \alpha_t) = \prod_{t=1}^{T} \frac{t}{t+p+1} = \frac{T!(p+1)!}{(T+p+1)!} = (p+1)! \prod_{j=1}^{p+1} \frac{1}{T+j} \leq \frac{(p+1)!}{(T+1)^{p+1}},
\]

(34)
which gives, for all \( i = 0, \ldots, p + 1, \)
\[
\sum_{t=0}^{T} \frac{A_T \alpha_t^i}{A_t} = \sum_{t=0}^{T} \frac{(p + 1)^i}{(t + p + 1)^i} \prod_{j=1}^{p+1} \frac{t + j}{T + j} = (p + 1)^i \prod_{j=1}^{p+1} \frac{1}{T + j} \sum_{t=0}^{T} \prod_{j=1}^{p+1} \frac{(t + j)}{(t + p + 1)^i}. \tag{35}
\]

To estimate the last sum we first prove that the elements of the sum are non-decreasing. Indeed, we have
\[
1 \leq 1 + \frac{1}{t + p + 1} \leq 1 + \frac{1}{t + j}, \quad \forall j \in [1, \ldots, p],
\]
and, hence, for all \( i = 1, \ldots, p + 1, \)
\[
\left(1 + \frac{1}{t + p + 1}\right)^i \leq \prod_{j=1}^{p+1} \left(1 + \frac{1}{t + j}\right)
\]
\[
\iff \left(\frac{t + p + 2}{t + p + 1}\right)^i \leq \prod_{j=1}^{p+1} \frac{t + j + 1}{t + j}
\]
\[
\iff \prod_{j=1}^{p+1} \frac{(t + j)}{(t + p + 1)^i} \leq \frac{\prod_{j=1}^{p+1} (t + 1 + j)}{(t + p + 2)^i}
\]
Thus, we have shown that the summands in the RHS of (35) are growing, whence we get the next upper bound for the sum, for all \( i = 0, \ldots, p + 1, \)
\[
\sum_{t=0}^{T} \frac{A_T \alpha_t^i}{A_t} \leq (p + 1)^i \prod_{j=1}^{p+1} \frac{1}{T + j} \sum_{t=0}^{T} \prod_{j=1}^{p+1} \frac{(t + j)}{(t + p + 1)^i}
\]
\[
\leq (p + 1)^i \prod_{j=1}^{p+1} \frac{1}{T + j} \cdot (T + 1) \cdot \frac{\prod_{j=1}^{p+1} (T + j)}{(T + p + 1)^i} \leq \frac{(T + 1)(p + 1)^i}{(T + p + 1)^i}. \tag{36}
\]
Additionally, for all \( i = 1, \ldots, p + 1, \)
\[
\sum_{t=0}^{T} \frac{A_T \alpha_t^i}{A_t} \leq \frac{(T + 1)(p + 1)^i}{(T + p + 1)^i} \leq \frac{(p + 1)^i}{(T + p + 1)^{i-1}} \tag{37}
\]
From (32) we get the statement of the Theorem:
\[
f(x_{T+1}) - f(x^*) \leq \frac{\delta_1}{\gamma} \sum_{t=0}^{T} \frac{A_T \alpha_t}{A_t} + \gamma \delta_1 D^2 \sum_{t=0}^{T} \frac{A_T \alpha_t^2}{A_t} + \sum_{i=2}^{p} \frac{\delta_i D^i}{(i - 1)!} \sum_{t=1}^{T} \frac{A_T \alpha_t^i}{A_t}
\]
\[
+ \frac{(L_p + p \sigma)D^{p+1}}{(p + 1)!} \sum_{t=1}^{T} \frac{A_T \alpha_t^{p+1}}{A_t}
\]
Next, one can set 
\[ \gamma = \sqrt{(T + 1)(T + p + 1)} D(p + 1) = O\left(\frac{T}{D}\right). \]

Then we obtain the following convergence rate
\[
\begin{align*}
    f(x_{T+1}) - f(x^*) & \leq 2 \sqrt{\frac{T + 1}{T + p + 1}} (p + 1) \delta_1 D \\
    & \quad + \sum_{i=2}^{p} \frac{\delta_i D^i}{(i - 1)! (T + p + 1)^{i-1}} \frac{(p + 1)^i}{(p + 1)!} (T + p + 1)^p \\
    & \quad + \frac{(L_p + p\sigma) D^{p+1}}{p + 1} \frac{(p + 1)^{p+1}}{(p + 1)!} (T + p + 1)^p \\
    & = \sum_{i=1}^{p} O\left(\frac{\delta_i D^i}{T^{i-1}}\right) + O\left(\frac{L_p D^{p+1}}{T^p}\right).
\end{align*}
\]

This result provides an upper bound for the objective residual after \( T \) iterations of the Inexact Tensor Method. The last term in the RHS of (26) and (27) corresponds to the case of exact Tensor method, i.e. \( \delta_i = 0, i = 1, \ldots, p \), and provides a similar convergence rate to that of non-accelerated Tensor method in [44]. The other terms in the RHS show how inexactness in each derivative influences the convergence rate. In particular, if, for some \( i \)th derivative, the corresponding error \( \delta_i \) is non-zero, the convergence rate becomes worse since the convergence rate bound has term proportional to \( \delta_i/T^{i-1} \). Importantly, when the error \( \delta_1 \) in the gradient is non-zero, our bound does not guarantee decrease of the objective below the error \( O(\delta_1 D) \). The same effect can be observed for first-order methods [19]. We now give sufficient conditions for the errors \( \delta_i, i = 1, \ldots, p \) such that the method is still guaranteed to find an \( \varepsilon \)-solution. In particular, this result answers the following question. Assume that the errors in the derivatives can be controlled and made as small as we would like them to be. Then, how small should we take the error in each derivative if we would like to achieve objective residual smaller than \( \varepsilon \)? As we will see below in Section 7, for stochastic optimization problems, this control over the errors can be achieved by increasing the sample size for each derivative.

**Corollary 4.1:** Let assumptions of Theorem 4.1 hold and let \( \varepsilon > 0 \) be the desired solution accuracy. Further, let the levels of inexactness in Condition 2.1 satisfy the following
inequalities:
\[
\delta_1 \leq \frac{1}{2(p+1)^2} \frac{\varepsilon}{D},
\]
(38)
\[
\delta_i \leq \frac{(i-1)!}{(p+1)!} \left( \frac{L_p + p\sigma}{p!} \right)^{\frac{i-1}{p}} \left( \frac{\varepsilon}{D(p+1)} \right)^{\frac{p-i+1}{p}}, \quad \forall i \in [2, \ldots, p].
\]
(39)

Also, let the number of iterations \(T\) of Algorithm 1 satisfy
\[
T = \max \left\{ 1; \left( \frac{p+1}{p!} \right)^{\frac{1}{p}} \left( \frac{(L_p + p\sigma)D^{p+1}}{\varepsilon} \right)^{\frac{1}{p}} - p - 1 \right\},
\]
(40)
and \(\gamma\) be chosen as
\[
\gamma = \frac{T + p + 1}{D(p+1)} = \left( \frac{p+1}{p!} \frac{(L_p + p\sigma)D^{p+1}}{\varepsilon} \right)^{\frac{1}{p}}.
\]
(41)

Then \(x_{T+1}\) is an \(\varepsilon\)-solution of problem (1), i.e. \(f(x_{T+1}) - f(x^*) \leq \varepsilon\).

**Proof:** First, by taking \(\gamma\) from (40) we get
\[
\frac{\delta_1}{\gamma} (T+1) + \frac{\varepsilon \delta_1 D^2 (p+1)^2}{T + p + 1} \leq 2\delta_1 D(p+1) \leq \frac{\varepsilon}{p+1}
\]
By our choice of \(\gamma, T, \delta_i, i = 1, \ldots, p\) we have:
\[
\frac{\delta_1 D^i}{(i-1)!} \frac{(p+1)^i}{(T + p + 1)^{i-1}} \leq \frac{\varepsilon}{p+1} \quad i = 2, \ldots, p
\]
\[
\frac{(L_p + p\sigma)D^{p+1}}{(p+1)!} \frac{(p+1)^{p+1}}{(T + p + 1)^p} = \frac{\varepsilon}{p+1}.
\]
Then, from (26) we have:
\[
f(x_{T+1}) - f(x^*) \leq \varepsilon.
\]

Thus, we showed that the number of steps sufficient for the Inexact Tensor Method to find an \(\varepsilon\)-solution to problem (1) is the same as for non-accelerated exact tensor method [44] and is equal to \(O\left(\frac{L_p D^{p+1}}{\varepsilon^{1/p}}\right)\) under the assumption that the inexactness \(\delta_i\) of each derivative satisfies \(\delta_i = O\left(\frac{\varepsilon}{p+1}\right)\), \(i = 1, \ldots, p\). Interestingly, as the order of the derivative increases, the requirement for the accuracy of the derivative becomes less strict.

**Remark 4.2:** In the statement of Theorem 4.1 we defined \(D = \max_{x \in L'(x_0)} \|x - x_0\|\), where \(L'(x_0) = \left\{ x | f(x) \leq f(x_0) + \frac{T_\delta_1}{2^p} \right\}\). Choosing the inexactness level \(\delta_1\) and the number of iterations \(T\) as in Corollary 4.1, we have \(\frac{T_\delta_1}{2^p} \leq \frac{\varepsilon}{2(p+1)}\). Therefore, the Inexact Tensor Method is monotone up to a small inaccuracy \(\frac{\varepsilon}{2(p+1)}\).
5. Accelerated inexact tensor method

In this section, we use the estimating sequences technique to propose Accelerated Inexact Tensor Method and prove its convergence rate theorem providing faster rates than that of Algorithm 1.

**Algorithm 2** Accelerated Inexact Tensor Method

1: **Input:** convex function $f$ such that $\nabla^p f$ is $L_p$-Lipschitz; $x_0$ is starting point; constants $
abla f$ is $L_p$-Lipschitz; $x_0$ is starting point; constants

$$\sigma \geq \max \left\{ 1, \frac{3}{p} \right\} L_p, \eta_i \geq 3 \text{ for } i = 2, \ldots, p; \text{ nonnegative nondecreasing sequences} \{\tilde{\kappa}_i^t\}_{t \geq 0} \text{ for } i = 2, \ldots, p + 1, \text{ and}$$

$$\alpha_t = \frac{p + 1}{t + p + 1}, \quad A_t = \prod_{j=1}^{t} (1 - \alpha_j), \quad A_0 = 1. \quad (42)$$

2: **Precomputation:** Call the inexact oracle to compute $G_{x_0,i}$ for $i = 1, \ldots, p$ such that Condition 2.1 is satisfied. Compute

$$x_1 = \arg \min_{x \in \mathbb{R}^n} \{\phi_{x_0,p}(x) + \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} d_i(x - x_0) + \frac{\sigma}{(p-1)!} d_{p+1}(x - x_0)\} \quad (43)$$

$$y_1 = \arg \min_{x \in \mathbb{R}^n} \left\{ \psi_1(x) := f(x_1) + \sum_{i=2}^{p+1} \frac{\tilde{\kappa}_i^0}{(i-1)!} d_i(x - x_0) \right\}. \quad (44)$$

3: **for** $t \geq 1$ **do**

4: Set

$$v_t = (1 - \alpha_t)x_t + \alpha_t y_t, \quad (45)$$

5: Call the inexact oracle to compute $G_{v_t,i}$ for $i = 1, \ldots, p$ such that Condition 2.1 is satisfied and set

$$x_{t+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ \phi_{v_t,p}(x) + \sum_{i=2}^{p} \frac{\eta_i \delta_i}{(i-1)!} d_i(x - v_t) + \frac{\sigma}{(p-1)!} d_{p+1}(x - v_t) \right\}. \quad (46)$$

6: Compute

$$y_{t+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ \psi_{t+1}(x) := \psi_t(x) + \sum_{i=2}^{p+1} \frac{\tilde{\kappa}_i^t - \tilde{\kappa}_i^{t-1}}{(i-1)!} d_i(x - x_0) + \frac{\alpha_t}{A_t} \phi_{x_{t+1},1}(x) \right\}. \quad (47)$$

7: **end for**
In Algorithm 2 we utilize the estimating sequence technique to obtain acceleration. In this technique, to prove the convergence theorem (Theorem 5.1) one needs to construct a lower and an upper bound for the estimating sequence $\psi_t(x)$ based on the objective function. So, the full proof is organized as follows:

- Lemma 5.1 provides an upper bound for the estimating sequence $\psi_t(x)$ (44), (47);
- Lemma 5.5 provides a lower bound on $\psi_t(x)$ based on results of technical Lemmas 5.3-5.4;
- finally, everything is combined together in Theorem 5.1 in order to prove convergence and obtain convergence rate.

The following lemma shows that the sequence of functions $\psi_t(x)$ generated by Algorithm 2 can be upper bounded by the properly regularized objective function.

**Lemma 5.1:** If Condition 2.1 is satisfied and $\psi_t(x)$ is defined as in (44), (47) then, for any $x \in \mathbb{R}^n$, we have

$$
\psi_t(x) \leq \frac{f(x)}{A_{t-1}} + \delta_1 \|x_1 - x_0\| + \delta_1 \|x - x_0\| + \sum_{i=2}^{p} \frac{2\delta_i + \bar{\kappa}_{i}^{t-1}}{(i-1)!} d_i(x - x_0) + L_p + p\sigma + \bar{\kappa}_{p+1}^{t-1} \frac{1}{p!} d_{p+1}(x - x_0) + \sum_{i=1}^{t-1} \frac{\alpha_i}{A_i} (G_{x_{i+1},1} - \nabla f(x_{i+1}), x - x_{i+1}).
$$

**Proof:**

$$
f(x_1) \begin{array}{c} (13) \\ \leq \phi_{x_0,p}(x_1) + \delta_1 \|x_1 - x_0\| + \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} d_i(x_1 - x_0) \\
+ \frac{L_p}{p!} d_{p+1}(x_1 - x_0)
\end{array}
$$

$$
L_p \leq \sigma \begin{array}{c} \leq \phi_{x_0,p}(x_1) + \delta_1 \|x_1 - x_0\| + \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} d_i(x_1 - x_0) + \frac{\sigma}{p!} d_{p+1}(x_1 - x_0) \\
\leq \phi_{x_0,p}(x_1) + \delta_1 \|x_1 - x_0\| + \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} d_i(x_1 - x_0) + \frac{p\sigma}{p!} d_{p+1}(x_1 - x_0)
\end{array}
$$

$$
\sigma \leq \rho \sigma \begin{array}{c} \leq \phi_{x_0,p}(x) + \delta_1 \|x_1 - x_0\| + \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} d_i(x - x_0) + \frac{\rho \sigma}{p!} d_{p+1}(x - x_0)
\end{array}
$$

$$
\begin{array}{c} (43) \\ \leq f(x) + \delta_1 \|x_1 - x_0\| + \delta_1 \|x - x_0\| + \sum_{i=2}^{p} \frac{2\delta_i}{(i-1)!} d_i(x - x_0) \\
+ \frac{L_p + p\sigma}{p!} d_{p+1}(x - x_0).
\end{array}
$$
Therefore,

\[ \psi_1(x) = f(x_1) + \sum_{i=2}^{p+1} \frac{\kappa_i^0}{(i-1)!} d_i(x - x_0) \]

\[ \leq f(x) + \delta_1 \| x_1 - x_0 \| + \delta_1 \| x - x_0 \| + \sum_{i=2}^{p} \frac{2\delta_i + \kappa_i^0}{(i-1)!} d_i(x - x_0) \]

\[ + \frac{L_p + p\sigma + p\kappa_p^0}{p!} d_{p+1}(x - x_0). \] (48)

From (47) we have

\[ \psi_1(x) = \psi_1(x) + \sum_{i=2}^{p+1} \frac{\kappa_i^t - \kappa_i^0}{(i-1)!} d_i(x - x_0) + \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \phi_{x_{j+1},1}(x) \]

\[ = \psi_1(x) + \sum_{i=2}^{p+1} \frac{\kappa_i^t - \kappa_i^0}{(i-1)!} d_i(x - x_0) + \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \phi_{x_{j+1},1}(x) \]

\[ + \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \left( G_{x_{j+1},1} - \nabla f(x_{j+1}), x - x_{j+1} \right). \] (49)

From (42) we have that, for all \( j \geq 1, \) \( A_j = A_{j-1}(1 - \alpha_j), \) which leads to \( \frac{\alpha_j}{A_j} = \frac{1}{A_j} - \frac{1}{A_{j-1}}. \)

Hence, we have \( \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} = \frac{1}{A_{t-1}} - \frac{1}{A_0} \) and, using the convexity of the objective \( f, \) we get

\[ \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \phi_{x_{j+1},1}(x) \leq f(x) \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} = f(x) \left( \frac{1}{A_{t-1}} - \frac{1}{A_0} \right). \] (50)

Finally, combining all the inequalities from above and using that \( A_0 = 1, \) we obtain

\[ \psi_1(x) \leq \psi_1(x) + \sum_{i=2}^{p+1} \frac{\kappa_i^t - \kappa_i^0}{(i-1)!} d_i(x - x_0) + f(x) \left( \frac{1}{A_{t-1}} - \frac{1}{A_0} \right) \]

\[ + \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \left( G_{x_{j+1},1} - \nabla f(x_{j+1}), x - x_{j+1} \right) \] (49),(50)

\[ \leq \frac{f(x)}{A_{t-1}} + \delta_1 \| x_1 - x_0 \| + \delta_1 \| x - x_0 \| + \sum_{i=2}^{p} \frac{2\delta_i + \kappa_i^t}{(i-1)!} d_i(x - x_0) \]

\[ + \frac{L_p + p\sigma + p\kappa_p^t}{p!} d_{p+1}(x - x_0) + \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \left( G_{x_{j+1},1} - \nabla f(x_{j+1}), x - x_{j+1} \right). \] (51)
The next Lemma characterizes the progress of the tensor step (46) in Algorithm 2.

**Lemma 5.2:** Let \( \{x_t, v_t\}_{t \geq 1} \) be generated by Algorithm 2. Then, the following holds.

- If
  \[
  \delta_1 \geq \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-1} + \frac{L_p}{p!} \|x_{t+1} - v_t\|^p, \tag{52}
  \]
  then
  \[
  f(x_{t+1}) - f(x^*) \leq \left( \max \left\{ \frac{p\sigma}{L_p} ; \max_i \{\eta_i\} \right\} + 2 \right) \delta_1 \|x_{t+1} - x^*\|; \tag{53}
  \]
- else
  \[
  \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle \\
  \geq \frac{\sqrt{5}}{3} \min \left\{ \min_{i=2}^{p} \|\nabla f(x_{t+1})\| \frac{(i-1)!}{p(\eta_i + 2)\delta_i} \right\}. \tag{54}
  \]

**Proof:** For simplicity, we denote
\[
\zeta_{t+1} = \sum_{i=2}^{p} \frac{\eta_i \delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-2} + \frac{\sigma}{(p-1)!} \|x_{t+1} - v_t\|^{p-1}. \tag{54'}
\]
By the optimality condition in (46)
\[
0 = \nabla \phi_{v_t,p}(x_{t+1}) + \sum_{i=2}^{p} \frac{\eta_i \delta_i}{(i-1)!} \nabla d_i(x_{t+1} - v_t) + \frac{\sigma}{(p-1)!} \nabla d_{p+1}(x_{t+1} - v_t)
\]
\[
= \nabla \phi_{v_t,p}(x_{t+1}) + \zeta_{t+1}(x_{t+1} - v_t). \tag{55}
\]
From the optimality condition (55) we obtain
\[
\|\nabla f(x_{t+1})\| = \|\nabla f(x_{t+1}) - \nabla \phi_{v_t,p}(x_{t+1}) - \zeta_{t+1}(x_{t+1} - v_t)\|
\]
\[
\leq \|\nabla \phi_{v_t,p}(x_{t+1}) - \nabla f(x_{t+1})\| + \zeta_{t+1}\|x_{t+1} - v_t\|
\]
\[
\leq \delta_1 + \sum_{i=2}^{p} \frac{(\eta_i + 1)\delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-1} + \frac{L_p + p\sigma}{p!} \|x_{t+1} - v_t\|^p. \tag{56}
\]
If \( \delta_1 \) is dominating, i.e. (52) holds, then
\[
\|\nabla f(x_{t+1})\| \leq \delta_1 + \sum_{i=2}^{p} \frac{(\eta_i + 1)\delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-1} + \frac{L_p + p\sigma}{p!} \|x_{t+1} - v_t\|^p
\]
\[
\leq \left( \max \left\{ \frac{p\sigma}{L_p} ; \max_{i \in \{2, ..., p\}} \{\eta_i\} \right\} + 2 \right) \delta_1. \tag{52'}
\]
Since \( f \) is convex, we have
\[
f(x_{t+1}) - f(x^*) \leq \langle \nabla f(x_{t+1}), x_{t+1} - x^* \rangle \leq \| \nabla f(x_{t+1}) \| \| x_{t+1} - x^* \|
\]
\[
\leq \left( \max \left\{ \frac{p\sigma}{L_p}, \max_{i \in \{2, \ldots, p\}} \{ \eta_i \} \right\} + 2 \right) \delta_1 \| x_{t+1} - x^* \|.
\]
This proves the first statement.

Next, from the previous inequality and optimality condition (55), we get
\[
\text{above inequality with (14) we have that} \quad \delta_1 \leq \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} \| x_{t+1} - v_t \|^{i-1} + \frac{L_p}{p!} \| x_{t+1} - v_t \|^p.
\]
We start with getting an upper bound for \( \| \nabla \phi_{\nu_{t, p}}(x_{t+1}) - \nabla f(x_{t+1}) \| \). Combining the above inequality with (14) we have
\[
\| \nabla \phi_{\nu_{t, p}}(x_{t+1}) - \nabla f(x_{t+1}) \|
\]
\[
\leq \delta_1 + \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} \| x_{t+1} - v_t \|^{i-1} + \frac{L_p}{p!} \| x_{t+1} - v_t \|^p
\]
\[
\leq \sum_{i=2}^{p} \frac{2\delta_i}{(i-1)!} \| x_{t+1} - v_t \|^{i-1} + \frac{2L_p}{p!} \| x_{t+1} - v_t \|^p
\]
\[
\leq \sum_{i=2}^{p} \frac{2\eta_i \delta_i}{(i-1)!} \| x_{t+1} - v_t \|^{i-1} + \frac{2L_p}{p!} \| x_{t+1} - v_t \|^p
\]
\[
\leq \frac{2}{3} \sum_{i=2}^{p} \frac{\eta_i \delta_i}{(i-1)!} \| x_{t+1} - v_t \|^{i-1} + \frac{\sigma}{(p-1)!} \| x_{t+1} - v_t \|^p
\]
\[
\leq \frac{2}{3} \eta \| x_{t+1} - v_t \|
\]
Next, from the previous inequality and optimality condition (55), we get
\[
\| \nabla \phi_{\nu_{t, p}}(x_{t+1}) - \nabla f(x_{t+1}) \|^2 \leq \| \nabla f(x_{t+1}) + \zeta_{t+1} (x_{t+1} - v_t) \|^2
\]
\[
= 2\langle \nabla f(x_{t+1}), x_{t+1} - v_t \rangle \zeta_{t+1} + \zeta_{t+1}^2 \| x_{t+1} - v_t \|^2 + \| \nabla f(x_{t+1}) \|^2
\]

Hence,
\[
2\langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle \zeta_{t+1} \geq \| \nabla f(x_{t+1}) \|^2 + \frac{5}{9} \zeta_{t+1}^2 \| x_{t+1} - v_t \|^2
\]
\[
\geq \frac{2\sqrt{5}}{3} \zeta_{t+1} \| \nabla f(x_{t+1}) \| \| x_{t+1} - v_t \|
\]
Dividing both sides by \( 2\zeta_{t+1} \), we finally get
\[
\langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle \geq \frac{\sqrt{5}}{3} \| \nabla f(x_{t+1}) \| \| x_{t+1} - v_t \|.
\]
To bound $\|x_{t+1} - v_t\|$ from below using the gradient norm $\|\nabla f(x_{t+1})\|$, we use inequality (56):

$$\|\nabla f(x_{t+1})\| \leq \delta_1 + \sum_{i=2}^{p} \frac{(\eta_i + 1)\delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-1} + \frac{L_p + \rho\sigma}{p!} \|x_{t+1} - v_t\|^p$$

$$\leq \sum_{i=2}^{p} \frac{(\eta_i + 2)\delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-1} + \frac{2L_p + \rho\sigma}{p!} \|x_{t+1} - v_t\|^p.$$

Next, we consider $p-1$ cases depending on which term dominates in the RHS of the last inequality.

- If, for some $i = 2, \ldots, p$, the term $\frac{(\eta_i + 2)\delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-1}$ dominates the others, then we get the following bound

$$\|x_{t+1} - v_t\| \geq \left( \frac{\|\nabla f(x_{t+1})\|((i-1)!)^{\frac{1}{i-1}}}{p(\eta_i + 2)\delta_i} \right)^{\frac{1}{i-1}}.$$

- If the term $\frac{2L_p + \rho\sigma}{p!} \|x_{t+1} - v_t\|^p$ dominates the others, then

$$\|x_{t+1} - v_t\| \geq \left( \frac{\|\nabla f(x_{t+1})\|((p-1)!)^{\frac{1}{p-1}}}{2L_p + \rho\sigma} \right)^{\frac{1}{p-1}}.$$

Clearly, $\|x_{t+1} - v_t\|$ is bounded from below by the minimum of these $p$ lower bounds.

Combining this minimum bound with the bound (58) we finish the proof. 

We will also use the next technical lemma [31,42] on Fenchel conjugate for the $p$th power of the norm.

**Lemma 5.3:** Let $g(z) = \frac{p}{p} \|z\|^p$ for $p \geq 2$ and $g^*$ be its conjugate function i.e. $g^*(v) = \sup_{z} \{ \langle v, z \rangle - g(z) \}$. Then, we have

$$g^*(v) = \frac{p-1}{p} \left( \frac{\|v\|^p}{\theta} \right)^{\frac{1}{p-1}}$$

Moreover, for any $v, z \in \mathbb{R}^n$, we have $g(z) + g^*(v) - \langle z, v \rangle \geq 0$.

The next step is to provide a lower bound $\psi_t(x) \geq \psi_t^* := \min_x \psi_t(x) \geq \frac{f(x_t)}{\lambda_{t-1}} + \text{err}_t$ for all $x$, where $\text{err}_t$ is some error term that will be defined later. The convergence rate of Algorithm 2 will follow from this bound and Lemma 5.1. The proof of the desired lower bound is quite technical and requires several auxiliary lemmas. After that we combine all the technical results together to obtain convergence rate in the proof of Theorem 5.1. We start the technical derivations with the following result.
Lemma 5.4: Let \( h(x) \) be a convex function, \( x_0 \in \mathbb{R}^n \), \( \theta_i \geq 0 \) for \( i = 2, \ldots, p + 1 \) and

\[
\bar{x} = \arg \min_{x \in \mathbb{R}^n} \{ \bar{h}(x) = h(x) + \sum_{i=2}^{p+1} \theta_i d_i(x - x_0) \}.
\]

Then, for all \( x \in \mathbb{R}^n \),

\[
\bar{h}(x) \geq \bar{h}(\bar{x}) + \sum_{i=2}^{p+1} \left( \frac{1}{2} \right)^{i-2} \theta_i d_i(x - \bar{x}).
\]

Proof: In [42] it is shown that, for all \( x, y \in \mathbb{R}^n \) and for any \( i \geq 2 \),

\[
d_i(x) - d_i(y) - \langle \nabla d_i(y), x - y \rangle \geq \left( \frac{1}{2} \right)^{i-2} d_i(x - y).
\]

Using the convexity of \( h \), we have

\[
\bar{h}(x) = h(x) + \sum_{i=2}^{p+1} \theta_i d_i(x - x_0) \geq h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle + \sum_{i=2}^{p+1} \theta_i d_i(x - x_0)
\]

\[
\geq h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle
\]

\[
+ \sum_{i=2}^{p+1} \theta_i \left( d_i(\bar{x} - x_0) + \langle \nabla d_i(\bar{x} - x_0), x - \bar{x} \rangle + \left( \frac{1}{2} \right)^{i-2} d_i(x - \bar{x}) \right)
\]

\[
= \bar{h}(\bar{x}) + \langle \nabla \bar{h}(\bar{x}), x - \bar{x} \rangle + \sum_{i=2}^{p+1} \left( \frac{1}{2} \right)^{i-2} \theta_i d_i(x - \bar{x}) \geq \bar{h}(\bar{x})
\]

\[
+ \sum_{i=2}^{p+1} \left( \frac{1}{2} \right)^{i-2} \theta_i d_i(x - \bar{x}),
\]

where the last inequality holds by optimality condition since \( \bar{h}(x) \) is convex. □

Finally, the last technical step is the next Lemma which will be a part of the induction step to prove that \( \frac{f(x_t)}{A_{t-1}} + err_t \leq \min_x \psi_t(x) = \psi^*_t \).

Lemma 5.5: Let \( \{x_t, y_t\}_{t \geq 1} \) be generated by Algorithm 2 and define

\[
err_t := \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} (G_{x_{j+1,1}} - \nabla f(x_{j+1}), y_{j+1} - x_{j+1}).
\]  \hspace{1cm} (59)

Assume also that

\[
\psi^*_t := \min_x \psi_t(x) \geq \frac{f(x_t)}{A_{t-1}} + err_t.
\]  \hspace{1cm} (60)
Then,

\[ \psi_{t+1}^* \geq \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} (\nabla f(x_{t+1}), v_t - x_{t+1}) + \sum_{i=2}^{p+1} \frac{1}{2} \frac{\bar{k}_i^{t-1}}{(i-1)!} d_i (y_{t+1} - y_t) \]

\[ + \frac{\alpha_t}{A_t} (\nabla f(x_{t+1}), y_{t+1} - y_t) + \epsilon_{t+1}. \]

**Proof:** By definition,

\[ \psi_t(x) = f(x_1) + \sum_{i=2}^{p+1} \frac{\bar{k}_i^{t-1}}{(i-1)!} d_i (x - x_0) + \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \phi_{x_{j+1}, 1}(x). \]

Next, we apply Lemma 5.4 with the following choice of parameters: \( h(x) = f(x_1) + \sum_{j=1}^{t-1} \frac{\alpha_j}{A_j} \phi_{x_{j+1}, 1}(x) \), \( \theta_i = \frac{\bar{k}_i^{t-1}}{(i-1)!} \) for \( i = 2, \ldots, p + 1. \)

By (47), \( y_t = \arg\min_{x \in \mathbb{R}^p} \bar{h}(x) \), and we have

\[ \psi_t(x) \geq \psi_t^* + \sum_{i=2}^{p+1} \frac{1}{2} \frac{\bar{k}_i^{t-1}}{(i-1)!} d_i (x - y_t) \]

\[ \geq \frac{f(x_t)}{A_{t-1}} + \sum_{i=2}^{p+1} \frac{1}{2} \frac{\bar{k}_i^{t-1}}{(i-1)!} d_i (x - y_t) + \epsilon_t, \]

where the last inequality follows from the assumption of the lemma.

By the definition of \( \psi_{t+1}(x) \), the above inequality, and convexity of \( f \), we obtain

\[ \psi_{t+1}(x) = \psi_t(x) + \sum_{i=2}^{p+1} \frac{\bar{k}_i^t - \bar{k}_i^{t-1}}{(i-1)!} d_i (x - x_0) + \frac{\alpha_t}{A_t} \phi_{x_{t+1}, 1}(x) \]

\[ \geq \frac{f(x_t)}{A_{t-1}} + \sum_{i=2}^{p+1} \frac{1}{2} \frac{\bar{k}_i^{t-1}}{(i-1)!} d_i (x - y_t) + \frac{\alpha_t}{A_t} \phi_{x_{t+1}, 1}(x) + \epsilon_t \]

\[ \geq \frac{1}{A_{t-1}} (f(x_{t+1}) + (\nabla f(x_{t+1}), x_t - x_{t+1})) + \sum_{i=2}^{p+1} \frac{1}{2} \frac{\bar{k}_i^{t-1}}{(i-1)!} d_i (x - y_t) \]

\[ + \frac{\alpha_t}{A_t} \Phi_{x_{t+1}, 1}(x) + \frac{\alpha_t}{A_t} \Phi_{x_{t+1}, 1}(x) + \epsilon_t \]

Next, we consider the sum of two linear models from the last inequality:

\[ \frac{1}{A_{t-1}} (f(x_{t+1}) + (\nabla f(x_{t+1}), x_t - x_{t+1})) + \frac{\alpha_t}{A_t} \Phi_{x_{t+1}, 1}(x) \]

\[ = \frac{1}{A_{t-1}} (f(x_{t+1}) + (\nabla f(x_{t+1}), x_t - x_{t+1})) + \frac{\alpha_t}{A_t} (f(x_{t+1}) + (\nabla f(x_{t+1}), x_t - x_{t+1})) \]
\[
\begin{align*}
&= \frac{1 - \alpha_t}{A_t} f(x_{t+1}) + \frac{1 - \alpha_t}{A_t} \langle \nabla f(x_{t+1}), x_t - x_{t+1} \rangle \\
&\quad + \frac{\alpha_t}{A_t} f(x_{t+1}) + \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle \\
\end{align*}
\]

\[
\begin{align*}
\frac{f(x_{t+1})}{A_t} + \frac{1 - \alpha_t}{A_t} \left( \nabla f(x_{t+1}), v_t - \alpha_t y_t - x_{t+1} \right) + \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle \\
= \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle + \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), x - y_t \rangle.
\end{align*}
\]

Therefore,
\[
\psi_{t+1}(x) \geq \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle + \sum_{i=2}^{p+1} \left( \frac{1}{2} \right)^{i-2} \frac{\bar{k}_i^{-1}}{(i-1)!} d_i (x - y_t) \\
+ \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), x - y_t \rangle + \frac{\alpha_t}{A_t} \langle G_{x_{t+1},1} - \nabla f(x_{t+1}), x - x_{t+1} \rangle + \text{err}_t.
\]

Finally, by (47), we get
\[
\psi_{t+1}^* = \psi_{t+1}(y_{t+1}) \geq \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle \\
+ \sum_{i=2}^{p+1} \left( \frac{1}{2} \right)^{i-2} \frac{\bar{k}_i^{-1}}{(i-1)!} d_i (y_{t+1} - y_t) \\
+ \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), y_{t+1} - y_t \rangle + \text{err}_{t+1}.
\]

Finally, we are in a position to prove the convergence rate theorem for the Accelerated Inexact Tensor Method (Algorithm 2). The proof uses the following technical assumption.

**Assumption 5.1:** Let \( \{x_t, y_t\}_{t \geq 1} \) be generated from Algorithm 2. Then there exists \( \bar{R} > 0 \) such that
\[
\begin{align*}
\|x_1 - x_0\| \leq \bar{R}, \quad &\|x_t - x^*\| \leq \bar{R} \quad \forall \ t \geq 1, \\
\|y_t - x^*\| \leq \bar{R} \quad &\forall \ t \geq 1.
\end{align*}
\]

Note, that in the case of exact gradient, i.e. \( \delta_1 = 0 \), we do not need Assumption 5.1. Let also \( R \) be such that
\[
\|x_0 - x^*\| \leq R.
\]

**Theorem 5.1:** Let Assumption 5.1 hold, Condition 2.1 be satisfied, \( f(x) \) be convex and \( p \) times differentiable function with Lipschitz constant \( L_p \) for the \( p \)th derivative, \( \sigma \geq \)
\[
\max \left\{ \frac{1}{3p}; \frac{3}{p} \right\} L_p, \text{ and } \eta_i \geq 3, \ i = 1, \ldots, p. \text{ Assume also that Algorithm 2 makes } T \geq 1 \text{ iterations with parameters}
\]
\[
\tilde{\kappa}^t_i = \frac{p}{2} \left( \frac{(i-1)6}{i\sqrt{5}} \right)^{i-1} (\eta_i + 2) \frac{\alpha_i^t}{A_t} \delta_i, i = 2, \ldots, p,
\]
\[
\tilde{\kappa}^t_{p+1} = \frac{1}{2} \left( \frac{6p}{\sqrt{5}(p+1)} \right)^p \frac{\alpha_{p+1}^t}{A_t} (2L_p + p\sigma).
\]

If, for all iterations \( t \leq T \), we have
\[
\delta_1 \leq \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} \|x_t - v_t\|^{i-1} + \frac{L_p}{p!} \|x_t - v_t\|^p,
\]
then we have the following bound for the objective residual:
\[
f(x_T) - f(x^*) \leq 2\delta_1 \tilde{R} + \sum_{i=2}^{p} \frac{p}{2} \left( \frac{6(i-1)}{\sqrt{5}i} \right)^{i-1} \frac{(p+1)^i(\eta_i + 2)}{i!} \frac{\delta_i R^{i+1}}{(T+p+1)^i}
\]
\[
+ \frac{\left( \frac{6p}{\sqrt{5}} \right)^p}{p!} \frac{(2L_p + p\sigma) R^{p+1}}{(T+p+1)^{p+1}},
\]

(64)
or, in other words,
\[
f(x_T) - f(x^*) \leq O(\delta_1 \tilde{R}) + \sum_{i=2}^{p} O \left( \frac{\delta_i R^{i+1}}{T^i} \right) + O \left( \frac{L_p R^{p+1}}{T^{p+1}} \right).
\]

(65)

Otherwise, if for some iteration \( t \leq T \),
\[
\delta_1 \geq \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} \|x_t - v_t\|^{i-1} + \frac{L_p}{p!} \|x_t - v_t\|^p,
\]
then
\[
f(x_t) - f(x^*) \leq \left( \max \left\{ \frac{p\sigma}{L_p}; \max_{i\in\{2,\ldots,p\}} \{\eta_i\} \right\} + 2 \right) \delta_1 \tilde{R} = O(\delta_1 \tilde{R}).
\]

(66)

**Proof:** Firstly, assuming that for all iterations \( t \leq T \) it holds that
\[
\delta_1 \leq \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} \|x_t - v_t\|^{i-1} + \frac{L_p}{p!} \|x_t - v_t\|^p,
\]
we show by induction that, for all \( t \geq 0 \),
\[
\frac{f(x_t)}{A_{t-1}} + \text{err}_t \leq \psi_t^*,
\]
where \( err_t \) is defined in (59). From (44), since \( A_0 = 1 \), we have that \( \frac{f(x_t)}{A_t} \leq \psi_t^* \). Let us assume that

\[
\frac{f(x_t)}{A_t} + err_t \leq \psi_t^*
\]

and show that \( \frac{f(x_{t+1})}{A_t} + err_{t+1} \leq \psi_{t+1}^* \). By Lemma 5.5, we have

\[
\psi_{t+1}^* \geq \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle + \sum_{i=2}^{p+1} \left( \frac{1}{2} \right)^{i-2} \frac{k_{i-1}}{(i-1)!} d_i(y_{t+1} - y_t)
\]

\[
+ \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), y_{t+1} - y_t \rangle + \sum_{j=1}^{t} \frac{\alpha_j}{A_j} \langle G_{x_{j+1},1} - \nabla f(x_{j+1}), y_{j+1} - x_{j+1} \rangle.
\]

(67)

Therefore, to complete the induction step we need to show, that the sum of all terms in the RHS except \( \frac{f(x_{t+1})}{A_t} \) and error terms is non-negative.

Lemma 5.2 provides the lower bound for \( \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle \). Let us consider the case when the minimum in the RHS of (53) is attained at the first term with particular \( i = 2, \ldots, p \). By Lemma 5.3 with the following choice of the parameters

\[
z = y_t - y_{t+1}, \quad v = \frac{\alpha_t}{A_t} \nabla f(x_{t+1}), \quad \theta = \left( \frac{1}{2} \right)^{i-2} \frac{k_{i-1}}{(i-1)!},
\]

we have

\[
\frac{1}{i} \left( \frac{1}{2} \right)^{i-2} \frac{k_{i-1}}{(i-1)!} \| y_t - y_{t+1} \|^i + \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), y_{t+1} - y_t \rangle
\]

\[
\geq - \frac{i - 1}{i} \left( \frac{\| \nabla f(x_{t+1}) \|}{\left( \frac{1}{2} \right)^{i-2} \frac{k_{i-1}}{(i-1)!}} \right)^{\frac{1}{i-1}}.
\]

(68)

Hence,

\[
\frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle + \left( \frac{1}{2} \right)^{i-2} \frac{k_{i-1}}{(i-1)!} d_i(y_{t+1} - y_t)
\]

\[
+ \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), y_{t+1} - y_t \rangle
\]

\[
\overset{(68)}{\geq} \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle - \frac{i - 1}{i} \left( \frac{\| \nabla f(x_{t+1}) \|}{\left( \frac{1}{2} \right)^{i-2} \frac{k_{i-1}}{(i-1)!}} \right)^{\frac{1}{i-1}}
\]

\[
\overset{(53)}{\geq} \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \frac{\sqrt{5}}{3} \| \nabla f(x_{t+1}) \|^\frac{1}{i-1} \left( \frac{(i-1)!}{p(\eta_t + 2)\delta_i} \right)^\frac{1}{i-1} - \frac{i - 1}{i} \left( \frac{\| \nabla f(x_{t+1}) \|}{\left( \frac{1}{2} \right)^{i-2} \frac{k_{i-1}}{(i-1)!}} \right)^{\frac{1}{i-1}}
\]

\[
\geq \frac{f(x_{t+1})}{A_t},
\]
where the last inequality holds by our choice of the parameters

\[
\tilde{\kappa}_i^t \geq \frac{p}{2} \left( \frac{(i-1)6}{i \sqrt{5}} \right)^{i-1} (\eta_i + 2) \frac{\alpha_{i+1}^t}{A_{t+1}} \delta_i, \quad i = 2, \ldots, p. \tag{69}
\]

Next, we consider the case when the minimum in the RHS of (53) is achieved on the second term. Again, by Lemma 5.3 with the same choice of \( z, v \) and with \( \theta = \left( \frac{1}{2} \right)^{p-1} \tilde{\kappa}^{t-1}_{p+1} (p-1)! \), we have

\[
\frac{1}{p} \left( \frac{1}{2} \right)^{p-1} \tilde{\kappa}^{t-1}_{p+1} \frac{p!}{p!} \| y_t - y_{t+1} \|^{p+1} + \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), y_{t+1} - y_t \rangle \geq - \frac{p}{p+1} \left( \frac{\| x_t^\alpha \nabla f(x_{t+1}) \|^{p+1}}{\left( \frac{1}{2} \right)^{p-1} \tilde{\kappa}^{t-1}_{p+1} (p-1)!} \right)^{\frac{1}{p}}. \tag{70}
\]

Hence, we get

\[
\frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle + \frac{1}{2p-1} \frac{\alpha_t}{p!} d_{p+1}(y_{t+1} - y_t) + \frac{\alpha_t}{A_t} \langle \nabla f(x_{t+1}), y_{t+1} - y_t \rangle \geq \frac{f(x_{t+1})}{A_t} + \frac{1}{A_t} \langle \nabla f(x_{t+1}), v_t - x_{t+1} \rangle - \frac{p}{p+1} \left( \frac{\| x_t^\alpha \nabla f(x_{t+1}) \|^{p+1}}{\left( \frac{1}{2} \right)^{p-1} \tilde{\kappa}^{t-1}_{p+1} (p-1)!} \right)^{\frac{1}{p}} \geq \frac{f(x_{t+1})}{A_t},
\]

where the last inequality holds by our choice of \( \tilde{\kappa}^{t}_{p+1} \):

\[
\tilde{\kappa}^{t}_{p+1} \geq \frac{p}{2} \left( \frac{6p}{\sqrt{5}(p+1)} \right)^p \frac{\alpha_{p+1}^{t}}{A_{t+1}} (2L_p + p\sigma). \tag{71}
\]

To sum up, by our choice of the parameters \( \tilde{\kappa}^{t}_i \), \( i = 2, \ldots, p \), we obtain from (67) that

\[
\psi^*_{t+1} \geq \frac{f(x_{t+1})}{A_t} + err_{t+1}.
\]
Thus, by induction, we obtain that, for all \( t \geq 1 \),
\[
\frac{f(x_{T+1})}{A_t} + \sum_{j=1}^{t} \alpha_j \langle G_{x_{j+1},1} - \nabla f(x_{j+1}), y_{j+1} - x_{j+1} \rangle \leq \psi_{t+1}^\ast \leq \psi_{t+1}(x^\ast)
\]

Lemma 5.1 \( f(x^\ast) \leq \frac{f(x^e)}{A_t} + \delta_1 \|x_1 - x_0\| + \delta_1 \|x^\ast - x_0\| + \sum_{i=2}^{p} \frac{2 \delta_i + \kappa_i^t}{(i-1)!} d_i(x^\ast - x_0) + \frac{L_p + p \sigma + \kappa_{p+1}^T}{p!} d_{p+1}(x^\ast - x_0) + \sum_{j=1}^{t} \alpha_j \langle G_{x_{j+1},1} - \nabla f(x_{j+1}), x^\ast - x_{j+1} \rangle
\]

where the first inequality has been just proved by induction and the third one comes from Lemma 5.1.

We now bound the gradient error terms in this inequality:
\[
\sum_{j=1}^{t} \alpha_j \langle G_{x_{j+1},1} - \nabla f(x_{j+1}), x^\ast - x_{j+1} \rangle - \sum_{j=1}^{t} \alpha_j \langle G_{x_{j+1},1} - \nabla f(x_{j+1}), y_{j+1} - x_{j+1} \rangle = \sum_{j=1}^{t} \alpha_j \langle G_{x_{j+1},1} - \nabla f(x_{j+1}), x^\ast - y_{j+1} \rangle \leq \delta_1 \sum_{j=1}^{t} \alpha_j \|x^\ast - y_{j+1}\| \leq \delta_1 \sum_{j=1}^{T} \frac{\alpha_j}{A_j}.
\]

Finally, we obtain the following convergence rate bound
\[
f(x_T) - f(x^\ast) \leq \frac{f(x^e)}{A_T} \left( 2 \delta_1 \bar{R} + \sum_{i=2}^{p} \frac{2 \delta_i + \kappa_i^T}{i!} \bar{R}^i + \frac{L_p + p \sigma + \kappa_{p+1}^T}{(p+1)!} \bar{R}^{p+1} + \delta_1 \bar{R} \sum_{j=1}^{T} \frac{\alpha_j}{A_j} \right)
\]
\[
= \frac{f(x^e)}{A_T} \left( 2 \delta_1 \bar{R} + \sum_{i=2}^{p} 2 \delta_i + \frac{p}{2} \left( \frac{6(i-1)\delta}{\sqrt{5}} \right)^{i-1}(\eta_i + 2) \frac{\alpha_{T+1}^T}{A_{T+1}} \delta_i \bar{R}^i \right) + \frac{L_p + p \sigma + \frac{1}{2} \left( \frac{6p}{\sqrt{5}} \right)^i (p+1)! (2L_p + p \sigma) \alpha_{T+1}^T}{(p+1)!} \bar{R}^{p+1} + \delta_1 \bar{R} \sum_{j=1}^{T} \frac{\alpha_j}{A_j}
\]
\[
\leq \sum_{i=2}^{p} \frac{p}{i!} \left( \frac{6(i-1)\delta}{\sqrt{5}} \right)^{i-1} (p+1)^i (\eta_i + 2) \frac{\delta_i \bar{R}^i}{(T + p + 2)^i}
\]
\[
+ \frac{6p}{\sqrt{5}} \left( \frac{2L_p + p \sigma}{(p-1)!} \right) \bar{R}^{p+1} + 2 \delta_1 \bar{R}.
\]
Now, for the rest of the proof we assume that for some iteration $t \leq T$ we have:

$$\delta_1 \geq \sum_{i=2}^{p} \frac{\delta_i}{(i-1)!} \|x_{t+1} - v_t\|^{i-1} + \frac{L_p}{p!} \|x_{t+1} - v_t\|^p.$$

Then, by Lemma 5.2 we have

$$f(x_{t+1}) - f(x^*) \leq \left( \max \left\{ \frac{p\sigma}{L_p}, \max_i \{ \eta_i \} \right\} + 2 \right) \delta_1 \|x_{t+1} - x^*\|.$$

Thus, we obtain the second statement of the theorem.

In Theorem 5.1 has two cases. The first one provides an upper bound on the objective residual after $T$ iterations (64). The second one shows, that there can be the case, when on iteration $t$ we have the objective residual smaller than $\delta_1\bar{R}$ already (66). Note, that this condition may be verified on every step of Algorithm 2.

The last term in the RHS of (64) and (65) corresponds to the case of an Accelerated Tensor Method, i.e. $\delta_i = 0, i = 1, \ldots, p$ provides similar convergence rate as [44]. As in Inexact Tensor Method other terms correspond to the inexactness in derivatives. For $i = 2, \ldots, p$ we have additional terms proportional to $\frac{\delta_1}{p!}$, which are better than ones in non-accelerated case $\left( \frac{\delta_1}{p!} \right)$. Note, that if the gradient error $\delta_1$ is non-zero, our bound does not guarantee the decrease of the objective below $O(\delta_1\bar{R})$. In the case of exact gradient (i.e. $\delta_1 = 0$) we do not need additional Assumption 5.1.

Assuming that we can control the errors in derivatives, we show how small one should take the errors to make the objective residual smaller than $\varepsilon$.

**Corollary 5.1:** Let assumptions of Theorem 5.1 hold and let $\varepsilon > 0$ be the desired solution accuracy. Further, let the levels of inexactness in Condition 2.1 satisfy the following inequalities:

$$\delta_1 \leq \min \left\{ \frac{\varepsilon}{2(p+1)\bar{R}}, \frac{\varepsilon}{\max \left\{ \frac{p\sigma}{L_p}, \max_i \{ \eta_i \} \right\} + 2} \right\} = O \left( \frac{\varepsilon}{\bar{R}} \right), \quad (73)$$

$$\delta_i \leq \varepsilon^{\frac{p+1-i}{p+1}} \frac{6^{p+1-i}}{\sqrt{5}^{p+1}} \frac{p+1-i}{p+1} \left( \frac{\eta_i + 2}{p+1} \right)^{i-1} \left( \frac{p+1}{p+1-i-1} \right)^{i-1} (\eta_i + 2) = O \left( \varepsilon^{\frac{p+1-i}{p+1}} \right), \quad i = 2, \ldots, p. \quad (74)$$

Also, let the number of iterations $T$ of Algorithm 1 satisfy

$$T = \max \left\{ 1; \frac{\bar{R}}{\varepsilon^{\frac{1}{p+1}}} \left( \frac{6p}{\sqrt{5}} \right)^{\frac{1}{p+1}} \left( \frac{p+1}{p+1-i} \right)^{\frac{1}{p+1}} \left( \frac{2L_p + p\sigma}{\bar{R}} \right)^{\frac{1}{p+1}} - p - 1 \right\}. \quad (75)$$

Then $x_t$ is an $\varepsilon$-solution of the problem (1), i.e. $f(x_t) - f(x^*) \leq \varepsilon$ for some $t = 1, \ldots, T.$
**Proof:** Firstly, let us consider the case in Theorem 5.1 when

\[
 f(x_t) - f(x^*) \leq \left( \max_i \left\{ \frac{p\sigma}{L_{p_i}}; \max_i \{\eta_i\} \right\} + 2 \right) \delta_1 \bar{R}.
\]

By our choice of \(\delta_1\) we obtain

\[
 f(x_t) - f(x^*) \leq \epsilon.
\]

Secondly, we consider the case, when convergence rate is given by (64). Then, by our choice of \(\delta_i\) and \(T\)

\[
 2\delta_1 \bar{R} \leq \frac{\epsilon}{p+1},
\]

\[
 p \left( \frac{6(i-1)}{\sqrt{5i}} \right)^{i-1} \frac{(p+1)^i(p+2)}{i!} \delta_i R^{i+1} \leq \frac{\epsilon}{p+1},
\]

\[
 \frac{\left( \frac{6p}{\sqrt{5}} \right)^p}{p!} \left( 2L_p + p\sigma \right) R^{p+1} \leq \frac{\epsilon}{p+1},
\]

and, from (64), we have

\[
 f(x_{T+1}) - f(x^*) \leq \epsilon.
\]

Thus, the number of steps sufficient for the Accelerated Inexact Tensor Method to find an \(\epsilon\)-solution is the same as for accelerated exact tensor method [44] and is equal to

\[
 O \left( \left( \frac{L_{D_{p+1}}}{\epsilon} \right)^{\frac{1}{p+1}} \right)
\]

under the assumption that \(\delta_1 = O(\epsilon)\) and \(\delta_i = O\left( \frac{\epsilon}{p+1} \right)\) for \(i = 2, \ldots, p\). Note, that inexactness lever for \(i = 2, \ldots, p\) are less strict than in non-accelerated case \(O\left( \frac{\epsilon}{p+1} \right)\).

6. Implementation details

Inexact tensor optimization methods, introduced in Sections 4, 5, are based on the solution of auxiliary subproblems (23), (45) in each iteration. As we already proved, these problems are convex, and, therefore, can be solved by convex optimization methods. However, the complexity of solving these problems can slow down the convergence of inexact tensor methods. In this section, we show how to treat these problems in the particular case of \(p = 3\). To do that, we consider the third degree model which corresponds to \(p = 3\)

\[
 \zeta_x(h) \overset{\text{def}}{=} \omega_{x,3}(x + h) = \phi_{x,3}(x + h) + \frac{\delta_1}{2y} + \frac{\gamma \delta_1}{2} \|h\|^2 + \frac{\delta_2}{2} \|h\|^2 + \frac{\delta_3}{6} \|h\|^3 + \frac{\sigma}{8} \|h\|^4.
\]

We use the following notation for the first three approximate (sampled) derivatives \(g = G_{x,1}, B = G_{x,2}, T = G_{x,3}\), which leads to the following expression for the inexact Taylor polynomial:

\[
 \phi_x(h) \overset{\text{def}}{=} f(x) + g^\top h + \frac{1}{2} h^\top B h + \frac{1}{6} T[h]^3.
\]
Lemma 6.1: For any $h \in \mathbb{R}^n$ and $\tau > 0$, we have

$$-\frac{1}{\tau} B - \frac{1}{\tau} (\gamma \delta_1 + \delta_2) I - \delta_3 \|h\| I - \frac{\tau}{2} L_3 \|h\|^2 I \preceq T[h] \preceq \frac{1}{\tau} B + \frac{1}{\tau} (\gamma \delta_1 + \delta_2) I + \delta_3 \|h\| I + \frac{\tau}{2} L_3 \|h\|^2 I$$

(78)

Proof: From (18) and definition of $\xi_x(h)$ we obtain:

$$0 \leq \langle \nabla^2 f(x + h)u, u \rangle \leq \langle \nabla^2 \phi_x(h)u, u \rangle + \frac{L_3^2}{2} \|h\|^2 \|u\|^2 + \delta_2 \|u\|^2 + \delta_3 \|h\| \|u\|^2$$

$$\leq \langle (B + T[h])u, u \rangle + \frac{L_3^2}{2} \|h\|^2 \|u\|^2 + \delta_2 \|u\|^2 + \delta_3 \|h\| \|u\|^2$$

Replacing $h$ with $\tau h$ and dividing by $\tau$, we get

$$-\langle T[h]u, u \rangle \leq \frac{1}{\tau} \langle Bu, u \rangle + \frac{\tau}{2} L_3 \|h\|^2 \|u\|^2 + \frac{1}{\tau} \delta_2 \|u\|^2 + \delta_3 \|h\| \|u\|^2.$$  

Replacing $h$ by $-h$ we obtain:

$$\langle T[h]u, u \rangle \leq \frac{1}{\tau} \langle Bu, u \rangle + \frac{\tau}{2} L_3 \|h\|^2 \|u\|^2 + \frac{1}{\tau} \delta_2 \|u\|^2 + \delta_3 \|h\| \|u\|^2.$$  

From the last two inequalities we get (78).

In the considered case of $p = 3$, Algorithm 1 requires to solve the following minimization problem at each iteration:

$$\min_{h \in \mathbb{R}^n} \left\{ \xi_x(h) = \phi_x(h) + \frac{\delta_1}{2\gamma} + (\gamma \delta_1 + \delta_2) d_2(h) + \delta_3 d_3(h) + \frac{\sigma}{2} d_4(h) \right\}.$$  

(79)

For any $h \in \mathbb{R}^n$ :

$$\nabla^2 \xi_x(h) = \nabla^2 \phi_x(h) + (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h) + \delta_3 \nabla^2 d_3(h) + \frac{\sigma}{2} \nabla^2 d_4(h)$$

$$= B + T[h] + (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h) + \delta_3 \nabla^2 d_3(h) + \frac{\sigma}{2} \nabla^2 d_4(h)$$

(78)

$$\geq B + (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h) + \delta_3 \nabla^2 d_3(h) + \frac{\sigma}{2} \nabla^2 d_4(h)$$

$$- \frac{1}{\tau} B - \frac{1}{\tau} (\gamma \delta_1 + \delta_2) I - \delta_3 \|h\| I - \frac{\tau}{2} L_3 \|h\|^2 I$$

$$\geq \left( 1 - \frac{1}{\tau} \right) B + \left( 1 - \frac{1}{\tau} \right) (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h) + \left( 1 - \frac{1}{\tau} \right) \delta_3 \nabla^2 d_3(h)$$

$$+ \left( \frac{\sigma - \tau L_3^3}{2} \right) \nabla^2 d_4(h).$$  

(80)
On the other hand, for any $h \in \mathbb{R}^n$:

$$
\nabla^2 \xi_x(h) = \nabla^2 \phi_x(h) + (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h) + \delta_3 \nabla^2 d_3(h) + \frac{\sigma}{2} \nabla^2 d_4(h)
$$



\begin{align*}
\Rightarrow B + \frac{1}{\tau} B + \frac{1}{\tau} (\gamma \delta_1 + \delta_2) I + \delta_3 \|h\|^2 I \\
+ (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h) + \delta_3 \nabla^2 d_3(h) + \frac{\sigma}{2} \nabla^2 d_4(h)
\end{align*}

(78)

$$
\approx \left(1 + \frac{1}{\tau}\right) B + \left(1 + \frac{1}{\tau}\right) (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h)
+ \left(1 + \frac{1}{\tau}\right) \delta_3 \nabla^2 d_3(h) + \left(\frac{\sigma - \tau L_3}{2}\right) \nabla^2 d_4(h).
$$

(9)

(81)

Next, we set $\sigma = \tau^2 L_3$ with $\tau \geq 1$. In view of Theorem 3.1, model $\xi_x(h)$ is convex with this choice of $\sigma$.

Then, from (80), (81) we obtain

$$
\left(1 - \frac{1}{\tau}\right) B + \left(1 - \frac{1}{\tau}\right) (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h)
+ \left(1 - \frac{1}{\tau}\right) \delta_3 \nabla^2 d_3(h) + \left(\frac{\sigma - \tau L_3}{2}\right) \nabla^2 d_4(h)
\leq \nabla^2 \xi_x(h)
\leq \frac{1 + \tau}{1 - \tau} \left(\left(1 - \frac{1}{\tau}\right) B + \left(1 - \frac{1}{\tau}\right) (\gamma \delta_1 + \delta_2) \nabla^2 d_2(h)
+ \left(1 - \frac{1}{\tau}\right) \delta_3 \nabla^2 d_3(h) + \left(\frac{\sigma - \tau L_3}{2}\right) \nabla^2 d_4(h)\right).
$$

(82)

Let $\rho_x(h) = \frac{1}{2} \left(1 - \frac{1}{\tau}\right) (Bh, h) + (1 - \frac{1}{\tau}) (\gamma \delta_1 + \delta_2) d_2(h) + (1 - \frac{1}{\tau}) \delta_3 d_3(h) + \left(\frac{\sigma - \tau L_3}{2}\right) d_4(h)$.

Therefore, we have proved the relative strong convexity with constant $1$ and relative smoothness with constant $\kappa(\tau) = \frac{1 + \tau}{1 - \tau}$ of the function $\xi_x(h)$ with respect to the function $\rho_x(h)$ [38]:

$$
\nabla^2 \rho_x(h) \preceq \nabla^2 \xi_x(h) \preceq \frac{1 + \tau}{1 - \tau} \nabla^2 \rho_x(h).
$$

(83)

The relative smoothness condition allows to solve the auxiliary problem (79) very efficiently [38,44] by the iterates

$$
h_{k+1} = \arg \min_{h \in \mathbb{R}^n} \left\{ (\nabla \xi_x(h_k), h) + \kappa(\tau) \beta \rho_x(h_k, h) \right\}
$$

with linear rate of convergence.

Note, that in the accelerated method (Algorithm 2) subproblem (45) is different than minimizing the model $\xi_x(h)$ and has the form:

$$
\min_x \left\{ \phi_{x,3}(x + h) + \frac{\delta_2}{2} \|h\|^2 + \frac{\delta_3}{6} \|h\|^3 + \frac{\sigma}{8} \|h\|^4 \right\}.
$$

However, this objective function also satisfies relative strong convexity and relative smoothness conditions. Indeed, the term $\frac{\delta_1}{2\gamma}$ does not correspond to convexity or smoothness. Therefore, we can omit it and the analysis will be the same. We set $\gamma = 0$ in
Equation (76) and the reference function

\[
\rho_x(h) = \frac{1}{2} \left( 1 - \frac{1}{\tau} \right) \langle Bh, h \rangle + \left( 1 - \frac{1}{\tau} \right) \delta_2 d_2(h) + \left( 1 - \frac{1}{\tau} \right) \delta_3 d_3(h) + \left( \frac{\sigma - \tau L_3}{2} \right) d_4(h).
\]

Then, the objective in subproblem (45) also satisfies relative smoothness and strong convexity condition. Therefore, it can also be solved by the process (83).

According to [44] it is not necessary to calculate the full third derivative tensor \( T \) in the above derivations. It is sufficient to use an automatic differentiation technique to calculate third-order derivative in a certain direction.

7. Stochastic tensor and accelerated tensor methods

In this section, we apply Inexact Tensor Method and Accelerated Inexact Tensor Method to solve stochastic optimization problem in the online (2) and offline (3) settings. The main step to do that is to find sufficient conditions for Condition 2.1 to be satisfied in these two settings. To do that we first introduce an additional assumption on the objective function \( f \):

Assumption 7.1: The function \( f(x) \) and its derivatives \( \nabla f(x), \ldots, \nabla^{p-1} f(x) \) are Lipschitz continuous:

\[
\| \nabla^i f(x) - \nabla^i f(y) \| \leq L_i \| x - y \|, \quad i = 0, \ldots, p - 1.
\]

In the stochastic versions of Algorithms 1, 2 for the stochastic optimization problem (1) in the online (2) and offline (3) settings we sample stochastic derivatives in each iteration in order to form mini-batch approximations for the derivatives of \( f \).

More precisely, for \( S_1, S_2, \ldots, S_p \) being sample sets for each derivative, we set

\[
G_{x,i} = \frac{1}{|S_i|} \sum_{j \in S_i} \nabla^i f(x, \xi_j), \quad i = 1, \ldots, p.
\]

In the next subsections we show how to choose the size of the sample sets to satisfy Condition 2.1 for both online and offline settings.

7.1. Online setting

For the online setting, i.e. when (2) holds, we need one more assumption to be able to satisfy Condition 2.1.

Assumption 7.2: For all \( i = 1, 2, \ldots, p, \xi, \) and \( x \in \mathbb{R}^n \):

\[
\| \nabla^i f(x, \xi) - \nabla^i f(x) \| \leq M_i.
\]

From the following tensor concentration bound theorem we derive the required conditions.
Theorem 7.1 (Tensor Hoeffding Inequality [39]): Let $\mathcal{X}$ be a sum of $m$ i.i.d. tensors $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times \cdots \times d_k}$. Let $u_1, \ldots, u_k$ be such that $\|u_i\| = 1$ and assume that for each tensor $\mathcal{Y}_i$, it holds that $a \leq \mathcal{Y}_i(u_1, \ldots, u_k) \leq b$. Let $\sigma := (b - a)$. Then we have
\[
P(\|\mathcal{X} - \mathbb{E}\mathcal{X}\| \geq t) \leq k_0 \left( \frac{\sum_{i=1}^{k} d_i}{2} \right) \cdot 2 \exp \left( -\frac{t^2}{2m\sigma^2} \right)
\]
where $k_0 = \left( \frac{2^k}{\log(3/2)} \right)$.

This Theorem allows us to provide a sufficient condition on the sample sets $S_i$ in order to satisfy Condition 2.1. Note, that it is also possible to use Bernstein tensor inequality [40], which is sharper than Hoeffding. However, sample sizes estimates from these inequalities will differ only in logarithmic terms.

Lemma 7.1: Let Assumptions 2.1, 7.1, 7.2 be satisfied. Then, for any fixed small constants $\delta_i > 0$ we can choose the sizes of the sample sets $S_i$ in Equation (7) to be
\[
|S_i| = n_i = 2 \left( (L_{i-1} + M_i)^2 \cdot \delta_i^{-2} \right) \log \frac{2k_0\delta_i}{\delta_i}, \quad i = 1, \ldots, p
\]
so that with probability at least $1 - \delta$ Condition 2.1 is satisfied.

Proof: Using Assumptions 2.1, 7.2 and the triangle inequality, we obtain
\[
\|G_{x,i}\| = \frac{1}{n_i} \sum_{j=1}^{n_i} \|\nabla^i f(x, \xi_j)\| \leq \frac{1}{n_i} \sum_{j=1}^{n_i} \left( \|\nabla^i f(x, \xi_j) - \nabla^i f(x)\| + \|\nabla^i f(x)\| \right)
\]
\[
\leq M_i + L_{i-1} \overset{\text{def}}{=} \sigma_i.
\]
Then, the proof completely repeats the proof of Lemma 11 in [39]. We require the probability of a large deviation to be smaller than $\delta \in (0, 1]$: \[
P \left\{ \|G_{x,i} - \nabla^i f(x)\| > t \right\} \leq k_0^{n_i} \cdot 2 \exp \left( -\frac{t^2n_i}{2\sigma_i^2} \right) \leq \delta.
\]
Taking the logarithm of both sides, we get
\[
-\frac{t^2n_i}{2\sigma_i^2} \leq \log \frac{\delta}{2k_0^{n_i}}
\]
which is equivalent to
\[
n_i \geq \frac{2\sigma_i^2}{t^2} \log \frac{2k_0^{n_i}}{\delta}.
\]
Finally, we can simply choose $t = \delta_i$ in order to satisfy (6) in Condition 2.1 since then, with probability at least $1 - \delta$, $\forall y \in \mathbb{R}^n$,
\[
\|G_{x,i}[y - x]|^{i-1} - \nabla^i f(x)[y - x]|^{i-1}\| \leq \|G_{x,i} - \nabla^i f(x)\| \|y - x||^{i-1} \leq \delta_i \|y - x||^{i-1}.
\]
Next, we obtain the required batch sizes for our stochastic tensor methods in order to achieve convergence rates as in the case of exact derivatives. The following result immediately follows from the Lemma above and Corollary 4.1.

**Corollary 7.1:** Let Assumptions 2.1, 7.1, 7.2 be satisfied. Then, we can choose the sizes \( |S_i| \) of sample sets \( S_i \) in (7) to be

\[
n_i = |S_i| = 2 \left( (L_{i-1} + M_i)^2 \cdot \varepsilon^{-2p+1} \right) \log \frac{2k_0^{ni}}{\delta \varepsilon^{1/p}}, \quad i = 1 \ldots p
\]

so that with probability at least \( 1 - \delta \) Stochastic Tensor Method has complexity \( T = O(\frac{1}{\varepsilon^{1/p}}) \).

A similar result can be obtained for the Accelerated Stochastic Tensor Method.

**Corollary 7.2:** Let Assumptions 2.1, 7.1, 7.2 be satisfied. Then, we can choose the sizes \( |S_i| \) of sample sets \( S_i \) in (7) to be

\[
n_1 = |S_1| = 2 \left( (L_0 + M_1)^2 \cdot \varepsilon^{-2} \right) \log \frac{2k_0^{n_1}}{\delta \varepsilon^{1/(p+1)}}, \quad (84)
\]

\[
n_i = |S_i| = 2 \left( (L_{i-1} + M_i)^2 \cdot \varepsilon^{-2p+i} \right) \log \frac{2k_0^{n_i}}{\delta \varepsilon^{1/(p+1)}}, \quad i = 2, \ldots, p \quad (85)
\]

so that with probability at least \( 1 - \delta \) Accelerated Stochastic Tensor Method has complexity \( T = O(\frac{1}{\varepsilon^{1/(p+1)}}) \).

Corollaries 7.1, 7.2 show how to choose batch sizes for derivatives in inexact tensor methods to achieve convergence in exact cases for online setting. Gradient sample sizes of non-accelerated and accelerated tensor methods are the same. For the case of higher derivatives (of order \( \geq 2 \)), we need to use fewer samples for the accelerated method than for the non-accelerated one. Moreover, batch sizes are decreasing with the growth of derivatives order.

### 7.2. Offline setting

For the offline setting, i.e. when (3) holds, we use the following result to provide a sufficient condition for Condition 2.1 to hold.

**Theorem 7.2 (Tensor Hoeffding–Serfling Inequality [39]):** Let \( \mathcal{X} \) be a sum of \( m \) tensors \( \mathcal{Y}_i \in \mathbb{R}^{d_1} \times \cdots \times d_k \), sampled without replacement from a finite population \( \mathcal{A} \) of size \( N \). Let \( u_1, \ldots, u_k \) be such that \( \|u_i\| = 1 \) and assume that for each tensor \( i \), \( a \leq \mathcal{Y}_i(u_1, \ldots, u_k) \leq b \). Let \( \sigma := (b - a) \), then we have

\[
P(\| \mathcal{X} - \mathbb{E} \mathcal{X} \| \geq t) \leq k_0^{\sum_{i=1}^k d_i} \cdot 2 \exp \left( - \frac{t^2 m^2}{2 \sigma^2 (m + 1)(1 - m/N)} \right), \quad (86)
\]

where \( k_0 = \frac{2k}{\log(3/2)} \).

The next lemma follows from Theorem 7.2. The proof can be found in [39].
Lemma 7.2: Let Assumptions 2.1, 7.1 be satisfied. Then, for any fixed small constants $\delta_i > 0$ we can choose the sizes $|S_i|$ of sample sets $S_i$ in (7) to be

$$n_i = |S_i| = \frac{4L_i^2 \log \frac{2\kappa n_i}{\delta}}{\delta_i^2 + \frac{2L_i^2}{N} \log \frac{\delta}{2\kappa N}} = \tilde{O} \left( \frac{\delta_i^2}{L_i^2} + \frac{1}{m} \right)^{-1}, \quad i = 1, \ldots, n,$$

where $\tilde{O}$ hides logarithmic dependencies, so that with probability at least $1 - \delta$ Condition 2.1 holds.

Similar to the previous subsection we obtain the following results.

Corollary 7.3: Let Assumptions 2.1, 7.1 be satisfied. Then, we can choose the sizes $|S_i|$ of sample sets $S_i$ in (7) to be

$$n_i = |S_i| = \tilde{O} \left( \frac{\epsilon^{2p-i+1}}{L_i^2} + \frac{1}{m} \right)^{-1}, \quad i = 1 \ldots n$$

so that with probability at least $1 - \delta$ Stochastic Tensor Method has complexity $T = O(\frac{1}{\epsilon^{1/p}})$.

Corollary 7.4: Let Assumptions 2.1, 7.1 be satisfied. Then, we can choose the sizes $|S_i|$ of sample sets $S_i$ in (7) to be

$$n_1 = |S_1| = \tilde{O} \left( \frac{\epsilon}{L_0^2} + \frac{1}{m} \right)^{-1},$$
$$n_i = |S_i| = \tilde{O} \left( \frac{\epsilon^{2p-i+1}}{L_i^2} + \frac{1}{m} \right)^{-1}, \quad i = 2 \ldots n$$

so that with probability at least $1 - \delta$ Accelerated Stochastic Tensor Method has complexity $T = O(\frac{1}{\epsilon^{1/(p+1)}})$.

Corollaries 7.3, 7.4 are the counterpart of Corollaries 7.1, 7.2 and shows the analogical result off the line setting.

At the end of this section, let us discuss an example of the function that satisfies Assumptions 2.1 and 7.1. We will estimate the Lipshitz constant $L_i$ for the $i$th derivative, where $i = 0, 1, 2$. These constants can be used to calculate estimates in Corollaries 7.1, 7.2, 7.3, 7.4. We consider a classical convex machine learning problem, logistic regression.

Example 7.1: For a given $x_j \in \mathbb{R}^n$, $y_j \in \{-1; 1\}$, where $j \in \{1, \ldots, m\}$, we consider the following logistic regression loss function:

$$f(w) = \frac{1}{m} \sum_{j=1}^{m} \log(1 + \exp(-y_j w^\top x_j)).$$

Let us fix arbitrary point $w \in \mathbb{R}^n$ and direction $h \in \mathbb{R}^d$, then straightforward computations give the next upper-bounds for Lipschitz constants. For the Lipschitz-continuous function
constant $L_0$, we get

$$\nabla f(w) = -\frac{1}{m} \sum_{j=1}^{m} y_j \exp (-y_j w^\top x_j) x_i,$$

$$\|\nabla f(w)\| \leq \frac{1}{m} \sum_{i=1}^{m} \|x_i\| = L_0.$$

For the Lipschitz-continuous gradient constant $L_1$, we get

$$(\nabla^2 f(w) h, h) = \frac{1}{m} \sum_{j=1}^{m} \frac{\exp (-y_j w^\top x_j)}{(1 + \exp (-y_j w^\top x_j))^2} (x_j, h)^2 = \frac{1}{m} \sum_{j=1}^{m} \chi_1(y_j w^\top x_j) (x_j, h)^2,$$

$$\|\nabla^2 f(w)\| = \max_{\|h\| \leq 1} |(\nabla^2 f(x) h, h)| \leq \frac{1}{4} \lambda_{\text{max}} \left( \frac{1}{m} \sum_{j=1}^{m} x_j x_j^\top \right),$$

$$\leq \frac{1}{4} \left( \frac{1}{m} \sum_{j=1}^{m} \|x_j\|^2 \right) = L_1,$$

where in the first inequality we use the next upper-bound

$$\chi_1(t) = \frac{\exp (-t)}{(1 + \exp (-t))^2} \leq \frac{1}{4}.$$

For the Lipschitz-continuous Hessian constant $L_2$, we get

$$\nabla^3 f(w)[h]^3 = -\frac{1}{m} \sum_{j=1}^{m} y_j \exp (-y_j w^\top x_j) (\exp (-y_j w^\top x_j) - 1) (x_j, h)^3$$

$$= -\frac{1}{m} \sum_{j=1}^{m} y_j \chi_2(y_j w^\top x_j) (x_j, h)^3,$$

$$\|\nabla^3 f(w)\| = \max_{\|h\| \leq 1} |\nabla^3 f(w)[h]^3| \leq \frac{1}{6 \sqrt{3}} \left( \frac{1}{m} \sum_{j=1}^{m} \|x_j\|^3 \right) = L_2,$$

where in the first inequality we use the next upper-bound

$$\chi_2(t) = \frac{\exp (-t) \left( \exp (-t) \right)}{(1 + \exp (-t))^3} \leq \frac{1}{6 \sqrt{3}}.$$

If we normalize data vectors $x_i$ prior to learning, then we get the next constants

$$L_0 = 1, \quad L_1 = 0.25, \quad L_2 = \frac{1}{6 \sqrt{3}}.$$
Acknowledgements

The work of A. Agafonov, and D. Kamzolov was supported by a grant for research centres in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

The authors are very grateful to Yu.E. Nesterov, A.I. Golikov, Yu.G. Evtushenko and G. Scutari for fruitful discussions.

Disclosure statement

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

Notes on contributors

Artem Agafonov got B.Sc. and M.Sc. degrees from the Department of Control and Applied Mathematics of Moscow Institute of Physics and Technology (MIPT) in 2020 and 2022, respectively. Since 2022 he has been pursuing a Ph.D. in MBZUAI.

Dmitry Kamzolov is a Research Associate at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. Dmitry earned his B.Sc. and M.Sc. degrees from the Department of Control and Applied Mathematics at the Moscow Institute of Physics and Technology (MIPT) in 2016 and 2018, respectively. He also obtained the M.Sc. in Operations Research, Combinatorics and Optimization (ORCO) at the University of Grenoble Alpes (UGA). In 2020, he obtained his Ph.D. in Mathematical Modelling, Numerical Methods and Program Complexes from MIPT. Since 2021, he has been working as a Research Associate in the group led by Martin Takáč at MBZUAI, UAE. His current research interests focus on second-order and high-order optimization methods for both convex and non-convex problems, with application in machine learning and distributed computations.

Pavel Dvurechensky got B.Sc and M.Sc degrees from the Department of Control and Applied Mathematics of Moscow Institute of Physics and Technology in 2008 and 2010 respectively. He obtained a PhD from the same university in 2014. Since 2015 he has been a research associate at Weierstrass Institute for Applied Analysis and Stochastics in Berlin.

Alexander Gasnikov obtained B.Sc and M.Sc degrees from the Department of Control and Applied Mathematics at the Moscow Institute of Physics and Technology in 2004 and 2006, respectively. He earned a PhD from the same university in 2007 and a Doctor’s degree (Habilitation) in 2016, under the supervision of A. Shanin and Yu. Nesterov. Since 2020, he has held the position of professor at the Moscow Institute of Physics and Technology and the Higher School of Economics. Since 2021, he has led a scientific group at the Ivanikov Institute for System Programming RAS. Starting in 2022, he has been the head of the Laboratory of Mathematical Methods of Optimization and the head of the Department of Mathematical Foundations of Control. As of 2023, he has been appointed as the head of the Laboratory of Mathematical Foundations of Machine Learning at the Institute of Information Transmission Problems RAS and as the head of the Laboratory for Multi-scale Neurodynamics for Smart Systems at Skoltech. He has received awards from Yahoo, Yandex, the Moscow government, and Huawei. He is the author of more than 200 papers, including publications in Q1 journals (JOTA, SIAM, EJOR, OMS, etc.) and A* conferences (NeurIPS, ICML, COLT).

Martin Takáč is an Associate Professor and Deputy Department Chair of Machine Learning Department at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE. Before joining MBZUAI, he was an Associate Professor in the Department of Industrial and Systems Engineering at Lehigh University, where he has been employed since 2014. He received his B.S. (2008) and M.S. (2010) degrees in Mathematics from Comenius University, Slovakia, and Ph.D. (2014) degree
in Mathematics from the University of Edinburgh, United Kingdom. His current research interests include designing and analyzing algorithms for machine learning, AI for science, understanding protein-DNA interactions, and using ML for energy.

**ORCID**

Dmitry Kamzolov http://orcid.org/0000-0001-8488-9692
Pavel Dvurechensky http://orcid.org/0000-0003-1201-2343
Martin Takáč http://orcid.org/0000-0001-7455-2025

**References**

[1] A. Agafonov, D. Kamzolov, P. Dvurechensky, and A. Gasnikov, *Inexact tensor methods and their application to stochastic convex optimization*, preprint 2020. arXiv:2012.15636.

[2] N. Agarwal and E. Hazan, *Lower bounds for higher-order convex optimization*, preprint 2017. arXiv:1710.10329.

[3] Y. Arjevani, O. Shamir, and R. Shiff, *Oracle complexity of second-order methods for smooth convex optimization*, Math. Program. 178(1-2) (2019), pp. 327–360.

[4] M. Baes, *Estimate sequence methods: extensions and approximations*. Institute for Operations Research, ETH, Zürich, Switzerland, 2009.

[5] S. Bellavia and G. Gurioli, *Stochastic analysis of an adaptive cubic regularization method under inexact gradient evaluations and dynamic hessian accuracy*, Optimization 71(1) (2022), pp. 227–261.

[6] S. Bellavia, G. Gurioli, and B. Morini, *Adaptive regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization*, IMA J. Numerical Anal. 41(1) (2020), pp. 764–799.

[7] S. Bellavia, G. Gurioli, B. Morini, and P.L. Toint, *Adaptive regularization algorithms with inexact evaluations for nonconvex optimization*, SIAM. J. Optim. 29(4) (2019), pp. 2881–2915.

[8] S. Bellavia, G. Gurioli, B. Morini, and P.L. Toint, *Adaptive regularization for nonconvex optimization using inexact function values and randomly perturbed derivatives*, J. Complex. 68 (2022), p. 101591.

[9] S. Bubeck, Q. Jiang, Y.T. Lee, Y. Li, and A. Sidford, *Near-optimal method for highly smooth convex optimization*, in *Conference on Learning Theory*, Phoenix, USA, PMLR, 2019, pp. 492–507.

[10] C. Cartis, N.I. Gould, and P.L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results*, Math. Program. 127(2) (2011), pp. 245–295.

[11] C. Cartis, N.I.M. Gould, and P.L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity*, Math. Program. 130(2) (2011), pp. 295–319.

[12] C. Cartis, N.I. Gould, and P.L. Toint, *Universal regularization methods: varying the power, the smoothness and the accuracy*, SIAM. J. Optim. 29(1) (2019), pp. 595–615.

[13] C. Cartis, N.I. Gould, and P.L. Toint, *Improved second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*, preprint 2017. arXiv:1708.04044.

[14] C. Cartis and K. Scheinberg, *Global convergence rate analysis of unconstrained optimization methods based on probabilistic models*, Math. Program. 169 (2018), pp. 337–375.

[15] P.L. Chebyshev, *Collected Works [in Russian]*, Vol. 5, Izd. Akad. NaukSSSR, Moscow-Leningrad, 1951.

[16] M. Cohen, J. Diakonikolas, and L. Orecchia, *On acceleration with noise-corrupted gradients*, in *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, J. Dy and A. Krause, eds., PMLR, Stockholmsmässan, Stockholm, 2018, pp. 1019–1028.
[17] A. Conn, N. Gould, and P. Toint, Trust Region Methods, Society for Industrial and Applied Mathematics, 2000. https://doi.org/10.1137/1.9780898719857.
[18] A. d’Aspremont, Smooth optimization with approximate gradient, SIAM. J. Optim. 19(3) (2008), pp. 1171–1183.
[19] O. Devolder, Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization, Ph.D. diss., ICTEAM and CORE, Université Catholique de Louvain, 2013.
[20] N. Doikov and Y. Nesterov, Convex optimization based on global lower second-order models, in Advances in Neural Information Processing Systems 33 proceedings (NeurIPS 2020), Curran Associates, Inc., 2020.
[21] D. Dvinskikh and A. Gasnikov, Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems, J. Inverse Ill Posed Problems 29(3) (2021), pp. 385–405.
[22] P. Dvurechensky and A. Gasnikov, Stochastic intermediate gradient method for convex problems with stochastic inexact oracle, J. Optim. Theory. Appl. 171(1) (2016), pp. 121–145.
[23] Y.G. Evtushenko, Optimization and Fast Automatic Differentiation, Computing Center of RAS, Moscow, 2013.
[24] A. Gasnikov, D. Dvinskikh, P. Dvurechensky, D. Kamzolov, D. Pasechnyky, V. Matykhin, N. Tupitsa, and A. Chernov, Accelerated meta-algorithm for convex optimization, Comput. Math. Math. Phys. 61(1) (2020), pp. 17–28.
[25] A.V. Gasnikov and P.E. Dvurechensky, Stochastic intermediate gradient method for convex optimization problems, Doklady Math. 93(2) (2016), pp. 148–151.
[26] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, C.A. Uribe, B. Jiang, H. Wang, S. Zhang, S. Bubeck, Q. Jiang, Y.T. Lee, Y. Li, and A. Sidford, Near optimal methods for minimizing convex functions with lipschitz pth derivatives, in Proceedings of the Thirty-Second Conference on Learning Theory, volume 99 of Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, eds., PMLR, Phoenix, 2019, pp. 1392–1393.
[27] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and C.A. Uribe, Optimal tensor methods in smooth convex and uniformly convex optimization, in Proceedings of the Thirty-Second Conference on Learning Theory, volume 99 of Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, eds., PMLR, Phoenix, 2019, pp. 1374–1391.
[28] A. Gasnikov, E. Gorbunov, D. Kovalev, A. Mohammed, and E. Chernousova, The global rate of convergence for optimal tensor methods in smooth convex optimization, Computer Res. Modell. 10(6) (2018), pp. 737–753.
[29] A. Gasnikov, E. Gorbunov, D. Kovalev, A. Mokhammed, and E. Chernousova, Reachability of optimal convergence rate estimates for high-order numerical convex optimization methods, in Doklady Mathematics, Vol. 99, Moscow, Russia, Springer, 2019, pp. 91–94.
[30] A. Gasnikov, Universal gradient descent, Modern numerical optimization methods. MCCME, 2020.
[31] S. Ghadimi, H. Liu, and T. Zhang, Second-order methods with cubic regularization under inexact information, 2017. arXiv:1710.05782.
[32] F. Hanzely, N. Doikov, Y. Nesterov, and P. Richtarik, Stochastic subspace cubic newton method, in International Conference on Machine Learning, PMLR, 2020, pp. 4027–4038.
[33] K.H. Hoffmann and H.J. Kornstaedt, Higher-order necessary conditions in abstract mathematical programming, J. Optim. Theory. Appl. 26(4) (1978), pp. 533–568.
[34] B. Jiang, H. Wang, and S. Zhang, An optimal high-order tensor method for convex optimization, in Conference on Learning Theory, Phoenix, USA, PMLR, 2019, pp. 1799–1801.
[35] D. Kamzolov and A. Gasnikov, Near-optimal hyperfast second-order method for convex optimization and its sliding, preprint 2020. arXiv:2002.09050.
[36] J.M. Kohler and A. Lucchi, Sub-sampled cubic regularization for non-convex optimization, in International Conference on Machine Learning, Sydney, Australia, PMLR, 2017, pp. 1895–1904.
[37] D. Kovalev, K. Mishchenko, and P. Richtárik, Stochastic newton and cubic newton methods with simple local linear-quadratic rates, preprint 2019. arXiv:1912.01597.
[38] H. Lu, R.M. Freund, and Y. Nesterov, Relatively smooth convex optimization by first-order methods, and applications, SIAM. J. Optim. 28(1) (2018), pp. 333–354.
[39] A. Lucchi and J. Kohler, *A sub-sampled tensor method for nonconvex optimization*, IMA J. Numerical Anal. 43(3) (2023), pp. 2856–2891.

[40] Z. Luo, L. Qi, and P.L. Toint, *Tensor bernstein concentration inequalities with an application to sample estimators for high-order moments*, Front. Math. China 15 (2020), pp. 367–384.

[41] R.D. Monteiro and B.F. Svaiter, *An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods*, SIAM. J. Optim. 23(2) (2013), pp. 1092–1125.

[42] Y. Nesterov, *Accelerating the cubic regularization of newton’s method on convex problems*, Math. Program. 112(1) (2008), pp. 159–181.

[43] Y. Nesterov, *Lectures on Convex Optimization*, Vol. 137. Springer, 2018. https://doi.org/10.1007/978-3-319-91578-4.

[44] Y. Nesterov, *Implementable tensor methods in unconstrained convex optimization*, Math. Program. 186 (2019), pp. 157–183.

[45] Y. Nesterov, *Inexact high-order proximal-point methods with auxiliary search procedure*, SIAM J. Optim. 31 (2021), pp. 2807–2828.

[46] Y. Nesterov, *Superfast second-order methods for unconstrained convex optimization*, J. Optim. Theory. Appl. 191 (2021), pp. 1–30.

[47] Y. Nesterov and B. Polyak, *Cubic regularization of newton method and its global performance*, Math. Program. 108(1) (2006), pp. 177–205.

[48] S. Park, S.H. Jung, and P.M. Pardalos, *Combining stochastic adaptive cubic regularization with negative curvature for nonconvex optimization*, J. Optim. Theory. Appl. 184(3) (2020), pp. 953–971.

[49] B.T. Polyak, *Introduction to Optimization. Optimization Software*, Vol. 1, Publications Division Inc., New York, 1987.

[50] A. Rodomanov and D. Kropotov, *A superlinearly-convergent proximal newton-type method for the optimization of finite sums*, in *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, M. F. Balcan and K. Q. Weinberger, eds., PMLR, New York, NY, 2016, pp. 2597–2605.

[51] F. Stonyakin, A. Tyurin, A. Gasnikov, P. Dvurechensky, A. Agafonov, D. Dvinskikh, M. Alkousa, D. Pasechnyuk, S. Artamonov, and V. Piskunova, *Inexact model: a framework for optimization and variational inequalities*. Optimization Methods and Software 36(6) (2021), pp. 1155–1201. 10.1080/10556788.2021.1924714.

[52] L.N. Trefethen, *Approximation Theory and Approximation Practice*, Vol. 164, SIAM, 2019. https://doi.org/10.1137/1.9781611975949.

[53] N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M.I. Jordan, *Stochastic cubic regularization for fast nonconvex optimization*, in *Advances in Neural Information Processing Systems*, Vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Montreal, Canada., Curran Associates, Inc., 2018, pp. 2899–2908.

[54] Z. Wang, Y. Zhou, Y. Liang, and G. Lan, *A note on inexact condition for cubic regularized newton’s method*, preprint 2018. arXiv:1808.07384v1.

[55] Z. Wang, Y. Zhou, Y. Liang, and G. Lan, *Cubic regularization with momentum for nonconvex optimization*, in *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, R. P. Adams and V. Gogate, eds., PMLR, Tel Aviv, 2020, pp. 313–322.

[56] P. Xu, F. Roosta, and M.W. Mahoney, *Newton-type methods for non-convex optimization under inexact hessian information*, Math. Program. 184(1-2) (2020), pp. 35–70.

[57] Y. Zhang and L. Xiao, *Disco: distributed optimization for self-concordant empirical loss*, in *International Conference on Machine Learning*, PMLR, Lille, France, 2015, pp. 362–370.

[58] D. Zhou, P. Xu, and Q. Gu, *Stochastic variance-reduced cubic regularized Newton methods*, in *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, J. Dy and A. Krause, eds., PMLR, Stockholmsmössan, Stockholm, 2018, pp. 5990–5999.