Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Inexact tensor methods and their application to stochastic convex optimization

Artem Agafonov[1], Dmitry Kamzolov[1], Pavel Dvurechensky[2], Alexander Gasnikov[1,2,3]

submitted: March 1, 2021

[1] Moscow Institute of Physics and Technology
Institutskiy Pereulok, 9
Dolgoprudny
Moscow Region 141701
Russian Federation
E-Mail: agafonov.ad@phystech.edu
dkamzolov@yandex.ru

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: pavel.dvurechensky@wias-berlin.de
alexander.gasnikov@wias-berlin.de

[3] Institute for Information Transmission Problems of RAS
Bolshoy Karetny per. 19, build.1
127051 Moscow
Russian Federation
E-Mail: gasnikov@yandex.ru

No. 2818
Berlin 2021

2020 *Mathematics Subject Classification.* 90C30, 90C25, 68Q25.

Key words and phrases. High-order methods, tensor methods, convex optimization, inexact derivatives, stochastic optimization.

The work of A. Agafonov and D. Kamzolov was fulfilled in Sirius (Sochi) in August 2020. The research of D. Kamzolov was partially supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, project no. 0714-2020-0005. The research of A. Gasnikov was partially supported by RFBR, project number 19-31-51001. The work of A. Agafonov was supported by Andrei M. Raigorodskii Scholarship in Optimization.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Inexact tensor methods and their application to stochastic convex optimization

Artem Agafonov, Dmitry Kamzolov, Pavel Dvurechensky, Alexander Gasnikov

Abstract

We propose a general non-accelerated tensor method under inexact information on higher-order derivatives, analyze its convergence rate, and provide sufficient conditions for this method to have similar complexity as the exact tensor method. As a corollary, we propose the first stochastic tensor method for convex optimization and obtain sufficient mini-batch sizes for each derivative.

1 Introduction

The idea of using the derivatives of the order p higher than two in optimization methods is known at least since 1970's [33]. In numerical analysis it was used much earlier, see the student work of P.L. Chebyshev [12] and more recent reviews [21, 59]. Despite theoretical advantages, the practical application of such tensor methods was limited until recent work [43] 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 [43] if the minimization problem is convex, then in each iteration of the tensor method one needs 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 [48]. This leads to a method with faster convergence than that of the accelerated Cubic regularized Newton's method with approximately the same cost of one iteration. These ideas were further developed to obtain accelerated versions of tensor methods [25, 46] and very fast second-order methods [47, 35, 46].

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 [63, 51]. A more general setting, which contains the finite-sum setting as a special case, is the setting of general stochastic optimization, for which second-order methods are also developed in the literature [60]. Tensor methods for such problems are also developed, but with the focus on non-convex problems [6, 39]. Thus, motivated by the lack of results on tensor methods for stochastic convex optimization, in this paper we study stochastic convex optimization problems, 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 analyze such methods under a particular condition for 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 method for convex optimization.

1.1 Problem Statement

We consider convex optimization problems of the following form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \tag{1}$$

DOI 10.20347/WIAS.PREPRINT.2818 Berlin 2021

under inexact information on its derivatives. 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 $f(\mathbf{x})$ is given as

$$f(\mathbf{x}) := \mathbf{E}_{\xi\sim\mathcal{D}}[f(\mathbf{x};\xi)], \tag{2}$$

where the random variable ξ is sampled from distribution $\mathcal{D}$ and an optimization procedure has access only to stochastic realizations of the function $f(\mathbf{x};\xi)$ via its high-order derivatives. In the *offline* setting the function f has *finite-sum* form:

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \tag{3}$$

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 ξ to be uniformly distributed over $i = 1, ..., m$. At the same time, we distinguish these two settings since the analysis of the proposed methods in the second setting can be made under different 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 methods originating from [48]. This line of works includes the development of accelerated methods [41, 40, 42], extensions of trust region methods [14, 9, 8, 10, 11], methods with inexact evaluations of gradients and Hessians [29, 61, 5, 18] with application to stochastic optimization in the online and offline settings. Stochastic second-order methods for convex optimization [51, 37, 31] and non-convex optimization [60, 65, 62, 49] 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 which use the derivative of the order p higher than 2, we refer to them as tensor methods. The idea of such methods was proposed quite long ago [33], and accelerated methods for convex optimization were also proposed in [4]. Recent interest to this type of methods in convex optimization was probably motivated by the lower bounds obtained in [3, 1]. In [43] it was shown that appropriately regularized Taylor expansion of a convex function is convex, which leads to implementability of such methods which minimize this regularized Taylor expansion. Accelerated tensor methods were also proposed, yet with a remaining gap between upper and lower complexity bounds. In the same paper the author also shows how tensor methods with $p = 3$ can be implemented when one solves the auxiliary problem with Bregman projected gradient method in the relative smoothness setting. Near-optimal, i.e. optimal up to a logarithmic factor, tensor methods for convex optimization were recently proposed in a number of works [26, 28, 25, 24, 7, 34, 27, 46, 23]. These developments allowed to propose faster second-order methods via implementing third-order methods with inexact third derivative [47, 35, 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 [6, 39]. To the best of our knowledge, there is no analysis of tensor methods in the literature for stochastic convex smooth enough problems.

Our analysis is based on inexact versions of tensor methods for convex optimization, which use inexact derivatives of higher-order. First-order methods with inexact gradients are well-developed in the literature, see, for example, [50, 15, 16, 20, 22, 13, 19, 57] and references therein. Some results on inexact second-order methods are listed above. The paper [46] proposes an analysis of third-order method for convex optimization with inexact third derivative. In [6] the authors analyze inexact tensor methods for non-convex optimization. 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.

1.3 Our contribution

Motivated by stochastic optimization methods, we propose and analyze Inexact Tensor Method of the general order $p \geq 2$. The idea of the algorithm is to use inexact values of the derivatives up to the order p to construct a regularized inexact Taylor expansion. For the resulting method we provide sufficient conditions for the inexactness of the derivatives for this method in order to have similar complexity as the exact tensor method. This allows to prove sublinear $O(1/k^p)$ convergence rate and corresponding $O(\varepsilon^{-1/p})$ iteration complexity, which are the same as for the non-accelerated exact tensor methods.

As a corollary, we propose a stochastic tensor method for stochastic convex optimization problems in the online and offline settings. The idea of the algorithm is to sample in each iteration mini-batches of derivatives up to the order p and use them to construct regularized stochastic Taylor expansion. For the resulting method we prove sublinear $O(1/k^p)$ convergence rate and corresponding $O(\varepsilon^{-1/p})$ iteration complexity, which are the same as for the deterministic tensor methods. Interestingly, we show that to reach this, the sufficient mini-batch size decreases as the order of the derivative increases.

We also consider a particular case $p = 3$, for which in the spirit of [43], we describe how to implement the resulting inexact tensor method.

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 function. We prove its convexity and show that it majorizes the objective function. Section 4 is dedicated to the Inexact Tensor Method itself and its convergence. Next, in Section 5 we introduce the smooth version of the Inexact Tensor Method. Then, we discuss implementation details (Section 6). And finally, in Section 7 we introduce the Stochastic Tensor Method and prove how to satisfy sampling conditions. Future work is discussed in the very last Section.

2 Preliminaries

We consider stochastic convex optimization problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \mathbf{E}_{\xi\sim\mathcal{D}}[f(\mathbf{x};\xi)]. \tag{4}$$

We refer to this problem as online setting. As a particular case of this problem, we consider the following problem with the objective given as a finite-sum, which we call offline setting,

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}). \tag{5}$$

We denote the i-th directional derivative of function f at $\mathbf{x}$ along directions $\mathbf{s}_1, \dots, \mathbf{s}_i \in \mathbb{R}^n$ as

$$\nabla^i f(\mathbf{x})[\mathbf{s}_1, \dots, \mathbf{s}_i].$$

For example, $\nabla f(\mathbf{x})[\mathbf{s}_1] = \langle \nabla f(\mathbf{x}), \mathbf{s}_1\rangle$ and $\nabla^2 f(\mathbf{x})[\mathbf{s}_1, \mathbf{s}_2] = \langle \nabla^2 f(\mathbf{x})\mathbf{s}_1, \mathbf{s}_2\rangle$. If all directions are the same we write $\nabla^i f(\mathbf{x})[\mathbf{s}]^i$. By $\|\cdot\|$ we denote the tensor norm recursively induced by the Euclidean norm on the space of p-th order tensors

$$\|\mathbf{T}\| = \max_{\|\mathbf{s}_1\|=\ldots=\|\mathbf{s}_p\|=1} \{|\mathbf{T}[\mathbf{s}_1, \dots, \mathbf{s}_p]|\},$$

where $\mathbf{T}$ is p-th order tensor [10].

Assumption 1. Function f is convex and p times differentiable on $\mathbb{R}^n$ and its p-th derivative is Lipschitz continuous, i.e. for all $\mathbf{x}, \mathbf{y} \in \mathcal{L}(x_0)$[1]:

$$\|\nabla^p f(\mathbf{x}) - \nabla^p f(\mathbf{y})\| \le L_p \|\mathbf{x} - \mathbf{y}\|,$$

where $\mathcal{L}(\mathbf{x}_0) \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \le f(\mathbf{x}_0)\}$.

Following the previous works, we construct Tensor methods based on the Taylor approximation of the function $f(\mathbf{x})$, which can be written as follows:

$$\Phi_{\mathbf{x},p}(\mathbf{s}) \overset{\text{def}}{=} f(\mathbf{x}) + \sum_{i=1}^{p} \frac{1}{i!}\nabla^i f(\mathbf{x})[\mathbf{s}]^i, \quad \mathbf{s} \in \mathbb{R}^n. \tag{6}$$

Since the full Taylor expansion of f requires computing high-order derivatives which could be expensive to calculate, it is natural to use their approximations to construct an inexact model of the objective:

$$\phi_{\mathbf{x},p}(\mathbf{s}) = f(\mathbf{x}) + \sum_{i=1}^{p} \frac{1}{i!}\mathbf{G}_{\mathbf{x},i}[\mathbf{s}]^i, \tag{7}$$

where $\mathbf{G}_{\mathbf{x},i}$ are approximations to the derivatives $\nabla^i f(\mathbf{x})$. Inspired by the work [39] for the non-convex setting, we consider the following condition on the accuracy of the approximations for the derivatives

Condition 1. Given the target accuracy ε for the solution of the problem (1), there exist numbers $\kappa_i > 0$, $i = 1, \dots, p$ such that for all $\mathbf{s} \in \mathcal{L}(\mathbf{x}_0)$:

$$\|(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-1}\| \le \kappa_i \varepsilon^{(p-i+1)/p} \|\mathbf{s}\|^{i-1}, i = 1, \dots, p. \tag{8}$$

[1] Here and later in other assumptions we require $\mathbf{x}$ to be from Lebesgue set $\mathcal{L}(\mathbf{x}_0)$, since the Stochastic Tensor Method is monotone under Condition 1. In Section 7 we will show that Condition 1 can be satisfied with a high probability in both online and offline settings.

First, we analyze inexact Tensor methods under this condition. Then, motivated by stochastic optimization problems, we focus on stochastic approximations of the derivatives through sampling. More precisely, for $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_p$ being sample sets, we construct the sampled approximations as

$$\mathbf{G}_{\mathbf{x},i} = \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \nabla^i f(\mathbf{x}, \xi_j), \quad i = 1, .., p. \tag{9}$$

In Section 7 we show how to choose the size of sample sets $\mathcal{S}_i$ to satisfy Condition 1 with high probability without requiring knowledge of the length of the step $\|\mathbf{s}\|$. This allows us to construct the desired Stochastic Tensor Method and analyze its complexity.

In our methods we use the following power prox funtions:

$$d_p(\mathbf{x}) = \frac{1}{p}\|\mathbf{x}\|^p, \; p \geq 2. \tag{10}$$

Note that

$$\nabla d_p(\mathbf{x}) = \|\mathbf{x}\|^{p-2}\mathbf{x},$$

$$\nabla^2 d_p(\mathbf{x}) = (p-2)\|\mathbf{x}\|^{p-4}\mathbf{x}\mathbf{x}^* + \|\mathbf{x}\|^{p-2} I \succeq \|\mathbf{x}\|^{p-2} I, \tag{11}$$

where I is the identity matrix in $\mathbb{R}^n$

As it is shown, e.g. in [43], Assumption 1 allows to control the quality of the approximation of the objective f by its Taylor polynomial:

$$|f(x+\mathbf{s}) - \Phi_{\mathbf{x},p}(\mathbf{s})| \leq \frac{L_p}{(p+1)!}\|\mathbf{s}\|^{p+1}, \; \mathbf{x}, \mathbf{s} \in \mathbb{R}^n. \tag{12}$$

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

$$\|\nabla f(x+\mathbf{s}) - \nabla \Phi_{\mathbf{x},p}(\mathbf{s})\| \leq \frac{L_p}{p!}\|\mathbf{s}\|^p, \tag{13}$$

$$\|\nabla^2 f(x+\mathbf{s}) - \nabla^2 \Phi_{\mathbf{x},p}(\mathbf{s})\| \leq \frac{L_p}{(p-1)!}\|\mathbf{s}\|^{p-1}, \tag{14}$$

for all $\mathbf{x}, \mathbf{s} \in \mathbb{R}^n$.

3 Inexact Tensor Model

In this section, we analyze inexact Taylor polynomial (7) first to show under Condition 1 approximation bounds in the spirit of (12), (13), (14) but on the quality of the approximation of f by $\phi_{\mathbf{x},p}$, and, then, based on such bounds to 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 analyze in the next section.

The following lemma gives a counterpart of (12) 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 $\phi_{\mathbf{x},p}(\mathbf{s})$.

Lemma 3.1. *For any* $\mathbf{x}, \; \mathbf{s} \in \mathbb{R}^n$ *we have*

$$|f(\mathbf{x}+\mathbf{s}) - \phi_{\mathbf{x},p}(\mathbf{s})| \leq \sum_{i=1}^{p} \frac{1}{i!}\kappa_i \varepsilon^{\frac{p-i+1}{p}}\|\mathbf{s}\|^i + \frac{L_p}{(p+1)!}\|\mathbf{s}\|^{p+1}. \tag{15}$$

Proof. For any $\mathbf{x}, \mathbf{s} \in \mathbb{R}^n$:

$$|f(\mathbf{x}+\mathbf{s})-\phi_{\mathbf{x},p}(\mathbf{s})| \leq |f(\mathbf{x}+\mathbf{s})-\Phi_{\mathbf{x},p}(\mathbf{s})|+|\Phi_{\mathbf{x},p}(\mathbf{s})-\phi_{\mathbf{x},p}(\mathbf{s})| \overset{(12)}{\leq} \frac{L_p}{(p+1)!}\|\mathbf{s}\|^{p+1}+|\Phi_{\mathbf{x},p}(\mathbf{s})-\phi_{\mathbf{x},p}(\mathbf{s})|.$$

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

$$|\Phi_{\mathbf{x},p}(\mathbf{s}) - \phi_{\mathbf{x},p}(\mathbf{s})| \overset{(6),(7)}{=} |\sum_{i=1}^{p} \frac{1}{i!}(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^i|$$
$$\leq \sum_{i=1}^{p} \frac{1}{i!}\|(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-1}\|\|\mathbf{s}\| \overset{(8)}{\leq} \sum_{i=1}^{p} \frac{1}{i!}\kappa_i \varepsilon^{\frac{p-i+1}{p}}\|\mathbf{s}\|^i.$$

Combining both inequalities above finishes the proof. □

The following lemma gives a counterpart of (13), (14) when the inexact derivatives are used.

Lemma 3.2. *For any* $\mathbf{x},\ \mathbf{s} \in \mathbb{R}^n$ *we have*

$$\|\nabla f(\mathbf{x}+\mathbf{s}) - \nabla \phi_{\mathbf{x},p}(\mathbf{s})\| \leq \sum_{i=1}^{p} \frac{1}{(i-1)!}\kappa_i \varepsilon^{\frac{p-i+1}{p}}\|\mathbf{s}\|^{i-1} + \frac{L_p}{p!}\|\mathbf{s}\|^p, \tag{16}$$

$$\|\nabla^2 f(\mathbf{x}+\mathbf{s}) - \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s})\| \leq \sum_{i=2}^{p} \frac{1}{(i-2)!}\kappa_i \varepsilon^{\frac{p-i+1}{p}}\|\mathbf{s}\|^{i-2} + \frac{L_p}{(p-1)!}\|\mathbf{s}\|^{p-1}. \tag{17}$$

Proof. Firstly, let us prove the bound for the first derivatives. For any $\mathbf{x}, \mathbf{s} \in \mathbb{R}^n$:

$$\|\nabla f(\mathbf{x}+\mathbf{s}) - \nabla \phi_{\mathbf{x},p}(\mathbf{s})\| \leq \|\nabla f(\mathbf{x}+\mathbf{s}) - \nabla \Phi_{\mathbf{x},p}(\mathbf{s})\| + \|\nabla \Phi_{\mathbf{x},p}(\mathbf{s}) - \nabla \phi_{\mathbf{x},p}(\mathbf{s})\|$$
$$\overset{(13)}{\leq} \frac{L_p}{p!}\|\mathbf{s}\|^p + \|\nabla \Phi_{\mathbf{x},p}(\mathbf{s}) - \nabla \phi_{\mathbf{x},p}(\mathbf{s})\|$$

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

$$\|\nabla \Phi_{\mathbf{x},p}(\mathbf{s}) - \nabla \phi_{\mathbf{x},p}(\mathbf{s})\| \overset{(6),(7)}{=} |\sum_{i=1}^{p} \frac{1}{(i-1)!}(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-1}|$$
$$\leq \sum_{i=1}^{p} \frac{1}{(i-1)!}\|(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-1}\| \overset{(8)}{\leq} \sum_{i=1}^{p} \frac{1}{(i-1)!}\kappa_i \varepsilon^{\frac{p-i+1}{p}}\|\mathbf{s}\|^{i-1}.$$

Combining both inequalities above we get (16).

Now, we will obtain the bound for the second derivatives. For any $\mathbf{x},\ \mathbf{s} \in \mathbb{R}^n$ we have

$$\|\nabla^2 f(\mathbf{x}+\mathbf{s}) - \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s})\| \leq \|\nabla^2 f(\mathbf{x}+\mathbf{s}) - \nabla^2 \Phi_{\mathbf{x},p}(\mathbf{s})\| + \|\nabla^2 \Phi_{\mathbf{x},p}(\mathbf{s}) - \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s})\|$$
$$\overset{(14)}{\leq} \frac{L_p}{(p-1)!}\|\mathbf{s}\|^{p-1} + \|\nabla^2 \Phi_{\mathbf{x},p}(\mathbf{s}) - \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s})\|.$$

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

$$\|\nabla^2 \Phi_{\mathbf{x},p}(\mathbf{s}) - \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s})\| \overset{(6),(7)}{=} |\sum_{i=2}^{p} \frac{1}{(i-2)!}(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-2}|$$
$$\leq \sum_{i=2}^{p} \frac{1}{(i-2)!}\|(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-2}\| \leq \sum_{i=2}^{p} \frac{1}{(i-2)!}\kappa_i \varepsilon^{\frac{p-i+1}{p}}\|\mathbf{s}\|^{i-2},$$

where the last inequality is valid due to conditions (8) and the definition of the matrix norm:

$$\|(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-2}\| \overset{\text{def}}{=} \sup_{\mathbf{s}\in\mathbb{R}^n} \frac{\|((\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-2})^T s\|}{\|s\|} \leq \kappa_i \varepsilon^{\frac{p-i+1}{p}} \|\mathbf{s}\|^{i-2}.$$

Thus, we have obtained (17).

□

We finish this section by constructing a global upper bound for the objective f based on the inexact Taylor expansion $\phi_{\mathbf{x},p}$ which is regularized by functions d_i, $i = 1, ..., p$.

Theorem 3.3. *For any* $\mathbf{s},\ \mathbf{x} \in \mathbb{R}^n$*:*

$$0 \preccurlyeq \nabla^2 f(\mathbf{x}+\mathbf{s}) \preccurlyeq \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s}) + \sum_{i=2}^{p} \frac{1}{(i-2)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} \|\mathbf{s}\|^{i-2} I + \frac{L_p}{(p-1)!} \|\mathbf{s}\|^{p-1} I. \tag{18}$$

Moreover, for any $\mathbf{s},\ \mathbf{x} \in \mathbb{R}^n$ *, and* $\sigma \geq L_p$ *the function*

$$\omega_{\mathbf{x},p}(\mathbf{s}) = \phi_{\mathbf{x},p}(\mathbf{s}) + \sum_{i=1}^{p} \frac{1}{(i-1)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} d_i(\mathbf{s}) + \frac{\sigma}{(p-1)!} d_{p+1}(\mathbf{s}) \tag{19}$$

is convex and it majorizes the function f*:*

$$f(\mathbf{x}+\mathbf{s}) \leq \omega_{\mathbf{x},p}(\mathbf{s}) \ \ \forall \mathbf{s} \in \mathbb{R}^n. \tag{20}$$

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

$$\begin{aligned}\left\langle \left(\nabla^2 f(\mathbf{x}+\mathbf{s}) - \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s})\right)\mathbf{h}, \mathbf{h}\right\rangle &\leq \left\|\nabla^2 f(\mathbf{x}+\mathbf{s}) - \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s})\right\| \cdot \|\mathbf{h}\|^2 \\ &\overset{(17)}{\leq} \left(\sum_{i=2}^{p} \frac{1}{(i-2)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} \|\mathbf{s}\|^{i-2} + \frac{L_p}{(p-1)!} \|\mathbf{s}\|^{p-1}\right) \|\mathbf{h}\|^2\end{aligned}$$

Further,

$$\begin{aligned}0 \preccurlyeq \nabla^2 f(\mathbf{x}+\mathbf{s}) &\preccurlyeq \nabla^2 \phi_{\mathbf{x},p}(\mathbf{s}) + \sum_{i=2}^{p} \frac{1}{(i-2)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} \|\mathbf{s}\|^{i-2} I + \frac{L_p}{(p-1)!} \|\mathbf{s}\|^{p-1} I \\ &\preccurlyeq \sum_{i=2}^{p} \frac{1}{(i-2)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} \|\mathbf{s}\|^{i-2} I + \frac{\sigma}{(p-1)!} \|\mathbf{s}\|^{p-1} I = \nabla^2 \omega_{\mathbf{x},p}(\mathbf{s}).\end{aligned}$$

Therefore, $\omega_{\mathbf{x},p}(\mathbf{s})$ is convex. Finally,

$$\begin{aligned}f(\mathbf{x}+\mathbf{s}) &\overset{(15)}{\leq} \phi_{\mathbf{x},p}(\mathbf{s}) + \sum_{i=1}^{p} \frac{1}{i!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} \|\mathbf{s}\|^{i} + \frac{L_p}{(p+1)!} \|\mathbf{s}\|^{p+1} \\ &= \phi_{\mathbf{x},p}(\mathbf{s}) + \sum_{i=1}^{p} \frac{1}{(i-1)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} d_i(\mathbf{s}) + \frac{L_p}{p!} d_{p+1}(\mathbf{s}) \\ &\leq \phi_{\mathbf{x},p}(\mathbf{s}) + \sum_{i=1}^{p} \frac{1}{(i-1)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} d_i(\mathbf{s}) + \frac{\sigma}{(p-1)!} d_{p+1}(\mathbf{s}) = \omega_{\mathbf{x},p}(\mathbf{s}).\end{aligned}$$

□

Thus, we have build a regularized inexact Taylor polynomial $\omega_{\mathbf{x},p}(\mathbf{s})$ defined in (19) as an model of the objective f. Theorem 3.3 claims that this model satisfies two main conditions:

- Model $\omega_{\mathbf{x},p}(\mathbf{s})$ is a global upper bound for the function f:

$$f(\mathbf{x}+\mathbf{s}) \overset{(20)}{\leq} \omega_{\mathbf{x},p}(\mathbf{s}). \tag{21}$$

- Model $\omega_{\mathbf{x},p}(\mathbf{s})$ is convex.

4 Inexact Tensor Method

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

$$\mathbf{T}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^n} \omega_{\mathbf{x},p}(\mathbf{s}). \tag{22}$$

Let us prove the next technical lemma.

Lemma 4.1. *For any $\mathbf{x}\in\mathbb{R}^n$ and $\sigma \geq L_p$*

$$f(\mathbf{x}+\mathbf{T}(\mathbf{x})) \leq \min_{\mathbf{s}\in\mathbb{R}^n}\left\{ f(\mathbf{x}+\mathbf{s}) + 2\sum_{i=1}^{p}\frac{1}{(i-1)!}\kappa_i\varepsilon^{\frac{p-i+1}{p}}d_i(\mathbf{s}) + \frac{L_p+p\sigma}{(p-1)!}d_{p+1}(\mathbf{s})\right\} \tag{23}$$

Proof.

$$f(\mathbf{x}+\mathbf{T}(\mathbf{x})) \overset{(21)}{\leq} \min_{\mathbf{s}\in\mathbb{R}^n}\left\{ \phi_{\mathbf{x},p}(\mathbf{s}) + \sum_{i=1}^{p}\frac{1}{(i-1)!}\kappa_i\varepsilon^{\frac{p-i+1}{p}}d_i(\mathbf{s}) + \frac{\sigma}{p!}d_{p+1}(\mathbf{s})\right\}$$
$$\overset{(15)}{\leq} \min_{\mathbf{s}\in\mathbb{R}^n}\left\{ f(\mathbf{x}+\mathbf{s}) + 2\sum_{i=1}^{p}\frac{1}{(i-1)!}\kappa_i\varepsilon^{\frac{p-i+1}{p}}d_i(\mathbf{s}) + \frac{L_p+p\sigma}{(p-1)!}d_{p+1}(\mathbf{s})\right\}.$$

□

Now we are in a position to estimate the convergence rate of the Inexact Tensor Method.

Algorithm 1 Inexact Tensor Method

1: **Input:** convex function f such that $\nabla^p f$ is L_p-Lipschitz; ε is target objective residual; x_0 is starting point; constant $\sigma \geq L_p$.
2: **for** $k \geq 0$ **do**
3: Call the inexact oracle to compute $\mathbf{G}_{\mathbf{x}_k,i}$ for $i = 1,\dots,p$ such that Condition 1 is satisfied.
4: Obtain $\mathbf{s}_k$ and make the step:

$$\begin{aligned}\mathbf{s}_k &= \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^n}\omega_{\mathbf{x},p}(\mathbf{s})\\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \mathbf{s}_k.\end{aligned} \tag{24}$$

5: **end for**

In the view of Eq. (20) the process $\mathbf{x}+\mathbf{T}(\mathbf{x})$ guarantees that $f(\mathbf{x}+\mathbf{T}(\mathbf{x})) \leq f(\mathbf{x})$, i.e. Algorithm 1 is monotone. Therefore, in Assumption 1 we can only require $\mathbf{x}$ to be from the Lebesgue set of the objective function f or a vicinity of it if the method is stochastic.

Theorem 4.2. *If Condition 1 is satisfied and $f(\mathbf{x})$ is convex p times differentiable function with Lipschitz constant L_p for p-th derivative and $\sigma \geq L_p$ then after $T+1$ iterations of Algorithm 1 we have the following bound for the objective residual:*

$$f(\mathbf{x}_{T+1}) - f(\mathbf{x}_*) \leq 2\sum_{i=1}^{p} \frac{\kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i}{i!} \frac{(p+1)^i}{(T+p+1)^{i-1}} + \frac{(L_p + p\sigma)}{(p+1)!} \frac{(p+1)^{p+1}}{(T+p+1)^p} D^{p+1}, \quad (25)$$

where $D = \max\limits_{\mathbf{x} \in \mathcal{L}(\mathbf{x}_0)} \|\mathbf{x} - \mathbf{x}_*\|$.

Proof.

$$f(\mathbf{x}_1) \overset{(23)}{\leq} \min_{\mathbf{s} \in \mathbb{R}^n} \left\{ f(\mathbf{x}_1 + \mathbf{s}) + 2\sum_{i=1}^{p} \frac{1}{(i-1)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} d_i(\mathbf{s}) + \frac{L_p + p\sigma}{(p-1)!} d_{p+1}(\mathbf{s}) \right\}$$
$$\leq f(\mathbf{x}_*) + 2\sum_{i=1}^{p} \frac{1}{i!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i + \frac{L_p + p\sigma}{(p+1)!} D^{p+1},$$

For any $t > 1$:

$$f(\mathbf{x}_{t+1}) \overset{(23)}{\leq} \min_{\mathbf{s} \in \mathbb{R}^n} \left\{ f(\mathbf{x}_{t+1} + \mathbf{s}) + 2\sum_{i=1}^{p} \frac{1}{(i-1)!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} d_i(\mathbf{s}) + \frac{L_p + p\sigma}{(p-1)!} d_{p+1}(\mathbf{s}) \right\}$$
$$\leq \min_{\alpha_t \in [0,1]} \left\{ f(\mathbf{x}_t + \alpha_t(\mathbf{x}_* - \mathbf{x}_t)) + 2\sum_{i=1}^{p} \frac{1}{i!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p+1)!} (\alpha_t D)^{p+1} \right\}$$
$$\leq \min_{\alpha_t \in [0,1]} \left\{ f(\mathbf{x}_t) - \alpha_t(f(\mathbf{x}_t) - f(\mathbf{x}_*)) + 2\sum_{i=1}^{p} \frac{1}{i!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p+1)!} (\alpha_t D)^{p+1}. \right\}$$

Therefore,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_*) \leq (1-\alpha_t)(f(\mathbf{x}_t) - f(\mathbf{x}_*)) + 2\sum_{i=1}^{p} \frac{1}{i!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p+1)!} (\alpha_t D)^{p+1}. \quad (26)$$

Let us choose a sequence $\{A_i\}$ as follows:

$$A_t = \begin{cases} 1, t = 0 \\ \prod\limits_{i=1}^{t} (1 - \alpha_i), \ t \geq 1. \end{cases} \quad (27)$$

We divide both sides of (26) by A_t:

$$\frac{1}{A_t} \left(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_*) \right)$$
$$\leq \frac{1}{A_t} \left((1-\alpha_t)(f(\mathbf{x}_t) - f(x_*)) + 2\sum_{i=1}^{p} \frac{1}{i!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p+1)!} (\alpha_t D)^{p+1} \right)$$
$$= \frac{1}{A_{t-1}} (f(\mathbf{x}_t) - f(x_*)) + \frac{1}{A_t} \left(2\sum_{i=1}^{p} \frac{1}{i!} \kappa_i \varepsilon^{\frac{p-i+1}{p}} (\alpha_t D)^i + \frac{L_p + p\sigma}{(p+1)!} (\alpha_t D)^{p+1} \right).$$

Summing both sides from $t = 1$ to $t = T$ we obtain:

$$f(\mathbf{x}_{T+1}) - f(\mathbf{x}_*) \leq 2\sum_{i=1}^{p} \frac{\kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i}{i!} \sum_{t=1}^{T} \frac{A_T \alpha_t^i}{A_t} + \frac{(L_p + p\sigma) D^{p+1}}{(p+1)!} \sum_{t=1}^{T} \frac{A_T \alpha_t^{p+1}}{A_t}.$$

Let us take

$$\alpha_t = \frac{p+1}{t+p+1} \tag{28}$$

and compute the following sums

$$A_T = \prod_{t=1}^{T} (1 - \alpha_t) = \prod_{t=1}^{T} \frac{t}{t+p+1} = \frac{T!(p+1)!}{(T+p+1)!} = (p+1)! \prod_{t=1}^{p+1} \frac{1}{T+t} \geq \frac{(p+1)!}{(T+1)^{p+1}},$$

which gives

$$\begin{aligned}
\sum_{t=1}^{T} \frac{A_T \alpha_t^i}{A_t} &= \sum_{i=1}^{T} \frac{(p+1)^i \prod_{t=1}^{p+1}(T+t)}{(t+p+1)^i (p+1)!} \cdot (p+1)! \prod_{j=1}^{p+1} \frac{1}{t+i} \\
&= (p+1)^i \sum_{t=1}^{T} \frac{\prod_{j=1}^{p+1}(t+j)}{(t+p+1)^i} \prod_{t=1}^{p+1} \frac{1}{T+t} \\
&\leq \frac{(p+1)^i}{(T+p+1)^{i-1}}.
\end{aligned}$$

Hence,

$$f(\mathbf{x}_{T+1}) - f(\mathbf{x}_*) \leq 2\sum_{i=1}^{p} \frac{\kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i}{i!} \frac{(p+1)^i}{(T+p+1)^{i-1}} + \frac{(L_p + p\sigma)}{(p+1)!} \frac{(p+1)^{p+1}}{(T+p+1)^p} D^{p+1}. \tag{29}$$

□

This result provides an upper bound for the objective residual after t iterations of the Inexact Tensor Method. The right term corresponds to the case of exact Tensor method, i.e. $\kappa_i = 0$, $i = 1, ..., p$, and provides similar convergence rate as for non-accelerated Tensor method in [43]. The left terms show how inexactness in each derivative influences the convergence rate. In particular, if $\kappa_i > 0$ only for one $i \in [1, p]$, the corresponding convergence rate slows down to $1/t^{i-1}$. At the same time since we are interested in accuracy ε, the right term says us that we should take $t + p + 1$ of the order $(L_p D^{p+1}/\varepsilon)^{1/p}$ to obtain the whole r.h.s. smaller than ε. In this case, we obtain

$$\begin{aligned}
f(\mathbf{x}_{T+1}) - f(\mathbf{x}_*) &\leq 2\sum_{i=1}^{p} \frac{\kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i}{i!} \frac{(p+1)^i}{(T+p+1)^{i-1}} + \frac{(L_p + p\sigma)}{(p+1)!} \frac{(p+1)^{p+1}}{(T+p+1)^p} D^{p+1} \\
&\leq 2\sum_{i=1}^{p} \frac{\kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i}{i!} \frac{(p+1)^i}{L_p^{\frac{i-1}{p}} D^{\frac{(i-1)(p+1)}{p}} \varepsilon^{-\frac{i-1}{p}}} + \varepsilon \\
&= 2\sum_{i=1}^{p} \kappa_i \varepsilon^{\frac{p-i+1}{p} + \frac{i-1}{p}} D^{i - \frac{(i-1)(p+1)}{p}} L_p^{-\frac{i-1}{p}} \frac{(p+1)^i}{i!} + \varepsilon \\
&= 2\sum_{i=1}^{p} \kappa_i \varepsilon D^{\frac{p-i+1}{p}} L_p^{-\frac{i-1}{p}} \frac{(p+1)^i}{i!} + \varepsilon
\end{aligned} \tag{30}$$

We now can ask the question on how to control the parameters κ_i, $i = 1, ..., p$ in order to achieve the objective residual ε with the same iteration complexity $\varepsilon^{-1/p}$ as for the exact Tensor Method in [43]. The answer is given by the following result.

Corollary 4.3. *Let us choose*

$$\kappa_i = \frac{L^{\frac{i-1}{p}} i!}{D^{\frac{p-i+1}{p}}}. \tag{31}$$

Then, after $T + 1$ *iterations of Algorithm 1 chosen such that* T *satisfies*

$$(T + p + 1)^p = \frac{(p+1)^{p+2}}{(p+1)!} \frac{L_p + p\sigma}{\varepsilon} D^{p+1} \tag{32}$$

we get

$$f(\mathbf{x}_{T+1}) - f(\mathbf{x}_*) \leq \varepsilon.$$

Proof. From Theorem 4.2

$$\begin{aligned} f(\mathbf{x}_{T+1}) - f(\mathbf{x}_*) &\leq 2 \sum_{i=1}^{p} \frac{\kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i}{i!} \frac{(p+1)^i}{(T+p+1)^{i-1}} + \frac{(L_p + p\sigma)}{(p+1)!} \frac{(p+1)^{p+1}}{(T+p+1)^p} D^{p+1} \\ &\overset{(32)}{\leq} 2 \sum_{i=1}^{p} \frac{\kappa_i \varepsilon^{\frac{p-i+1}{p}} D^i}{i!} \frac{(p+1)^i}{L_p^{\frac{i-1}{p}} D^{\frac{(i-1)(p+1)}{p}} \varepsilon^{-\frac{i-1}{p}}} + \frac{\varepsilon}{p+1} \\ &= 2 \sum_{i=1}^{p} \kappa_i \varepsilon D^{\frac{p-i+1}{p}} L_p^{-\frac{i-1}{p}} \frac{(p+1)^i}{i!} + \frac{\varepsilon}{p+1} \qquad (33) \\ &= \varepsilon. \qquad (34) \end{aligned}$$

□

Thus, we showed that the number of steps needed for the Inexact Tensor Method to find an ε solution to the problem (1) is $O\left(\left(\frac{L_p D^{p+1}}{\varepsilon}\right)^{1/p}\right)$ under the appropriate assumption on the inexactness.

5 Smooth model

One of the inconveniences of the inexact model $\omega_{\mathbf{x},p}(\mathbf{s})$ is that it uses odd powers of $\|s\|$, which can lead to computationally expansive iterations (24). Thus, in this section we introduce a smooth version of the model $\omega_{\mathbf{x},p}(\mathbf{s})$ which uses only even powers of $\|s\|$. We obtain the smooth model from the definition of the model $\omega_{\mathbf{x},p}(\mathbf{s})$ (19), using the next inequality

$$\|x\| \leq \frac{\|x\|^2}{2\alpha} + \frac{\alpha}{2} \tag{35}$$

with $\alpha = \varepsilon^{1/p}$.

Firstly, let us consider the case when p is odd. Applying inequality (35) to the definition (19) we obtain:

$$\omega_{\mathbf{x},p}(\mathbf{s}) \leq \phi_{\mathbf{x},p}(\mathbf{s}) + \frac{1}{2}\kappa_1 \varepsilon^{\frac{p+1}{p}} + \sum_{i=1}^{\frac{p-1}{2}} \left(\frac{1}{2}\frac{\kappa_{2i-1}}{(2i-1)!} + \frac{\kappa_{2i}}{(2i)!} + \frac{1}{2}\frac{\kappa_{2i+1}}{(2i+1)!} \right) \varepsilon^{\frac{p-2i+1}{p}} \|s\|^{2i} + \left(\frac{1}{2}\frac{\kappa_p}{p!} + \frac{\sigma}{(p+1)!} \right) \|s\|^{p+1} \stackrel{\text{def}}{=} \zeta^o_{\mathbf{x},p}(\mathbf{s}).$$

Now, assume that p is even. We are going to use the same technique, but it will lead to a slightly different result:

$$\omega_{\mathbf{x},p}(\mathbf{s}) \leq \phi_{\mathbf{x},p}(\mathbf{s}) + \frac{1}{2}\kappa_1 \varepsilon^{\frac{p+1}{p}} + \sum_{i=1}^{\frac{p-2}{2}} \left(\frac{1}{2}\frac{\kappa_{2i-1}}{(2i-1)!} + \frac{\kappa_{2i}}{(2i)!} + \frac{1}{2}\frac{\kappa_{2i+1}}{(2i+1)!} \right) \varepsilon^{\frac{p-2i+1}{p}} \|s\|^{2i} + \left(\frac{1}{2}\frac{\kappa_{p-1}}{(p-1!} + \frac{\kappa_p}{p!} + \frac{1}{2(p+1)}\frac{\sigma}{(p-1)!} \right) \varepsilon^{\frac{1}{p}} \|s\|^p + \frac{\sigma}{2(p+1)!}\varepsilon^{-\frac{1}{p}} \|s\|^{p+2} \stackrel{\text{def}}{=} \zeta^e_{\mathbf{x},p}(\mathbf{s}).$$

Finally,

$$\zeta_{\mathbf{x},p}(\mathbf{s}) \stackrel{\text{def}}{=} \begin{cases} \zeta^o_{\mathbf{x},p}(\mathbf{s}), & \text{if } p \text{ is odd,} \\ \zeta^e_{\mathbf{x},p}(\mathbf{s}), & \text{if } p \text{ is even.} \end{cases}$$

Since the model ζ_p is a global upper bound for the model $\omega_{\mathbf{x},p}(\mathbf{s})$ (19), the statement of Theorem 3.3 also holds for the smooth model ζ_p. Therefore, the smooth version of Algorithm (1) differs only in the step (24) and reads as

$$\begin{aligned} \mathbf{s}_k &= \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^n} \zeta_{\mathbf{x},p}(\mathbf{s}); \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \mathbf{s}_k. \end{aligned} \tag{36}$$

The complexity of an iterative method based on these equalities is the same as the complexity of the non-smooth version described in the previous section. The proof of this fact is exactly the same as the proof of Theorem 4.2 up to constant coefficients of the model.

6 Implementation Details

Inexact tensor optimization methods, introduced in Sections 4, 5, are based on the solution of auxiliary subproblems (24), (36) 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 the Inexact Tensor Methods. In this section, we show how to treat these problems in the particluar case of $p = 3$. To do that, we consider the third degree smooth model $\zeta(\mathbf{s})$ which corresponds to $p = 3$

$$\zeta_{\mathbf{x}}(\mathbf{h}) \stackrel{\text{def}}{=} \zeta^o_{\mathbf{x},3}(\mathbf{h}) = \phi_{\mathbf{x},3}(\mathbf{h}) + \frac{\kappa_g \varepsilon^{\frac{4}{3}}}{2} + \left(\frac{\kappa_g}{2} + \frac{\kappa_b}{2} + \frac{\kappa_t}{12} \right) \varepsilon^{\frac{2}{3}} \|\mathbf{h}\|^2 + \left(\frac{\kappa_t}{12} + \frac{\sigma}{8} \right) \|\mathbf{h}\|^4, \tag{37}$$

where $\kappa_1 = \kappa_g$, $\kappa_2 = \kappa_b$, $\kappa_3 = \kappa_t$. We also introduce the following notation for the first three approximate (sampled) derivatives $\mathbf{G}_{\mathbf{x},1} = \mathbf{g}$, $\mathbf{G}_{\mathbf{x},2} = \mathbf{B}$, $\mathbf{G}_{\mathbf{x},3} = \mathbf{T}$ (see Eq.(9)). In this case formula (7) can be rewritten in the following form

$$\phi_{\mathbf{x}}(\mathbf{s}) \stackrel{\text{def}}{=} f(\mathbf{x}) + \mathbf{g}^T\mathbf{h} + \frac{1}{2}\mathbf{h}^T\mathbf{B}\mathbf{h} + \frac{1}{6}\mathbf{T}[\mathbf{h}]^3. \tag{38}$$

Lemma 6.1. *For any $\mathbf{h} \in \mathbb{R}^n$ and $\tau > 0$, we have*

$$-\frac{1}{\tau}\mathbf{B} - \frac{\tau}{2}L_3\|\mathbf{h}\|^2 - \frac{1}{\tau}\kappa_b\varepsilon^{\frac{2}{3}} - \kappa_t\varepsilon^{\frac{1}{3}}\|\mathbf{h}\| \preccurlyeq \mathbf{T} \preccurlyeq \frac{1}{\tau}\mathbf{B} + \frac{\tau}{2}L_3\|\mathbf{h}\|^2 + \frac{1}{\tau}\kappa_b\varepsilon^{\frac{2}{3}} + \kappa_t\varepsilon^{\frac{1}{3}}\|\mathbf{h}\|. \quad (39)$$

Proof. From (18) we obtain:

$$\begin{aligned} 0 \le \langle \nabla^2 f(\mathbf{x}+\mathbf{h})\mathbf{u},\mathbf{u}\rangle &\le \langle \nabla^2\phi(\mathbf{h})\mathbf{u},\mathbf{u}\rangle + \frac{L_3}{2}\|\mathbf{h}\|^2\|\mathbf{u}\|^2 + \kappa_b\varepsilon^{\frac{2}{3}}\|\mathbf{u}\|^2 + \kappa_t\varepsilon^{\frac{1}{3}}\|\mathbf{h}\|\|\mathbf{u}\|^2 \\ &\le \langle (\mathbf{B}+\mathbf{T}[\mathbf{h}])\mathbf{u},\mathbf{u}\rangle + \frac{L_3}{2}\|\mathbf{h}\|^2 + \kappa_b\varepsilon^{\frac{2}{3}}\|\mathbf{u}\|^2 + \kappa_t\varepsilon^{\frac{1}{3}}\|\mathbf{h}\|\|\mathbf{u}\|^2 \end{aligned}$$

Replacing $\mathbf{h}$ with $\tau\mathbf{h}$ and dividing by τ, we get

$$-\langle \mathbf{T}[\mathbf{h}]\mathbf{u},\mathbf{u}\rangle \le \frac{1}{\tau}\langle \mathbf{B}\mathbf{u},\mathbf{u}\rangle + \frac{\tau}{2}L_3\|\mathbf{h}\|^2\|\mathbf{u}\|^2 + \frac{1}{\tau}\|u\|^2\kappa_b\varepsilon^{\frac{2}{3}} + \kappa_t\varepsilon^{\frac{1}{3}}\|\mathbf{h}\|\|\mathbf{u}\|^2.$$

Replacing $\mathbf{h}$ by $-\mathbf{h}$ we obtain:

$$\langle \mathbf{T}[\mathbf{h}]\mathbf{u},\mathbf{u}\rangle \le \frac{1}{\tau}\langle \mathbf{B}\mathbf{u},\mathbf{u}\rangle + \frac{\tau}{2}L_3\|\mathbf{h}\|^2\|\mathbf{u}\|^2 + \frac{1}{\tau}\|u\|^2\kappa_b\varepsilon^{\frac{2}{3}} + \kappa_t\varepsilon^{\frac{1}{3}}\|\mathbf{h}\|\|\mathbf{u}\|^2.$$

From the last two inequalities we get (38).

□

Let's build an analogue of Lemma 6.1 for model $\zeta(\mathbf{h})$.

Lemma 6.2. *For any $\mathbf{h} \in \mathbb{R}^n$ and $\tau > 0$, we have*

$$\begin{aligned} &-\frac{1}{\tau}\mathbf{B} - \frac{\tau}{2}\left(L_3+\kappa_t\right)\|\mathbf{h}\|^2 - \frac{1}{\tau}\underbrace{\left(\kappa_b\varepsilon^{\frac{2}{3}} + \frac{1}{2}\kappa_t\varepsilon^{\frac{1}{3}}\right)}_{\tilde{C}_2} \\ &\preccurlyeq \mathbf{T} \preccurlyeq \frac{1}{\tau}\mathbf{B} + \frac{\tau}{2}\left(L_3+\kappa_t\right)\|\mathbf{h}\|^2 + \frac{1}{\tau}\underbrace{\left(\kappa_b\varepsilon^{\frac{2}{3}} + \frac{1}{2}\kappa_t\varepsilon^{\frac{1}{3}}\right)}_{\tilde{C}_2} \end{aligned} \quad (40)$$

Proof. Almost the same as the proof of Lemma 6.1. Just use inequality (35) with $\alpha = \varepsilon^{1/3}$ for the lower and upper bounds. □

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

$$\zeta(\mathbf{h}) = \phi(\mathbf{h}) + \underbrace{\left(\frac{\sigma}{2} + \frac{\kappa_t}{3}\right)}_{C_4} d_4(\mathbf{h}) + \underbrace{\left(\kappa_g\varepsilon^{\frac{2}{3}} + \kappa_b\varepsilon^{\frac{2}{3}} + \frac{\kappa_t\varepsilon^{\frac{2}{3}}}{6}\right)}_{C_2} d_2(\mathbf{h}) + \frac{\kappa_g\varepsilon^{\frac{4}{3}}}{2} \to \min_{\mathbf{h}\in\mathbb{R}^n}. \quad (41)$$

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

$$\begin{aligned}\nabla^2\zeta(\mathbf{h}) &= \nabla^2\phi(\mathbf{h}) + C_4\nabla^2 d_4(\mathbf{h})C_2\nabla^2 d_2(\mathbf{h}) = \mathbf{B} + \mathbf{T}[\mathbf{h}] + C_4\nabla^2 d_4(\mathbf{h}) + C_2\nabla^2 d_2(\mathbf{h}) \\ &\overset{(40)}{\succcurlyeq} \mathbf{B} + C_4\nabla^2 d_4(\mathbf{h}) + C_2\nabla^2 d_2(\mathbf{h}) - \frac{1}{\tau}\mathbf{B} - \frac{1}{\tau}\tilde{C}_2 - \frac{\tau}{2}(L_3+\kappa_t)\|s\|^2 \\ &\overset{(11)}{\succcurlyeq} \mathbf{B}\left(1-\frac{1}{\tau}\right) + \frac{(3\sigma+2\kappa_t)}{6}\nabla^2 d_4(\mathbf{h}) - \frac{\tau(L_3+\kappa_t)}{2}\nabla^2 d_4(\mathbf{h}) \\ &\quad + \left(\frac{2\kappa_t}{3} - \frac{\tau\kappa_t}{2}\right)\nabla^2 d_4(\mathbf{h}) + \left(C_2 - \frac{1}{\tau}\tilde{C}_2\right)\nabla^2 d_2(\mathbf{h}) \\ &\succcurlyeq \left(1-\frac{2}{\tau}\right)\mathbf{B} + \left(1-\frac{2}{\tau}\right)C_2\nabla^2 d_2(\mathbf{h}) + \left(1-\frac{2}{\tau}\right)\frac{2\kappa_t}{3}\nabla^2 d_4(\mathbf{h}) \\ &\quad + \left(\frac{(3\sigma+2\kappa_t) - 3\tau(L_3+\kappa_t)}{6}\right)\nabla^2 d_4(\mathbf{h}),\end{aligned}$$

where the last inequality comes from the following inequality

$$2C_2 \geq \tilde{C}_2. \tag{42}$$

Let $\rho_{\mathbf{x}}(\mathbf{h}) = \frac{1}{2}\left(1-\frac{2}{\tau}\right)\langle \mathbf{B}\mathbf{h}, \mathbf{h}\rangle + \frac{\sigma - L_3\tau}{2}d_4(\mathbf{h}) + \left(1-\frac{2}{\tau}\right)C_2 d_2(\mathbf{h}) + \left(\frac{(3\sigma+2\kappa_t)-3\tau(L_3+\kappa_t)}{6}\right)d_4(\mathbf{h})$. Therefore we have proved the strong relative convexity:

$$\nabla^2\zeta(\mathbf{h}) \succcurlyeq \nabla^2\rho_{\mathbf{x}}(\mathbf{h}). \tag{43}$$

On the other hand,

$$\begin{aligned}\nabla^2\zeta(\mathbf{h}) &= \mathbf{B} + \mathbf{T}[\mathbf{h}] + C_4\nabla^2 d_4(\mathbf{h}) + C_2\nabla^2 d_2(\mathbf{h}) \\ &\overset{(40)}{\preccurlyeq} \mathbf{B} + C_4\nabla^2 d_4(\mathbf{h}) + C_2\nabla^2 d_2(\mathbf{h}) + \frac{1}{\tau}\mathbf{B} + \frac{1}{\tau}\tilde{C}_2 + \frac{\tau}{2}(L_3+\kappa_t)\|s\|^2 \\ &\overset{(11)}{\preccurlyeq} \left(1+\frac{1}{\tau}\right)\mathbf{B} + \frac{3\sigma+2\kappa_t}{6}\nabla^2 d_4(\mathbf{h}) + \frac{\tau(L_3+\kappa_t)}{2}\nabla^2 d_4(\mathbf{h}) + \left(C_2 + \frac{1}{\tau}\tilde{C}_2\right)\nabla^2 d_2(\mathbf{h}) \\ &\preccurlyeq \left(1+\frac{2}{\tau}\right)\mathbf{B} + \left(1+\frac{2}{\tau}\right)C_2\nabla^2 d_2(\mathbf{h}) + \left(\frac{(3\sigma+2\kappa_t) - 3\tau(L_3+\kappa_t)}{6}\right)\nabla^2 d_4(\mathbf{h})\end{aligned} \tag{44}$$

where the last inequality comes from (42).

Let us choose σ and κ_t such that the following inequality holds $2\sigma + 2\kappa_t = 3\tau^2(L_3+\kappa_t)$ for $\tau > 2$ in (41). With that choice of σ a model $\zeta(\mathbf{h})$ is convex since $\sigma = \frac{3}{2}\tau^2(L_3+\kappa_t) - \kappa_t > L_3$ for $\tau > 2$.

Under the above choice of the parameters, from (44) we have that

$$\begin{aligned}\nabla^2\zeta(\mathbf{h}) \preccurlyeq \left(\frac{\tau+2}{\tau-2}\right)\Bigg(&\left(1-\frac{2}{\tau}\right)\mathbf{B} + \left(1-\frac{2}{\tau}\right)C_2\nabla^2 d_2(\mathbf{h}) + \left(1-\frac{2}{\tau}\right)\frac{2\kappa_t}{3}\nabla^2 d_4(\mathbf{h}) \\ &+ \left(\frac{\sigma-\tau L_3}{2}\right)\nabla^2 d_4(\mathbf{h})\Bigg) \preccurlyeq \left(\frac{\tau+2}{\tau-2}\right)\nabla^2\rho_{\mathbf{x}}(\mathbf{h}).\end{aligned}$$

Thus, we have proved the strong relative convexity with constant 1 and relative smoothness with constant $\frac{\tau+2}{\tau-2}$ of $\zeta(\mathbf{h})$ w.r.t. $\rho_{\mathbf{x}}(\mathbf{h})$.

The relative smoothness condition allows to solve the auxiliary problem (41) very efficiently [43, 38] by the iterates

$$\mathbf{h}_{k+1} = \arg\min_{\mathbf{h}\in\mathbb{R}^n}\left\{\langle\nabla\zeta(\mathbf{h}_k), \mathbf{h}\rangle + \kappa(\tau)\beta_{\rho_x}(\mathbf{h}_k, \mathbf{h})\right) \tag{45}$$

with linear rate of convergence.

According to [43] it is not necessary to calculate the full third derivative tensor $\mathbf{T}$ 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 Methods

In this section we apply 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 1 to be satisfied in these two settings. To do that we first introduce an additional assumption on the objective function f:

Assumption 2. The derivatives $f(x), \nabla f(x), \dots, \nabla^{p-1} f(x)$ are Lipschitz continious for all $i = 1, \dots, p-1$ and $\mathbf{x}, \mathbf{y} \in \mathcal{L}(\mathbf{x}_0)$:

$$\|\nabla^i f(\mathbf{x}) - \nabla^i f(\mathbf{y})\| \leq L_i \|\mathbf{x} - \mathbf{y}\|.$$

7.1 Stochastic Tensor Method

Our stochastic version of Algorithm 1 for the stochastic optimization problem (1) in the online (2) and offline (3) settings is the following algorithm.

Algorithm 2 Stochastic Tensor Method

1: **Input:** convex function f such that $\nabla^p f$ is L_p-Lipschitz; ε is objective residual; x_0 is starting point; constant $\sigma \geq L_p$.
2: **for** $k \geq 0$ **do**
3: Sample derivatives $\mathbf{G}_{\mathbf{x}_k, i}$ given in (9) for $i = 1, \dots, p$ such that Condition 1 is satisfied, see Lemma 7.2 for the online setting and Lemma 7.4 for the offline setting.
4: Obtain $\mathbf{s}_k$ and make the step:

$$\begin{aligned} \mathbf{s}_k &= \operatorname*{argmin}_{\mathbf{s} \in \mathbb{R}^n} \omega_{\mathbf{x}, p}(\mathbf{s}) \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \mathbf{s}_k. \end{aligned} \tag{46}$$

5: **end for**

In the next subsections we show how to choose the size of sample sets to satisfy Condition 1 for both online and offline settings.

7.2 Online setting

In this setting we need one more assumption to be able to satisfy Condition 1 in the case of stochastic optimization problem.

Assumption 3. For all $i = 1, 2, \dots, p$ and $\mathbf{x} \in \mathcal{L}(x_0)$:

$$\|\nabla^i f(\mathbf{x}, \xi) - \nabla^i f(\mathbf{x})\| \leq M_i.$$

From the following tensor concentration bound theorem we derive required conditions.

Theorem 7.1 (Tensor Hoeffding Inequality [39]). *Let $\mathcal{X}$ be a sum of n i.i.d. tensors $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times \cdots \times d_k}$. Let $\mathbf{u}_1, \dots \mathbf{u}_k$ be such that $\|\mathbf{u}_i\| = 1$ and assume that for each tensor $i, a \leq \mathcal{Y}_i(\mathbf{u}_1, \dots, \mathbf{u}_k) \leq b$. Let $\sigma := (b-a)$. Then we have*

$$P(\|\mathcal{X} - \mathbb{E}\mathcal{X}\| \geq t) \leq k_0^{\left(\sum_{i=1}^k d_i\right)} \cdot 2\exp\left(-\frac{t^2}{2n\sigma^2}\right)$$

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

This Theorem allows us to provide a sufficient condition on the sample sets $\mathcal{S}_i$ in order to satisfy Condition 1.

Lemma 7.2. *Let Assumptions 1, 2, 3 be satisfied. Then, for any fixed small constants $\kappa_i > 0$ we can choose the sizes of the sample sets $\mathcal{S}_i$ in equation* (9) *to be*

$$|\mathcal{S}_i| = n_i = \tilde{\mathcal{O}}\left(\frac{(L_{i-1} + M_i)^2}{\kappa_i^2} \cdot \varepsilon^{-2(p-i+1)/p}\right)$$

so that with probability $1 - \delta$ Condition 1 is satisfied.

Proof. Using Assumptions 1, 3 and the triangle inequality, we obtain

$$\|\mathbf{G}_{\mathbf{x},i}\| = \frac{1}{n_i}\sum_{j=1}^{n_i} \|\nabla^i f(\mathbf{x}, \xi_j)\| \leq \frac{1}{n_i}\sum_{j=1}^{n_i} \left(\|\nabla^i f(\mathbf{x}, \xi_j) - \nabla^i f(\mathbf{x})\| + \|\nabla^i f(\mathbf{x})\|\right) \leq M_i + L_{i-1} \stackrel{\text{def}}{=} \sigma_i.$$

Then, the proof completely replies the proof of Lemma 11 in [39]. We require the probability of a deviation larger or equal to t to be lower than $\delta \in (0, 1]$.

$$\mathbf{P}\{\left\|\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x})\right\| > t\} \leq k_0^{ni} \cdot 2\exp\left(-\frac{t^2 n_i}{2\sigma_i^2}\right) \leq \delta.$$

Taking the log of both sides, we get

$$-\frac{t^2 n_i}{2\sigma_t^2} \leq \log\frac{\delta}{2k_0^{ni}}$$

which is equivalent to

$$n_i \geq \frac{2\sigma_t^2}{t^2}\log\frac{2k_0^{ni}}{\delta}.$$

Finally, we can simply choose $t = \kappa_i \epsilon^{\frac{p-i+1}{p}}$ in order to satisfy Eq. (8) since $\forall \mathbf{s} \in \mathbb{R}^n$

$$\begin{aligned}\left\|\mathbf{G}_{\mathbf{x},i}[\mathbf{s}]^{i-1} - \nabla^i f(\mathbf{x})[\mathbf{s}]^{i-1}\right\| &\leq \left\|\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x})\right\| \|\mathbf{s}\|^{i-1} \\ &\leq \kappa_i \epsilon^{\frac{p-i+1}{p}} \|\mathbf{s}\|^{i-1}\end{aligned}$$

□

7.3 Offline setting

In this setting we use the following result to provide a sufficient condition for Condition 1 to hold.

Theorem 7.3 (Tensor Hoeffding-Serfling Inequality [39])**.** *Let $\mathcal{X}$ be a sum of n tensors $\mathcal{Y}_i \in \mathbb{R}^{d_1 \times \ldots \times d_k}$, sampled without replacement from a finite population $\mathcal{A}$ of size N. Let $\boldsymbol{u}_i, ..., \boldsymbol{u}_k$ be such that $\|\boldsymbol{u}_i\| = 1$ and assume that for each tensor i, $a \leq \mathcal{Y}_i(\boldsymbol{u}_i, ..., \boldsymbol{u}_k) \leq b$. Let $\sigma := (b-a)$, then we have*

$$P(\|\mathcal{X} - \mathsf{E}\mathcal{X}\| \geq t) \leq k_0^{\sum_{i=1}^k d_i} \cdot 2\exp\left(-\frac{t^2 n^2}{2\sigma^2(n+1)(1-n/N)}\right), \tag{47}$$

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

The next lemma follows from Theorem 7.3. The proof can be found in [39].

Lemma 7.4. *Let Assumptions 1, 2 be satisfied. Then, for any fixed small constants $\kappa_i > 0$ we can choose the sizes $|\mathcal{S}_i|$ of sample sets $\mathcal{S}_i$ in* (9) *to be*

$$n_i = |\mathcal{S}_i| = \tilde{\mathcal{O}}\left(\frac{\kappa_i^2 \varepsilon^{2(p-i+1)/p}}{L_{i-1}^2} + \frac{1}{m}\right)^{-1}$$

so that with probability $1-\delta$ Condition 1 holds.

8 Future Work

The most interesting and natural generalization of the results mentioned above is their generalization to accelerated tensor methods [41, 42, 43, 26, 28, 25, 17, 27, 47, 44, 35, 46, 23]. For second-order tensor methods this was partially done in [29].

Below we briefly formulate the main scheme of the paper and discuss possible accelerated generalizations. Assume that (see (8))

$$\|(\mathbf{G}_{\mathbf{x},i} - \nabla^i f(\mathbf{x}))[\mathbf{s}]^{i-1}\| = O\left(\delta_i \|\mathbf{s}^{i-1}\|\right), i = 1, \ldots, p.$$

Then, our result in this paper shows that for an inexact non accelerated tensor method which uses $\mathbf{G}_{\mathbf{x},i}$ instead of $\nabla^i f(\mathbf{x})$, we can obtain the following convergence guarantee (see also [27] for $p=1$ and [29, 18] for $p=2$):

$$f(\mathbf{x}_T) - f(\mathbf{x}_*) = O\left(\frac{L_p D^{p+1}}{T^p} + \frac{\delta_p D^p}{T^{p-1}} + ... + \frac{\delta_3 D^3}{T^2} + \frac{\delta_2 D^2}{T} + \delta_1 D\right),$$

where D is a diameter of the ball (with center at $\mathbf{x}_*$) that contains all $\{\mathbf{x}_t\}_{t=0}^T$. The best rate of convergence takes place when $\delta_i = 0$, $i = 1, ..., p$, which makes it sufficient to make $T \sim \varepsilon^{-1/p}$ steps to obtain $f(\mathbf{x}_T) - f(\mathbf{x}_*) \leq \varepsilon$. If $\delta_i \sim \varepsilon^{\frac{p-i+1}{p}}$ we still can take $T \sim \varepsilon^{-1/p}$ and achieve the same error ε. In the stochastic optimization setting, the sufficient batch sizes r_i for i-th derivatives can be obtained from these bounds by using a quite expected result, that $\delta_i \sim \frac{1}{\sqrt{r_i}}$.

For the accelerated tensor methods introduced in [41, 4, 43], i.e. the ones which have convergence rate $1/k^{p+1}$ instead of $1/k^p$, we may expect that (see [15, 13, 19] for $p=1$ and [29] for $p=2$):

$$f(\mathbf{x}_T) - f(\mathbf{x}_*) = O\left(\frac{L_p R^{p+1}}{T^{p+1}} + \frac{\delta_p R^p}{T^p} + ... + \frac{\delta_3 R^3}{T^3} + \frac{\delta_2 R^2}{T^2} + \delta_1 R\right),$$

where $R \geq \|\mathbf{x}_0 - \mathbf{x}_*\|$ and T is not very large.

For optimal (near-optimal) accelerated tensor methods [40, 25], i.e. the ones which have convergence rate $1/k^{(3p+1)/2}$ instead of $1/k^p$, the conjectured bound is

$$f(\mathbf{x}_T) - f(\mathbf{x}_*) = O\left(\frac{L_p R^{p+1}}{T^{\frac{3p+1}{2}}} + \frac{\delta_p R^p}{T^{\frac{3(p-1)+1}{2}}} + ... + \frac{\delta_3 R^3}{T^{\frac{7}{2}}} + \frac{\delta_2 R^2}{T^2} + \delta_1 R\right).$$

There are several arguments from [36], that allow to understand this result as oracle complexity separation result for the function of sum-type with different oracles for different terms. These terms have corresponding constants $L_{i-1} = \delta_i$. So we can consider the considered result to be generalization of [36].

Another possible generalization is strongly convex and uniformly convex optimization problems [24] and also Hólder higher-order derivatives setting [56]. We expect that the results for strongly convex problems allow to make this technique more applicable in Machine Learning due to regularization arguments [54, 53].

One more direction of possible generalizations is the statistical preconditioning of centralized distributed methods with tensor algorithm used by a master node [55, 63, 64, 32]. Here we have exact gradient and statistically estimated Hessian and third-order derivatives. Moreover, we expect that these results can be generalized to decentralized setup by using the approach of [52]. So, this may improve the best known bounds in (strongly) convex decentralized distributed optimization with similar terms [58] and makes the bounds close to the lower bounds [2], see also [30] for discussion and more references.

Finally, we note that all the results described above can be obtained under an assumption of inexact solution to auxiliary problems in the described methods [45, 17, 36, 44, 35, 46, 23].

References

[1] Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. *arXiv preprint arXiv:1710.10329*, 2017.

[2] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. *Advances in neural information processing systems*, 28:1756–1764, 2015.

[3] Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1-2):327–360, 2019.

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

[5] Stefania Bellavia, Gianmarco Gurioli, and Benedetta Morini. Adaptive cubic regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization. *IMA Journal of Numerical Analysis*, 04 2020. drz076.

[6] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Philippe L. Toint. Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2881–2915, 2019.

[7] Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization. In *Conference on Learning Theory*, pages 492–507. PMLR, 2019.

[8] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319, Dec 2011.

[9] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[10] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Improved second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *arXiv preprint arXiv:1708.04044*, 2017.

[11] Coralia. Cartis, Nick I. Gould, and Philippe L. Toint. Universal regularization methods: Varying the power, the smoothness and the accuracy. *SIAM Journal on Optimization*, 29(1):595–615, 2019.

[12] Pafnuty L. Chebyshev. *Collected Works [in Russian]*, volume 5. Izd. Akad. NaukSSSR: Moscow-Leningrad., 1951.

[13] Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1019–1028, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. arXiv:1805.12591.

[14] Andrew Conn, Nicholas Gould, and Philippe Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000.

[15] Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

[16] Olivier Devolder. *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*. PhD thesis, PhD thesis, ICTEAM and CORE, Université Catholique de Louvain, 2013.

[17] Nikita Doikov and Yurii Nesterov. Contracting proximal methods for smooth convex optimization. *SIAM Journal on Optimization*, 30(4):3146–3169, 2020.

[18] Nikita Doikov and Yurii Nesterov. Convex optimization based on global lower second-order models. *Advances in Neural Information Processing Systems 33 proceedings (NeurIPS 2020)*, 2020.

[19] Darina Dvinskikh and Alexander Gasnikov. Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems. *Ill posed Inverse problems*, 2021.

[20] Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.

[21] Yurii G. Evtushenko. Optimization and fast automatic differentiation. *Computing Center of RAS, Moscow.*, 2013.

[22] A. V. Gasnikov and P. E. Dvurechensky. Stochastic intermediate gradient method for convex optimization problems. *Doklady Mathematics*, 93(2):148–151, Mar 2016.

[23] Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Dmitry Kamzolov, Dmitry Pasechnyk, Vladislav Matykhin, Nazarii Tupitsa, and Alexei Chernov. Accelerated meta-algorithm for convex optimization. *Computational Mathematics and Mathematical Physics*, 61(1), 2020.

[24] Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, and César A. Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1374–1391, Phoenix, USA, 25–28 Jun 2019. PMLR. arXiv:1809.00382.

[25] Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A. Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1392–1393, Phoenix, USA, 25–28 Jun 2019. PMLR. arXiv:1809.00382.

[26] ALexander Gasnikov, Eduard Gorbunov, Dmitry Kovalev, Ahmed Mohammed, and Elena Chernousova. The global rate of convergence for optimal tensor methods in smooth convex optimization. *Computer research and modelling*, 10(6):737–753, 2018.

[27] A.V. Gasnikov. *Universal gradient descent*. Modern numerical optimization methods. MCCME, 2020.

[28] AV Gasnikov, EA Gorbunov, DA Kovalev, AAM Mokhammed, and EO Chernousova. Reachability of optimal convergence rate estimates for high-order numerical convex optimization methods. In *Doklady Mathematics*, volume 99, pages 91–94. Springer, 2019.

[29] Saeed Ghadimi, Han Liu, and Tong Zhang. Second-order methods with cubic regularization under inexact information, 2017. arXiv:1710.05782.

[30] Eduard Gorbunov, Alexander Rogozin, Aleksandr Beznosikov, Darina Dvinskikh, and Alexander Gasnikov. Recent theoretical advances in decentralized distributed convex optimization. *arXiv preprint arXiv:2011.13259*, 2020.

[31] Filip Hanzely, Nikita Doikov, Yurii Nesterov, and Peter Richtarik. Stochastic subspace cubic newton method. In *International Conference on Machine Learning*, pages 4027–4038. PMLR, 2020.

[32] Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. *arXiv preprint arXiv:2002.10726*, 2020.

[33] K. H. Hoffmann and H. J. Kornstaedt. Higher-order necessary conditions in abstract mathematical programming. *Journal of Optimization Theory and Applications*, 26(4):533–568, Dec 1978.

[34] Bo Jiang, Haoyue Wang, and Shuzhong Zhang. An optimal high-order tensor method for convex optimization. In *Conference on Learning Theory*, pages 1799–1801. PMLR, 2019.

[35] Dmitry Kamzolov and Alexander Gasnikov. Near-optimal hyperfast second-order method for convex optimization and its sliding. *arXiv preprint arXiv:2002.09050*, 2020.

[36] Dmitry Kamzolov, Alexander Gasnikov, and Pavel Dvurechensky. On the optimal combination of tensor optimization methods. *arXiv preprint arXiv:2002.01004*, 2020.

[37] Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtárik. Stochastic newton and cubic newton methods with simple local linear-quadratic rates. *arXiv preprint arXiv:1912.01597*, 2019.

[38] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[39] Aurelien Lucchi and Jonas Kohler. A stochastic tensor method for non-convex optimization. *arXiv preprint arXiv:1911.10367*, 2019.

[40] Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

[41] Yurii Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[42] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[43] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, pages 1–27, 2019.

[44] Yurii Nesterov. Inexact accelerated high-order proximal-point methods. *CORE DP*, 8, 2020.

[45] Yurii Nesterov. Inexact basic tensor methods for some classes of convex optimization problems. *Optimization Methods and Software*, 0(0):1–29, 2020.

[46] Yurii Nesterov. Inexact high-order proximal-point methods with auxiliary search procedure. *CORE DP*, 10, 2020.

[47] Yurii Nesterov. Superfast second-order methods for unconstrained convex optimization. *CORE DP*, 7, 2020.

[48] Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[49] Seonho Park, Seung Hyun Jung, and Panos M Pardalos. Combining stochastic adaptive cubic regularization with negative curvature for nonconvex optimization. *Journal of Optimization Theory and Applications*, 184(3):953–971, 2020.

[50] Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

[51] Anton Rodomanov and Dmitry Kropotov. A superlinearly-convergent proximal newton-type method for the optimization of finite sums. volume 48 of *Proceedings of Machine Learning Research*, pages 2597–2605, New York, New York, USA, 20–22 Jun 2016. PMLR.

[52] Alexander Rogozin, Vladislav Lukoshkin, Alexander Gasnikov, Dmitry Kovalev, and Egor Shulgin. Towards accelerated rates for distributed optimization over time-varying networks. *arXiv preprint arXiv:2009.11069*, 2020.

[53] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability and stability in the general learning setting. In *COLT*, 2009.

[54] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[55] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008, 2014.

[56] Chaobing Song, Yong Jiang, and Yi Ma. Unified acceleration of high-order algorithms under holder continuity and uniform convexity. *arXiv preprint arXiv:1906.00582*, 2019.

[57] Fedor Stonyakin, Alexander Gasnikov, Alexander Tyurin, Dmitry Pasechnyuk, Artem Agafonov, Pavel Dvurechensky, Darina Dvinskikh, Alexey Kroshnin, and Victorya Piskunova. Inexact model: A framework for optimization and variational inequalities. *arXiv preprint arXiv:1902.00990*, 2019.

[58] Ying Sun, Amir Daneshmand, and Gesualdo Scutari. Distributed optimization based on gradient-tracking revisited: Enhancing convergence rate via surrogation. *arXiv preprint arXiv:1905.02637*, 2019.

[59] Lloyd N. Trefethen. *Approximation theory and approximation practice*, volume 164. SIAM, 2019.

[60] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2899–2908. Curran Associates, Inc., 2018.

[61] Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan. A note on inexact condition for cubic regularized newton's method. *arXiv preprint arXiv:1808.07384v1*, 2018.

[62] Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan. Cubic regularization with momentum for nonconvex optimization. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 313–322, Tel Aviv, Israel, 22–25 Jul 2020. PMLR.

[63] Yuchen Zhang and Lin Xiao. Disco: Distributed optimization for self-concordant empirical loss. In *International conference on machine learning*, pages 362–370, 2015.

[64] Yuchen Zhang and Lin Xiao. Communication-efficient distributed optimization of self-concordant empirical loss. In *Large-Scale and Distributed Optimization*, pages 289–341. Springer, 2018.

[65] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularized Newton methods. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5990–5999, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.