Near-Optimal Hyperfast Second-Order Method for Convex Optimization

Dmitry Kamzolov(✉)

Moscow Institute of Physics and Technology, Moscow, Russia
kamzolov.dmitry@phystech.edu

Abstract. In this paper, we present a new Hyperfast Second-Order Method with convergence rate $O(N^{-5})$ up to a logarithmic factor for the convex function with Lipshitz 3rd derivative. This method based on two ideas. The first comes from the superfast second-order scheme of Yu. Nesterov (CORE Discussion Paper 2020/07, 2020). It allows implementing the third-order scheme by solving subproblem using only the second-order oracle. This method converges with rate $O(N^{-4})$. The second idea comes from the work of Kamzolov et al. (arXiv:2002.01004). It is the inexact near-optimal third-order method. In this work, we improve its convergence and merge it with the scheme of solving subproblem using only the second-order oracle. As a result, we get convergence rate $O(N^{-5})$ up to a logarithmic factor. This convergence rate is near-optimal and the best known up to this moment.

Keywords: Tensor method · Inexact method · Second-order method · Complexity

1 Introduction

In recent years, it has been actively developing higher-order or tensor methods for convex optimization problems. The primary impulse was the work of Yu. Nesterov [23] about the possibility of the implementation tensor method. He proposed a smart regularization of Taylor approximation that makes subproblem convex and hence implementable. Also Yu. Nesterov proposed accelerated tensor methods [22,23], later A. Gasnikov et al. [4,11,12,18] proposed the near-optimal tensor method via the Monteiro–Svaiter envelope [21] with line-search and got a near-optimal convergence rate up to a logarithmic factor. Starting from 2018–2019 the interest in this topic rises. There are a lot of developments in tensor methods, like tensor methods for Hölder-continuous higher-order derivatives [15,28], proximal methods [6], tensor methods for minimizing the gradient norm of convex function [9,15], inexact tensor methods [14,19,24], and near-optimal composition of tensor methods for sum of two functions [19]. There are some results about local convergence and convergence for strongly convex functions [7,10,11]. See [10] for more references on applications of tensor method.

The work was funded by RFBR, project number 19-31-27001.

Y. Kochetov et al. (Eds.): MOTOR 2020, CCIS 1275, pp. 167–178, 2020.
https://doi.org/10.1007/978-3-030-58657-7_15

At the very beginning of 2020, Yurii Nesterov proposed a Superfast Second-Order Method [25] that converges with the rate $O(N^{-4})$ for a convex function with Lipshitz third-order derivative. This method uses only second-order information during the iteration, but assume additional smoothness via Lipshitz third-order derivative.[1] Here we should note that for the first-order methods, the worst-case example can't be improved by additional smoothness because it is a specific quadratic function that has all high-order derivatives bounded [24].[2] But for the second-order methods, one can see that the worst-case example does not have Lipshitz third-order derivative. This means that under the additional assumption, classical lower bound $O(N^{-2/7})$ can be beaten, and Nesterov proposes such a method that converges with $O(N^{-4})$ up to a logarithmic factor. The main idea of this method to run the third-order method with an inexact solution of the Taylor approximation subproblem by method from Nesterov with inexact gradients that converges with the linear speed. By inexact gradients, it becomes possible to replace the direct computation of the third derivative by the inexact model that uses only the first-order information. Note that for non-convex problems previously was proved that the additional smoothness might speed up algorithms [1,3,14,26,29].

In this paper, we propose a Hyperfast Second-Order Method for a convex function with Lipshitz third-order derivative with the convergence rate $O(N^{-5})$ up to a logarithmic factor. For that reason, firstly, we introduce Inexact Near-optimal Accelerated Tensor Method, based on methods from [4,19] and prove its convergence. Next, we apply Bregman-Distance Gradient Method from [14,25] to solve Taylor approximation subproblem up to the desired accuracy. This leads us to Hyperfast Second-Order Method and we prove its convergence rate. This method have near-optimal convergence rates for a convex function with Lipshitz third-order derivative and the best known up to this moment.

The paper is organized as follows. In Sect. 2 we formulate problem and introduce some basic facts and notation. In Sect. 3 we propose Inexact Near-optimal Accelerated Tensor Method and prove its convergence rate. In Sect. 4 we propose Hyperfast Second-Order Method and get its convergence speed.

2 Problem Statement and Preliminaries

In what follows, we work in a finite-dimensional linear vector space $E = \mathbb{R}^n$, equipped with a Euclidian norm $\|\cdot\| = \|\cdot\|_2$.

We consider the following convex optimization problem:

$$\min_x f(x), \tag{1}$$

[1] Note, that for the first-order methods in non-convex case earlier (see, [5] and references therein) it was shown that additional smoothness assumptions lead to an additional acceleration. In convex case, as far as we know these works of Yu. Nesterov [24,25] are the first ones where such an idea was developed.

[2] However, there are some results [30] that allow to use tensor acceleration for the first-order schemes. This additional acceleration requires additional assumptions on smoothness. More restrictive ones than limitations of high-order derivatives.

where $f(x)$ is a convex function with Lipschitz p-th derivative, it means that

$$\|D^p f(x) - D^p f(y)\| \leq L_p \|x - y\|. \tag{2}$$

Then Taylor approximation of function $f(x)$ can be written as follows:

$$\Omega_p(f, x; y) = f(x) + \sum_{k=1}^{p} \frac{1}{k!} D^k f(x) [y - x]^k, \ y \in \mathbb{R}^n. \tag{3}$$

By (2) and the standard integration we can get next two inequalities

$$|f(y) - \Omega_p(f, x; y)| \leq \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \tag{4}$$

$$\|\nabla f(y) - \nabla \Omega_p(f, x; y)\| \leq \frac{L_p}{p!} \|y - x\|^p. \tag{5}$$

3 Inexact Near-Optimal Accelerated Tensor Method

Problem (1) can be solved by tensor methods [23] or its accelerated versions [4,12,18,22]. This methods have next basic step:

$$T_{H_p}(x) = \underset{y}{\operatorname{argmin}} \left\{ \tilde{\Omega}_{p,H_p}(f, x; y) \right\},$$

where

$$\tilde{\Omega}_{p,H_p}(f, x; y) = \Omega_p(f, x; y) + \frac{H_p}{p!} \|y - x\|^{p+1}. \tag{6}$$

For $H_p \geq L_p$ this subproblem is convex and hence implementable.

But what if we can not solve exactly this subproblem. In paper [25] it was introduced Inexact pth-Order Basic Tensor Method (BTMI$_p$) and Inexact pth-Order Accelerated Tensor Method (ATMI$_p$). They have next convergence rates $O(k^{-p})$ and $O(k^{-(p+1)})$, respectively. In this section, we introduce Inexact pth-Order Near-optimal Accelerated Tensor Method (NATMI$_p$) with improved convergence rate $\tilde{O}(k^{-\frac{3p+1}{2}})$, where $\tilde{O}(\cdot)$ means up to logarithmic factor. It is an improvement of Accelerated Taylor Descent from [4] and generalization of Inexact Accelerated Taylor Descent from [19].

Firstly, we introduce the definition of the inexact subproblem solution. Any point from the set

$$\mathcal{N}^{\gamma}_{p,H_p}(x) = \left\{ T \in \mathbb{R}^n \ : \ \|\nabla \tilde{\Omega}_{p,H_p}(f, x; T)\| \leq \gamma \|\nabla f(T)\| \right\} \tag{7}$$

is the inexact subproblem solution, where $\gamma \in [0; 1]$ is an accuracy parameter. N^0_{p,H_p} is the exact solution of the subproblem.

Next we propose Algorithm 1.

Algorithm 1 . Inexact pth-Order Near-optimal Accelerated Tensor Method (NATMI)

Input: convex function $f : \mathbb{R}^n \to \mathbb{R}$ such that $\nabla^p f$ is L_p-Lipschitz, $H_p = \xi L_p$ where ξ is a scaling parameter, γ is a desired accuracy of the subproblem solution.

Set $A_0 = 0, x_0 = y_0$

for $k = 0$ **to** $k = K - 1$ **do**

Compute a pair $\lambda_{k+1} > 0$ and $y_{k+1} \in \mathbb{R}^n$ such that

$$\frac{1}{2} \leq \lambda_{k+1} \frac{H_p \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}}{(p-1)!} \leq \frac{p}{p+1},$$

where

$$y_{k+1} \quad \in \quad \mathcal{N}^{\gamma}_{p,H_p}(\tilde{x}_k) \tag{8}$$

and

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1} A_k}}{2}, \; A_{k+1} = A_k + a_{k+1}$$

$$\tilde{x}_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} x_k .$$

Update $x_{k+1} := x_k - a_{k+1} \nabla f(y_{k+1})$

return y_K

To get the convergence rate of Algorithm 1 we prove additional lemmas. The first lemma gets intermediate inequality to connect theory about inexactness and method's theory.

Lemma 1. *If* $y_{k+1} \in \mathcal{N}^{\gamma}_{p,H_p}(\tilde{x}_k)$, *then*

$$\|\nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1})\| \leq \frac{\gamma}{1-\gamma} \cdot \frac{(p+1)H_p + L_p}{p!} \|y_{k+1} - \tilde{x}_k\|^p. \tag{9}$$

Proof. From triangle inequality we get

$$\begin{aligned}
\|\nabla f(y_{k+1})\| &\leq \|\nabla f(y_{k+1}) - \nabla \Omega_p(f, \tilde{x}_k; y_{k+1})\| \\
&\quad + \|\nabla \Omega_p(f, \tilde{x}_k; y_{k+1}) - \nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1})\| + \|\nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1})\| \\
&\overset{(5),(6),(7)}{\leq} \frac{L_p}{p!} \|y_{k+1} - \tilde{x}_k\|^p + \frac{(p+1)H_p}{p!} \|y_{k+1} - \tilde{x}_k\|^p + \gamma \|\nabla f(y_{k+1})\|.
\end{aligned}$$

Hence,

$$(1-\gamma)\|\nabla f(y_{k+1})\| \leq \frac{(p+1)H_p + L_p}{p!} \|y_{k+1} - \tilde{x}_k\|^p.$$

And finally from (7) we get

$$\|\nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1})\| \leq \frac{\gamma}{1-\gamma} \cdot \frac{(p+1)H_p + L_p}{p!} \|y_{k+1} - \tilde{x}_k\|^p.$$

Next lemma plays the crucial role in the prove of the Algorithm 1 convergence. It is the generalization for inexact subpropblem of Lemma 3.1 from [4].

Lemma 2. *If* $y_{k+1} \in \mathcal{N}^{\gamma}_{p,H_p}(\tilde{x}_k)$, $H_p = \xi L_p$ *such that* $1 \geq 2\gamma + \frac{1}{\xi(p+1)}$ *and*

$$\frac{1}{2} \leq \lambda_{k+1} \frac{H_p \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}}{(p-1)!} \leq \frac{p}{p+1}, \text{ then} \tag{10}$$

$$\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}\nabla f(y_{k+1}))\| \leq \sigma \cdot \|y_{k+1} - \tilde{x}_k\|, \tag{11}$$

$$\sigma \geq \frac{p\xi + 1 - \xi + 2\gamma\xi}{(1-\gamma)2p\xi}, \tag{12}$$

where $\sigma \leq 1$.

Proof. Note, that by definition

$$\begin{aligned} \nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1}) &= \nabla \Omega_p(f, \tilde{x}_k; y_{k+1}) \\ &\quad + \frac{H_p(p+1)}{p!}\|y_{k+1} - \tilde{x}_k\|^{p-1}(y_{k+1} - \tilde{x}_k). \end{aligned} \tag{13}$$

Hence,

$$\begin{aligned} y_{k+1} - \tilde{x}_k = &\frac{p!}{H_p(p+1)\|y_{k+1} - \tilde{x}_k\|^{p-1}} \\ &\cdot \left(\nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1}) - \nabla \Omega_p(f, \tilde{x}_k; y_{k+1})\right). \end{aligned} \tag{14}$$

Then, by triangle inequality we get

$$\begin{aligned} &\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}\nabla f(y_{k+1}))\| = \|\lambda_{k+1}(\nabla f(y_{k+1}) - \nabla \Omega_p(f, \tilde{x}_k; y_{k+1})) \\ &+ \lambda_{k+1}\nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1}) \\ &+ \left(y_{k+1} - \tilde{x}_k + \lambda_{k+1}(\nabla \Omega_p(f, \tilde{x}_k; y_{k+1}) - \nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1}))\right)\| \\ &\overset{(5),(14)}{\leq} \lambda_{k+1}\frac{L_p}{p!}\|y_{k+1} - \tilde{x}_k\|^p + \lambda_{k+1}\|\nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1})\| \\ &+ \left|\lambda_{k+1} - \frac{p!}{H_p \cdot (p+1) \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}}\right| \\ &\cdot \|\nabla \tilde{\Omega}_{p,H_p}(f, \tilde{x}_k; y_{k+1}) - \nabla \Omega_p(f, \tilde{x}_k; y_{k+1})\| \end{aligned}$$

$$\begin{aligned} &\overset{(9),(13)}{\leq} \|y_{k+1} - \tilde{x}_k\| \left(\lambda_{k+1}\frac{L_p}{p!}\|y_{k+1} - \tilde{x}_k\|^{p-1}\right. \\ &\left. + \lambda_{k+1}\frac{\gamma}{1-\gamma} \cdot \frac{(p+1)H_p + L_p}{p!}\|y_{k+1} - \tilde{x}_k\|^{p-1}\right) \\ &+ \left|\lambda_{k+1} - \frac{p!}{H_p \cdot (p+1) \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}}\right| \cdot \frac{(p+1)H_p}{p!}\|y_{k+1} - \tilde{x}_k\|^p \end{aligned}$$

$$
\begin{aligned}
&= \|y_{k+1} - \tilde{x}_k\| \left(\frac{\lambda_{k+1}}{p!} \left(L_p + \frac{\gamma}{1-\gamma}((p+1)H_p + L_p) \right) \|y_{k+1} - \tilde{x}_k\|^{p-1} \right) \\
&+ \|y_{k+1} - \tilde{x}_k\| \left| \frac{\lambda_{k+1}(p+1)H_p}{p!} \|y_{k+1} - \tilde{x}_k\|^{p-1} - 1 \right| \\
&\overset{(10)}{\leq} \|y_{k+1} - \tilde{x}_k\| \left(\frac{\lambda_{k+1}}{p!} \left(L_p + \frac{\gamma}{1-\gamma}((p+1)H_p + L_p) \right) \|y_{k+1} - \tilde{x}_k\|^{p-1} \right) \\
&+ \|y_{k+1} - \tilde{x}_k\| \left(1 - \frac{\lambda_{k+1}(p+1)H_p}{p!} \|y_{k+1} - \tilde{x}_k\|^{p-1} \right) \\
&= \|y_{k+1} - \tilde{x}_k\| \left(1 + \frac{\lambda_{k+1}}{p!} \|y_{k+1} - \tilde{x}_k\|^{p-1} \right. \\
&\left. \cdot \left(L_p - (p+1)H_p + \frac{\gamma}{1-\gamma}((p+1)H_p + L_p) \right) \right).
\end{aligned}
$$

Hence, by (10) and simple calculations we get

$$
\begin{aligned}
\sigma &\geq 1 + \frac{1}{2pH_p} \left(L_p - (p+1)H_p + \frac{\gamma}{1-\gamma}((p+1)H_p + L_p) \right) \\
&= 1 + \frac{1}{2p\xi} \left(1 - (p+1)\xi + \frac{\gamma}{1-\gamma}((p+1)\xi + 1) \right) \\
&= 1 + \frac{1}{2p\xi} \left(1 - p\xi - \xi + \frac{\gamma p\xi + \gamma\xi + \gamma}{1-\gamma} \right) \\
&= 1 + \frac{1}{2p\xi} \left(\frac{1 - p\xi - \xi - \gamma + \gamma p\xi + \gamma\xi + \gamma p\xi + \gamma\xi + \gamma}{1-\gamma} \right) \\
&= 1 + \left(\frac{1 - p\xi - \xi + 2\gamma p\xi + 2\gamma\xi}{(1-\gamma)2p\xi} \right) \\
&= \frac{p\xi + 1 - \xi + 2\gamma\xi}{(1-\gamma)2p\xi}.
\end{aligned}
$$

Lastly, we prove that $\sigma \leq 1$. For that we need

$$
\begin{aligned}
(1-\gamma)2p\xi &\geq p\xi + 1 - \xi + 2\gamma\xi \\
(p+1)\xi &\geq 1 + 2\gamma\xi(1+p) \\
\frac{1}{2} - \frac{1}{2\xi(p+1)} &\geq \gamma.
\end{aligned}
$$

We have proved the main lemma for the convergence rate theorem, other parts of the proof are the same as [4]. As a result, we get the next theorem.

Theorem 1. *Let f be a convex function whose p^{th} derivative is L_p-Lipschitz and x_* denote a minimizer of f. Then Algorithm 1 converges with rate*

$$
f(y_k) - f(x_*) \leq \tilde{O}\left(\frac{H_p R^{p+1}}{k^{\frac{3p+1}{2}}} \right), \tag{15}
$$

where

$$R = \|x_0 - x^*\| \tag{16}$$

is the maximal radius of the initial set.

4 Hyperfast Second-Order Method

In recent work [25] it was mentioned that for convex optimization problem (1) with first order oracle (returns gradient) the well-known complexity bound $\left(L_1 R^2/\varepsilon\right)^{1/2}$ can not be beaten even if we assume that all $L_p < \infty$. This is because of the structure of the worth case function

$$f_p(x) = |x_1|^{p+1} + |x_2 - x_1|^{p+1} + ... + |x_n - x_{n-1}|^{p+1},$$

where $p = 1$ for first order method. It's obvious that $f_p(x)$ satisfy the condition $L_p < \infty$ for all natural p. So additional smoothness assumptions don't allow to accelerate additionally. The same thing takes place, for example, for $p = 3$. In this case, we also have $L_p < \infty$ for all natural p. But what is about $p = 2$? In this case $L_3 = \infty$. It means that $f_2(x)$ couldn't be the proper worth case function for the second-order method with additional smoothness assumptions. So there appears the following question: Is it possible to improve the bound $\left(L_2 R^3/\varepsilon\right)^{2/7}$? At the very beginning of 2020 Yu. Nesterov gave a positive answer. For this purpose, he proposed to use an accelerated third-order method that requires $\tilde{O}\left((L_3 R^4/\varepsilon)^{1/4}\right)$ iterations by using second-order oracle [23]. So all this means that if $L_3 < \infty$, then there are methods that can be much faster than $\tilde{O}\left(\left(L_2 R^3/\varepsilon\right)^{2/7}\right)$.

In this section, we improve convergence speed and reach near-optimal speed up to logarithmic factor. We consider problem (1) with $p = 3$, hence $L_3 < \infty$. In previous section, we have proved that Algorithm 1 converges. Now we fix the parameters for this method

$$p = 3, \quad \gamma = \frac{1}{2p} = \frac{1}{6}, \quad \xi = \frac{2p}{p+1} = \frac{3}{2}. \tag{17}$$

By (12) we get $\sigma = 0.6$ that is rather close to initial exact $\sigma_0 = 0.5$. For such parameters we get next convergence speed of Algorithm 1 to reach accuracy ε:

$$N_{out} = \tilde{O}\left(\left(\frac{L_3 R^4}{\varepsilon}\right)^{\frac{1}{5}}\right). \tag{18}$$

Note, that at every step of Algorithm 1 we need to solve next subproblem with accuracy $\gamma = 1/6$

$$\begin{aligned} \underset{y}{\operatorname{argmin}} \Big\{ & \langle \nabla f(x_i), y - x_i \rangle + \frac{1}{2} \nabla^2 f(x_i)[y - x_i]^2 \\ & + \frac{1}{6} D^3 f(x_i)[y - x_i]^3 + \frac{L_3}{4} \|y - x_i\|^4 \Big\}. \end{aligned} \tag{19}$$

In [14] it was proved, that problem (19) can be solved by Bregman-Distance Gradient Method (BDGM) with linear convergence speed. According to [25] BDGM can be improved to work with inexact gradients of the functions. This made possible to approximate $D^3 f(x)$ by gradients and escape calculations of $D^3 f(x)$ at each step. As a result, in [25] it was proved, that subproblem (19) can be solved up to accuracy $\gamma = 1/6$ with one calculation of Hessian and $O\left(\log\left(\frac{\|\nabla f(x_i)\|+\|\nabla^2 f(x_i)\|}{\varepsilon}\right)\right)$ calculation of gradient.

We use BDGM to solve subproblem from Algorithm 1 and, as a result, we get next Hyperfast Second-Order method as merging NATMI and BDGM.

Algorithm 2. Hyperfast Second-Order Method

Input: convex function $f : \mathbb{R}^n \to \mathbb{R}$ with L_3-Lipschitz 3rd-order derivative.
Set $A_0 = 0, x_0 = y_0$
for $k = 0$ **to** $k = K - 1$ **do**
  Compute a pair $\lambda_{k+1} > 0$ and $y_{k+1} \in \mathbb{R}^n$ such that

$$\frac{1}{2} \leq \lambda_{k+1} \frac{3L_3 \cdot \|y_{k+1} - \tilde{x}_k\|^2}{4} \leq \frac{3}{4},$$

  where $y_{k+1} \in \mathcal{N}^{1/6}_{3,3L_3/2}(\tilde{x}_k)$ solved by Algorithm 3 and

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1} A_k}}{2}, \; A_{k+1} = A_k + a_{k+1}$$

$$\tilde{x}_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_{k+1}}{A_{k+1}} x_k .$$

  Update $x_{k+1} := x_k - a_{k+1} \nabla f(y_{k+1})$
return y_K

In the Algorithm 3, $\beta_{\rho_k}(z_i, z)$ is a Bregman distance generated by $\rho_k(z)$

$$\beta_{\rho_k}(z_i, z) = \rho_k(z) - \rho_k(z_i) - \langle \nabla \rho_k(z_i), z - z_i \rangle .$$

By $g_{\varphi_k,\tau}(z)$ we take an inexact gradient of the subproblem (19)

$$g_{\varphi_k,\tau}(z) = \nabla f(\tilde{x}_k) + \nabla^2 f(\tilde{x}_k)[z - \tilde{x}_k] + \frac{1}{2} g^{\tau}_{\tilde{x}_k}(z) + L_3 \|z - \tilde{x}_k\|^2 (z - \tilde{x}_k) \quad (22)$$

and $g^{\tau}_{\tilde{x}_k}(z)$ is a inexact approximation of $D^3 f(\tilde{x}_k)[y - \tilde{x}_k]^2$

$$g^{\tau}_{\tilde{x}_k}(z) = \frac{1}{\tau^2} \left(\nabla f(\tilde{x}_k + \tau(z - \tilde{x}_k)) + \nabla f(\tilde{x}_k - \tau(z - \tilde{x}_k)) - 2\nabla f(\tilde{x}_k)\right). \quad (23)$$

In paper [25] it is proved, that we can choose

$$\delta = O\left(\frac{\varepsilon^{\frac{3}{2}}}{\|\nabla f(\tilde{x}_k)\|_*^{\frac{1}{2}} + \|\nabla^2 f(\tilde{x}_k)\|^{\frac{3}{2}} / L_3^{\frac{1}{2}}}\right),$$

Algorithm 3. Bregman-Distance Gradient Method

1: Set $z_0 = \tilde{x}_k$ and $\tau = \frac{3\delta}{8(2+\sqrt{2})\|\nabla f(\tilde{x}_k)\|}$

2: Set objective function

$$\varphi_k(z) = \langle \nabla f(\tilde{x}_k), z - \tilde{x}_k \rangle + \frac{1}{2}\nabla^2 f(\tilde{x}_k)[z - \tilde{x}_k]^2 + \frac{1}{6}D^3 f(\tilde{x}_k)[z - \tilde{x}_k]^3 + \frac{L_3}{4}\|z - \tilde{x}_k\|^4$$

3: Set feasible set

$$S_k = \left\{ z : \|z - \tilde{x}_k\| \leq 2\left(\tfrac{2+\sqrt{2}}{L_3}\|\nabla f(\tilde{x}_k)\|\right)^{\frac{1}{3}} \right\} \tag{20}$$

4: Set scaling function

$$\rho_k(z) = \tfrac{1}{2}\left\langle \nabla^2 f(\tilde{x}_k)(z - \tilde{x}_k),\, z - \tilde{x}_k \right\rangle + \tfrac{L_3}{4}\|z - \tilde{x}_k\|^4 \tag{21}$$

5: **for** $k \geq 0$ **do**

6: Compute the approximate gradient $g_{\varphi_k,\tau}(z_i)$ by (22).

7: **IF** $\|g_{\varphi_k,\tau}(z_i)\| \leq \frac{1}{6}\|\nabla f(z_i)\| - \delta$, then **STOP**

8: **ELSE** $z_{i+1} = \underset{z \in S_k}{\operatorname{argmin}} \left\{ \langle g_{\varphi_k,\tau}(z_i), z - z_i \rangle + 2\left(1 + \frac{1}{\sqrt{2}}\right)\beta_{\rho_k}(z_i, z) \right\}$,

9: **return** z_i

then total number of inner iterations equal to

$$T_k(\delta) = O\left(\ln \frac{G + H}{\varepsilon}\right), \tag{24}$$

where G and H are the uniform upper bounds for the norms of the gradients and Hessians computed at the points generated by the main algorithm. Finally, we get next theorem.

Theorem 2. *Let f be a convex function whose third derivative is L_3-Lipschitz and x_* denote a minimizer of f. Then to reach accuracy ε Algorithm 2 with Algorithm 3 for solving subproblem computes*

$$N_1 = \tilde{O}\left(\left(\frac{L_3 R^4}{\varepsilon}\right)^{\frac{1}{5}}\right) \tag{25}$$

Hessians and

$$N_2 = \tilde{O}\left(\left(\frac{L_3 R^4}{\varepsilon}\right)^{\frac{1}{5}} \log\left(\frac{G + H}{\varepsilon}\right)\right) \tag{26}$$

gradients, where G and H are the uniform upper bounds for the norms of the gradients and Hessians computed at the points generated by the main algorithm.

One can generalize this result on uniformly-strongly convex functions by using inverse restart-regularization trick from [13].

So, the main observation of this section is as follows: If $L_3 < \infty$, then we can use this hyperfast[3] second-order algorithm instead of considered in the paper optimal one to make our sliding faster (in convex and uniformly convex cases).

[3] Here we use terminology introduced in [25].

5 Conclusion

In this paper, we present Inexact Near-optimal Accelerated Tensor Method and improve its convergence rate. This improvement make it possible to solve the Taylor approximation subproblem by other methods. Next, we propose Hyperfast Second-Order Method and get its convergence speed $O(N^{-5})$ up to logarithmic factor. This method is a combination of Inexact Third-Order Near-Optimal Accelerated Tensor Method with Bregman-Distance Gradient Method for solving inner subproblem. As a result, we prove that our method has near-optimal convergence rates for given problem class and the best known on that moment.

In this paper, we developed near-optimal Hyperfast Second-Order method for sufficiently smooth convex problem in terms of convergence in function. Based on the technique from the work [9], we can also developed near-optimal Hyperfast Second-Order method for sufficiently smooth convex problem in terms of convergence in the norm of the gradient. In particular, based on the work [16] one may show that the complexity of this approach to the dual problem for 1-entropy regularized optimal transport problem will be $\tilde{O}\left(\left((\sqrt{n})^4/\varepsilon\right)^{1/5}\right) \cdot O(n^{2.5}) = O(n^{2.9}\varepsilon^{-1/5})$ a.o., where n is the linear dimension of the transport plan matrix, that could be better than the complexity of accelerated gradient method and accelerated Sinkhorn algorithm $O(n^{2.5}\varepsilon^{-1/2})$ a.o. [8,16]. Note, that the best theoretical bounds for this problem are also far from to be practical ones [2,17,20,27].

Acknowledgements. I would like to thank Alexander Gasnikov, Yurii Nesterov, Pavel Dvurechensky and Cesar Uribe for fruitful discussions.

References

1. Birgin, E.G., Gardenghi, J., Martínez, J.M., Santos, S.A., Toint, L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. **163**(1–2), 359–368 (2017)
2. Blanchet, J., Jambulapati, A., Kent, C., Sidford, A.: Towards optimal running times for optimal transport. arXiv preprint arXiv:1810.07717 (2018)
3. Bubeck, S., Jiang, Q., Lee, Y.T., Li, Y., Sidford, A.: Complexity of highly parallel non-smooth convex optimization. In: Advances in Neural Information Processing Systems, pp. 13900–13909 (2019)
4. Bubeck, S., Jiang, Q., Lee, Y.T., Li, Y., Sidford, A.: Near-optimal method for highly smooth convex optimization. In: Conference on Learning Theory, pp. 492–507 (2019)
5. Carmon, Y., Duchi, J., Hinder, O., Sidford, A.: Lower bounds for finding stationary points II: first-order methods. arXiv preprint arXiv:1711.00841 (2017)
6. Doikov, N., Nesterov, Y.: Contracting proximal methods for smooth convex optimization. arXiv preprint arXiv:1912.07972 (2019)
7. Doikov, N., Nesterov, Y.: Local convergence of tensor methods. arXiv preprint arXiv:1912.02516 (2019)
8. Dvurechensky, P., Gasnikov, A., Kroshnin, A.: Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. arXiv preprint arXiv:1802.04367 (2018)

9. Dvurechensky, P., Gasnikov, A., Ostroukhov, P., Uribe, C.A., Ivanova, A.: Near-optimal tensor methods for minimizing the gradient norm of convex function. arXiv preprint arXiv:1912.03381 (2019)
10. Gasnikov, A.: Universal gradient descent. arXiv preprint arXiv:1711.00394 (2017)
11. Gasnikov, A., Dvurechensky, P., Gorbunov, E., Vorontsova, E., Selikhanovych, D., Uribe, C.A.: Optimal tensor methods in smooth convex and uniformly convex optimization. In: Conference on Learning Theory, pp. 1374–1391 (2019)
12. Gasnikov, A., et al.: Near optimal methods for minimizing convex functions with Lipschitz p-th derivatives. In: Conference on Learning Theory, pp. 1392–1393 (2019)
13. Gasnikov, A.V., Kovalev, D.A.: A hypothesis about the rate of global convergence for optimal methods (Newton's type) in smooth convex optimization. Comput. Res. Model. **10**(3), 305–314 (2018)
14. Grapiglia, G.N., Nesterov, Y.: On inexact solution of auxiliary problems in tensor methods for convex optimization. arXiv preprint arXiv:1907.13023 (2019)
15. Grapiglia, G.N., Nesterov, Y.: Tensor methods for minimizing functions with hölder continuous higher-order derivatives. arXiv preprint arXiv:1904.12559 (2019)
16. Guminov, S., Dvurechensky, P., Nazary, T., Gasnikov, A.: Accelerated alternating minimization, accelerated Sinkhorn's algorithm and accelerated iterative Bregman projections. arXiv preprint arXiv:1906.03622 (2019)
17. Jambulapati, A., Sidford, A., Tian, K.: A direct tilde {O}(1/epsilon) iteration parallel algorithm for optimal transport. In: Advances in Neural Information Processing Systems, pp. 11355–11366 (2019)
18. Jiang, B., Wang, H., Zhang, S.: An optimal high-order tensor method for convex optimization. In: Conference on Learning Theory, pp. 1799–1801 (2019)
19. Kamzolov, D., Gasnikov, A., Dvurechensky, P.: On the optimal combination of tensor optimization methods. arXiv preprint arXiv:2002.01004 (2020)
20. Lee, Y.T., Sidford, A.: Solving linear programs with Sqrt (rank) linear system solves. arXiv preprint arXiv:1910.08033 (2019)
21. Monteiro, R.D., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIAM J. Optim. **23**(2), 1092–1125 (2013)
22. Nesterov, Y.: Lectures on Convex Optimization. SOIA, vol. 137. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91578-4
23. Nesterov, Y.: Implementable tensor methods in unconstrained convex optimization. Math. Program., 1–27 (2019). https://doi.org/10.1007/s10107-019-01449-1
24. Nesterov, Y.: Inexact accelerated high-order proximal-point methods. Technical report, Technical Report CORE Discussion paper 2020, Université catholique de Louvain, Center for Operations Research and Econometrics (2020)
25. Nesterov, Y.: Superfast second-order methods for unconstrained convex optimization. Technical report, Technical Report CORE Discussion paper 2020, Université catholique de Louvain, Center for Operations Research and Econometrics (2020)
26. Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Math. Program. **108**(1), 177–205 (2006)
27. Quanrud, K.: Approximating optimal transport with linear programs. arXiv preprint arXiv:1810.05957 (2018)

28. Song, C., Ma, Y.: Towards unified acceleration of high-order algorithms under hölder continuity and uniform convexity. arXiv preprint arXiv:1906.00582 (2019)
29. Wang, Z., Zhou, Y., Liang, Y., Lan, G.: Cubic regularization with momentum for nonconvex optimization. In: Proceedings of the Uncertainty in Artificial Intelligence (UAI) Conference (2019)
30. Wilson, A., Mackey, L., Wibisono, A.: Accelerating rescaled gradient descent. arXiv preprint arXiv:1902.08825 (2019)