



2019/21

DP

Nikita Doikov and Yurii Nesterov

Local convergence of tensor methods



CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: immaq-library@uclouvain.be

[https://uclouvain.be/en/research-institutes/
lidam/core/discussion-papers.html](https://uclouvain.be/en/research-institutes/lidam/core/discussion-papers.html)

CORE DISCUSSION PAPER

2019/21

Local convergence of tensor methods ^{*}

Nikita Doikov [†] and Yurii Nesterov [‡]

November 26, 2019

Abstract

In this paper, we study local convergence of high-order Tensor Methods for solving convex optimization problems with composite objective. We justify local superlinear convergence under the assumption of uniform convexity of the smooth component, having Lipschitz-continuous high-order derivative. The convergence both in function value and in the norm of minimal subgradient is established. Global complexity bounds for the Composite Tensor Method in convex and uniformly convex cases are also discussed. Lastly, we show how local convergence of the methods can be globalized using the inexact proximal iterations.

Keywords: convex optimization, high-order methods, tensor methods, local convergence, uniform convexity, proximal methods

^{*}The research results of this paper were obtained in the framework of ERC Advanced Grant 788368.

[†]Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Catholic University of Louvain (UCL); e-mail: Nikita.Doikov@uclouvain.be.

[‡]Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; e-mail: Yurii.Nesterov@uclouvain.be.

1 Introduction

Motivation. In Nonlinear Optimization, it seems to be a natural idea to increase the performance of numerical methods by employing high-order oracles. However, the main obstacle to this approach consists in a prohibiting complexity of the corresponding Taylor approximations formed by the high-order multidimensional polynomials, which are difficult to store, handle, and minimize. If we go just one step above the commonly used quadratic approximation, we get a multidimensional polynomial of degree three which is never convex. Consequently, its usefulness for optimization methods is questionable.

However, recently in [13] it was shown that the Taylor polynomials of *convex functions* have a very interesting structure. It appears that their augmentation by a power of Euclidean norm with a reasonably big coefficients gives us a global upper *convex* model of the objective function, which keeps all advantages of the local high-order approximation.

One of the classical and well-known results in Nonlinear Optimization is related to the local quadratic convergence of Newton's method [8, 14]. Later on, it was generalized to the case of *composite* optimization problems [9], where the objective is represented as a sum of two convex components: smooth, and possibly nonsmooth but simple.

A big step in the second-order optimization theory was made since [16], where Cubic regularization of the Newton method with its global complexity estimates was proposed. Additionally, the local superlinear convergence was justified.

Our paper is aimed to study local convergence of high-order methods, generalizing corresponding results from [16] in several ways. We establish local superlinear convergence of Tensor Method [13] of degree $p \geq 2$, in the case when the objective is composite, and its smooth part is uniformly convex of arbitrary degree q from the interval $2 \leq q < p - 1$. For strongly convex functions ($q = 2$), this gives the local convergence of degree p .

Contents. We formulate our problem of interest and define a step of the Regularized Composite Tensor Method in Section 2. Then, we declare some of its properties, which are required for our analysis.

In Section 3, we prove local superlinear convergence of the Tensor Method in function value, and in the norm of minimal subgradient, under the assumption of uniform convexity of the objective.

In Section 4, we discuss global behavior of the method and justify sublinear and linear global rates of convergence for convex and uniformly convex cases, respectively.

One application of our developments is provided in Section 5. We show how local convergence can be applied for computing an inexact step in proximal methods. A global sublinear rate of convergence for the resulting scheme is also given.

Notations and generalities. In what follows, we denote by $\mathbb{E}$ a finite-dimensional real vector space, and by $\mathbb{E}^*$ its dual space composed by linear functions on $\mathbb{E}$. For such a function $s \in \mathbb{E}^*$, we denote by $\langle s, x \rangle$ its value at $x \in \mathbb{E}$. Using a self-adjoint positive-definite operator $B : \mathbb{E} \rightarrow \mathbb{E}^*$ (notation $B = B^* \succ 0$), we can endow these spaces with mutually conjugate Euclidean norms:

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

For a smooth function $f : \text{dom } f \rightarrow \mathbb{R}$ with convex and open domain $\text{dom } f \subseteq \mathbb{E}$, denote by $\nabla f(x)$ its gradient, and by $\nabla^2 f(x)$ its Hessian evaluated at point $x \in \text{dom } f \subseteq \mathbb{E}$. Note

that

$$\nabla f(x) \in \mathbb{E}^*, \quad \nabla^2 f(x)h \in \mathbb{E}^*, \quad x \in \text{dom } f, \quad h \in \mathbb{E}.$$

For non-differentiable convex function $f(\cdot)$, we denote by $\partial f(x) \subset \mathbb{E}^*$ its subdifferential at the point $x \in \text{dom } f$.

In what follows, we often work with directional derivatives. For $p \geq 1$, denote by

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

the directional derivative of function f at x along directions $h_i \in \mathbb{E}$, $i = 1, \dots, p$. If all directions $h_1, \dots, h_p$ are the same, we apply a simpler notation

$$D^p f(x)[h]^p, \quad h \in \mathbb{E}.$$

Note that $D^p f(x)[\cdot]$ is a *symmetric p -linear form*. Its *norm* is defined in the standard way:

$$\begin{aligned} \|D^p f(x)\| &= \max_{h_1, \dots, h_p \in \mathbb{E}} \{D^p f(x)[h_1, \dots, h_p] : \|h_i\| \leq 1, i = 1, \dots, p\} \\ &= \max_{h \in \mathbb{E}} \left\{ \left| D^p f(x)[h]^p \right| : \|h\| \leq 1 \right\} \end{aligned} \quad (1.1)$$

(for the last equation see, for example, Appendix 1 in [15]). Similarly, we define

$$\|D^p f(x) - D^p f(y)\| = \max_{h \in \mathbb{E}} \left\{ \left| D^p f(x)[h]^p - D^p f(y)[h]^p \right| : \|h\| \leq 1 \right\}. \quad (1.2)$$

In particular, for any $x \in \text{dom } f$ and $h_1, h_2 \in \mathbb{E}$, we have

$$Df(x)[h_1] = \langle \nabla f(x), h_1 \rangle, \quad D^2 f(x)[h_1, h_2] = \langle \nabla^2 f(x)h_1, h_2 \rangle.$$

Thus, for the Hessian, our definition corresponds to a *spectral norm* of the self-adjoint linear operator (maximal module of all eigenvalues computed with respect to $B \succ 0$).

Finally, the Taylor approximation of function $f(\cdot)$ at $x \in \text{dom } f$ is defined as follows:

$$f(x+h) = \Omega_p(f, x; h) + o(\|h\|^p), \quad x+h \in \text{dom } f,$$

$$\Omega_p(f, x; y) \stackrel{\text{def}}{=} f(x) + \sum_{k=1}^p \frac{1}{k!} D^k f(x)[y-x]^k, \quad y \in \mathbb{E}.$$

Consequently, for all $y \in \mathbb{E}$ we have

$$\nabla \Omega_p(f, x; y) = \sum_{k=1}^p \frac{1}{(k-1)!} D^k f(x)[y-x]^{k-1}, \quad (1.3)$$

$$\nabla^2 \Omega_p(f, x; y) = \sum_{k=2}^p \frac{1}{(k-2)!} D^k f(x)[y-x]^{k-2}. \quad (1.4)$$

2 Main inequalities

In this paper, we consider the following *composite* minimization problem

$$\min_{x \in \text{dom } h} \left\{ F(x) = f(x) + h(x) \right\}, \quad (2.1)$$

where $h : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function and $f \in C^{p,p}(\text{dom } h)$ for a certain $p \geq 2$. In other words, we assume that the p th derivative of function f is Lipschitz continuous:

$$\|D^p f(x) - D^p f(y)\| \leq L_p \|x - y\|, \quad x, y \in \text{dom } h. \quad (2.2)$$

Assuming that $L_p < +\infty$, by the standard integration arguments we can bound the residual between function value and its Taylor approximation:

$$|f(y) - \Omega_p(f, x; y)| \leq \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad x, y \in \text{dom } h. \quad (2.3)$$

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

$$\|\nabla f(y) - \nabla \Omega_p(f, x; y)\|_* \leq \frac{L_p}{p!} \|y - x\|^p, \quad (2.4)$$

$$\|\nabla^2 f(y) - \nabla^2 \Omega_p(f, x; y)\| \leq \frac{L_p}{(p-1)!} \|y - x\|^{p-1}, \quad (2.5)$$

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

Let us define now one step of the *Regularized Composite Tensor Method* (RCTM) of degree $p \geq 2$:

$$T \equiv T_H(x) \stackrel{\text{def}}{=} \arg \min_{y \in \mathbb{E}} \left\{ \Omega_p(f, x; y) + \frac{H}{(p+1)!} \|y - x\|^{p+1} + h(y) \right\}. \quad (2.6)$$

It can be shown that for

$$\boxed{H \geq pL_p} \quad (2.7)$$

the auxiliary optimization problem in (2.6) is *convex* (see Theorem 1 in [13]). This condition is crucial for implementability of our methods and we always assume it to be satisfied.

Let us write down the first-order optimality condition for the auxiliary optimization problem in (2.6):

$$\langle \nabla \Omega_p(f, x; T) + \frac{H}{p!} \|T - x\|^{p-1} B(T - x), y - T \rangle + h(y) \geq h(T), \quad y \in \text{dom } h. \quad (2.8)$$

In other words, for vector

$$h'(T) \stackrel{\text{def}}{=} - \left(\nabla \Omega_p(f, x; T) + \frac{H}{p!} \|T - x\|^{p-1} B(T - x) \right) \quad (2.9)$$

we have $h'(T) \stackrel{(2.8)}{\in} \partial h(T)$. This fact explains our notation

$$F'(T) \stackrel{\text{def}}{=} \nabla f(T) + h'(t) \in \partial F(T). \quad (2.10)$$

Let us present some properties of the point $T = T_H(x)$. First of all, we need some bounds for the norm of vector $F'(T)$. Note that

$$\begin{aligned} \left\| F'(T) + \frac{H}{p!} \|T - x\|^{p-1} B(T - x) \right\|_* &\stackrel{(2.9)}{=} \left\| \nabla f(T) - \nabla \Omega_p(f, x; T) \right\|_* \\ &\stackrel{(2.4)}{\leq} \frac{L_p}{p!} \|T - x\|^p. \end{aligned} \quad (2.11)$$

Consequently,

$$\|F'(T)\|_* \leq \frac{L_p + H}{p!} \|T - x\|^p. \quad (2.12)$$

Secondly, we need two lemmas characterizing the gradients at point T .

Lemma 1 *For any $x \in \text{dom } h$ we have*

$$h(x) \geq h(T) + \langle \nabla f(T), T - x \rangle. \quad (2.13)$$

Proof:

Indeed,

$$\begin{aligned} -h'(T) &\stackrel{(2.9)}{=} \nabla \Omega_p(f, x; T) + \frac{H}{p!} \|T - x\|^{p-1} B(T - x) \\ &\stackrel{\text{def}}{=} \nabla f(x) + G_{x,p}(T - x), \end{aligned}$$

where $G_{x,p} = \nabla^2 f(x) + \sum_{k=3}^p \frac{1}{(k-1)!} D^{(k)} f(x) [T - x]^{k-2} + \frac{H}{p!} \|T - x\|^{p-1} B$. Note that

$$\begin{aligned} G_{x,p} &= \int_0^1 \left(\nabla^2 f(x) + \sum_{k=3}^p \frac{\tau^{k-2}}{(k-2)!} D^{(k)} f(x) [T - x]^{k-2} + \frac{H\tau^{p-1}}{(p-1)!} \|T - x\|^{p-1} B \right) d\tau \\ &\stackrel{(1.4)}{=} \int_0^1 \left(\nabla^2 \Omega_p(f, x; x + \tau(T - x)) + \frac{H\tau^{p-1}}{(p-1)!} \|T - x\|^{p-1} B \right) d\tau \\ &\stackrel{(2.5)}{\succeq} \int_0^1 \nabla^2 f(x + \tau(T - x)) d\tau. \end{aligned}$$

Therefore, applying the condition (2.8) with $y = x$, we get

$$\begin{aligned} h(T) &\leq h(x) + \langle \nabla f(x) + G_{x,p}(T - x), x - T \rangle \\ &\leq h(x) + \langle \nabla f(x), x - T \rangle - \left\langle \int_0^1 \nabla^2 f(x + \tau(T - x)) d\tau (T - x), T - x \right\rangle \\ &= h(x) + \langle \nabla f(x), x - T \rangle - \langle \nabla f(T) - \nabla f(x), T - x \rangle \\ &= h(x) - \langle \nabla f(T), T - x \rangle. \end{aligned} \quad \square$$

Lemma 2 Let $\beta > 1$ and $H = \beta L_p$. Then

$$\langle F'(T), x - T \rangle \geq \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}} \cdot \|F'(T)\|_*^{\frac{p+1}{p}} \cdot \frac{(\beta^2-1)^{\frac{p-1}{2p}}}{\beta} \cdot \frac{p}{(p^2-1)^{\frac{p-1}{2p}}}. \quad (2.14)$$

In particular, if $\beta = p$, then

$$\langle F'(T), x - T \rangle \geq \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}} \cdot \|F'(T)\|_*^{\frac{p+1}{p}}. \quad (2.15)$$

Proof:

Denote $r = \|T - x\|$, $h = \frac{H}{p!}$, and $l = \frac{L_p}{p!}$. Then inequality (2.11) can be written as follows:

$$\|F'(T) + hr^{p-1}B(T - x)\|_*^2 \leq l^2 r^{2p}.$$

This means that

$$\langle F'(T), x - T \rangle \geq \frac{1}{2hr^{p-1}} \|F'(T)\|_*^2 + \frac{r^{2p}(h^2 - l^2)}{2hr^{p-1}}. \quad (2.16)$$

Denote

$$a = \frac{1}{2h} \|F'(T)\|_*^2, \quad b = \frac{h^2 - l^2}{2h}, \quad \tau = r^{p-1}, \quad \alpha = \frac{p+1}{p-1}.$$

Then inequality (2.16) can be rewritten as follows:

$$\langle F'(T), x - T \rangle \geq \frac{a}{\tau} + b\tau^\alpha \geq \min_{t>0} \left\{ \frac{a}{t} + bt^\alpha \right\} = (1 + \alpha) \left(\frac{a}{\alpha} \right)^{\frac{\alpha}{1+\alpha}} b^{\frac{1}{1+\alpha}}.$$

Taking into account that $1 + \alpha = \frac{2p}{p-1}$ and $\frac{\alpha}{1+\alpha} = \frac{p+1}{2p}$, and using the actual meaning of a , b , and α , we get

$$\begin{aligned} \langle F'(T), x - T \rangle &\geq \frac{2p}{p-1} \cdot \frac{\|F'(T)\|_*^{\frac{p+1}{p}}}{(2h)^{\frac{p+1}{2p}}} \cdot \frac{(p-1)^{\frac{p+1}{2p}}}{(p+1)^{\frac{p+1}{2p}}} \cdot \frac{(h^2 - l^2)^{\frac{p-1}{2p}}}{(2h)^{\frac{p-1}{2p}}} \\ &= \|F'(T)\|_*^{\frac{p+1}{p}} \cdot \frac{(h^2 - l^2)^{\frac{p-1}{2p}}}{h} \cdot \frac{p}{(p+1)^{\frac{p+1}{2p}} (p-1)^{\frac{p-1}{2p}}} \\ &= \|F'(T)\|_*^{\frac{p+1}{p}} \cdot \frac{(h^2 - l^2)^{\frac{p-1}{2p}}}{h} \cdot \frac{p}{(p^2 - 1)^{\frac{p-1}{2p}} (p+1)^{\frac{1}{p}}}. \end{aligned}$$

It remains to note that

$$\frac{(h^2 - l^2)^{\frac{p-1}{2p}}}{h} = \frac{(H^2 - L_p^2)^{\frac{p-1}{2p}}}{H} \cdot (p!)^{\frac{1}{p}} = \frac{(\beta^2 - 1)^{\frac{p-1}{2p}}}{\beta} \cdot \left(\frac{p!}{L_p} \right)^{\frac{1}{p}}.$$

□

3 Local convergence

The main goal of this paper consists in analyzing the local behavior of the *Regularized Composite Tensor Method* (RCTM):

$$x_0 \in \text{dom } h, \quad x_{k+1} = T_H(x_k), \quad k \geq 0, \quad (3.1)$$

as applied to the problem (2.1). In order to prove local superlinear convergence of this scheme, we need one more assumption.

Assumption 1 *The smooth part of the objective function in problem (2.1) is uniformly convex of degree $q \geq 2$:*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \sigma_q \|x - y\|^q, \quad x, y \in \text{dom } h, \quad (3.2)$$

for certain $\sigma_q > 0$.

It is well known that this assumption guarantees the uniform convexity of the objective function in the problem (2.1) (see, for example, Lemma 4.2.1 in [14]):

$$F(y) \geq F(x) + \langle g_x, y - x \rangle + \frac{\sigma_q}{q} \|y - x\|^q, \quad y \in \text{dom } h, \quad (3.3)$$

where g_x is an arbitrary subgradient from $\partial F(x)$. Therefore,

$$\begin{aligned} F^* &= \min_{y \in \text{dom } h} F(y) \geq \min_{y \in \mathbb{E}} \left\{ F(x) + \langle g_x, y - x \rangle + \frac{\sigma_q}{q} \|y - x\|^q \right\} \\ &= F(x) - \frac{q-1}{q} \left(\frac{1}{\sigma_q} \right)^{\frac{1}{q-1}} \|g_x\|_*^{\frac{q}{q-1}}. \end{aligned} \quad (3.4)$$

This simple inequality gives us the following local convergence rate for RCTM.

Theorem 1 *For any $k \geq 0$ we have*

$$F(x_{k+1}) - F^* \leq (q-1)q^{\frac{p-q+1}{q-1}} \left(\frac{1}{\sigma_q} \right)^{\frac{p+1}{q-1}} \left(\frac{L_{p+H}}{p!} \right)^{\frac{q}{q-1}} \left[F(x_k) - F^* \right]^{\frac{p}{q-1}}. \quad (3.5)$$

Proof:

Indeed, for any $k \geq 0$ we have

$$\begin{aligned} F(x_k) - F^* &\geq F(x_k) - F(x_{k+1}) \\ &\stackrel{(3.3)}{\geq} \langle F'(x_{k+1}), x_k - x_{k+1} \rangle + \frac{\sigma_q}{q} \|x_k - x_{k+1}\|^q \\ &\stackrel{(2.14)}{\geq} \frac{\sigma_q}{q} \|x_k - x_{k+1}\|^q \stackrel{(2.12)}{\geq} \frac{\sigma_q}{q} \left(\frac{p!}{L_{p+H}} \|F'(x_{k+1})\|_* \right)^{\frac{q}{p}} \\ &\stackrel{(3.4)}{\geq} \frac{\sigma_q}{q} \left(\frac{p!}{L_{p+H}} \right)^{\frac{q}{p}} \left(\frac{q\sigma_q^{\frac{1}{q-1}}}{q-1} (F(x_{k+1}) - F^*) \right)^{\frac{q-1}{p}}. \end{aligned}$$

And this is exactly inequality (3.5). $\square$

Corollary 1 *If $p > q - 1$, then method (3.1) has local superlinear rate of convergence for problem (2.1).*

Proof:

Indeed, in this case $\frac{p}{q-1} > 1$. □

For example, if $q = 2$ (strongly convex function) and $p = 2$ (Cubic Regularization of the Newton Method), then the rate of convergence is quadratic. If $q = 2$, and $p = 3$, then the local rate of convergence is cubic, etc.

Let us study now the local convergence of the method (3.1) in terms of the norm of gradient. For any $x \in \text{dom } h$ denote

$$\eta(x) \stackrel{\text{def}}{=} \min_{g \in \partial h(x)} \|\nabla f(x) + g\|_*. \quad (3.6)$$

If $\partial h(x) = \emptyset$, we set $\eta(x) = +\infty$.

Theorem 2 *For any $k \geq 0$ we have*

$$\eta(x_{k+1}) \leq \|F'(x_{k+1})\|_* \leq \frac{L_p + H}{p!} \left[\frac{1}{\sigma_q} \eta(x_k) \right]^{\frac{p}{q-1}}. \quad (3.7)$$

Proof:

Indeed, in view of inequality (2.13), we have

$$\begin{aligned} \langle \nabla f(x_k), x_k - x_{k+1} \rangle &= \langle \nabla f(x_k) - \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \\ &\geq \langle \nabla f(x_k) - \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + h(x_{k+1}) - h(x_k) \\ &\stackrel{(3.2)}{\geq} \sigma_q \|x_k - x_{k+1}\|^q + h(x_{k+1}) - h(x_k) \\ &\geq \sigma_q \|x_k - x_{k+1}\|^q + \langle g_k, x_{k+1} - x_k \rangle, \end{aligned}$$

where g_k is an arbitrary vector from $\partial h(x_k)$. Therefore,

$$\sigma_q \|x_k - x_{k+1}\|^q \leq \langle \nabla f(x_k) + g_k, x_k - x_{k+1} \rangle,$$

and we conclude that $\eta(x_k) \geq \sigma_q \|x_k - x_{k+1}\|^{q-1}$. It remains to use inequality (2.12). □

As we can see, the condition for superlinear convergence of the method (3.1) in terms of the norm of the gradient is the same as in Corollary 1: we need to have $\frac{p}{q-1} > 1$, that is $p > q - 1$. Moreover, the local rate of convergence has the same order as that for the residual of the function value.

According to Theorem 1, the region of superlinear convergence of RCTM in terms of the function value is as follows:

$$\mathcal{Q} = \left\{ x \in \text{dom } h : F(x) - F^* \leq \frac{1}{q} \cdot \left(\frac{\sigma_q^{p+1}}{(q-1)^{q-1}} \cdot \left(\frac{p!}{L_p + H} \right)^q \right)^{\frac{1}{p-q+1}} \right\}. \quad (3.8)$$

Alternatively, by Theorem 2, in terms of the norm of minimal subgradient (3.6), the region of quadratic convergence looks as follows:

$$\mathcal{G} = \left\{ x \in \text{dom } h : \eta(x) \leq \left(\sigma_q^p \cdot \left(\frac{p!}{L_p + H} \right)^{q-1} \right)^{\frac{1}{p-q+1}} \right\}. \quad (3.9)$$

Note that these sets can be very different. Indeed, set $\mathcal{Q}$ is a closed and convex neighborhood of the point x^* . At the same time, the structure of the set $\mathcal{G}$ can be very complex since in general the function $\eta(x)$ is discontinuous. Let us look at simple example where $h(x) = \text{Ind}_Q(x)$, the indicator function of a closed convex set Q .

Example 1 Consider the following optimization problem:

$$\min_{x \in \mathbb{R}^2} \left\{ x^{(2)} : \|x\|^2 \stackrel{\text{def}}{=} (x^{(1)})^2 + (x^{(2)})^2 \leq 1 \right\}. \quad (3.10)$$

Clearly, in this problem $x^* = (0, -1)$ and it can be written in the form (2.1) with

$$f(x) = x^{(2)}, \quad h(x) = \begin{cases} +\infty, & \text{if } \|x\| > 1, \\ 0, & \text{otherwise.} \end{cases}$$

Note that for $x \in \text{dom } h \equiv \{x : \|x\| \leq 1\}$, we have

$$\partial h(x) = \begin{cases} 0, & \text{if } \|x\| < 1, \\ \{\gamma x, \gamma \geq 0\}, & \text{if } \|x\| = 1. \end{cases}$$

Therefore, if $\|x\| < 1$, then $\eta(x) = \|\nabla f(x)\| = 1$. If $\|x\| = 1$, then

$$\begin{aligned} \eta^2(x) &\stackrel{(3.6)}{=} \min_{\gamma \geq 0} \{(\gamma x^{(1)})^2 + (1 + \gamma x^{(2)})^2\} = \min_{\gamma \geq 0} \{1 + 2\gamma x^{(2)} + \gamma^2\} \\ &= 1 - (-x^{(2)})_+^2, \end{aligned}$$

where $(a)_+ = \max\{a, 0\}$. Thus, in any neighborhood of x^* , $\eta(x)$ vanishes only along the boundary of the feasible set. $\square$

Thus, the question arises how the Tensor Method (3.1) could come to the region $\mathcal{G}$. The answer follows from the inequalities derived in Section 2. Indeed,

$$\|F'(x_{k+1})\|_* \stackrel{(2.12)}{\leq} \frac{L_p + H}{p!} \|x_k - x_{k+1}\|^p,$$

$$F(x_k) - F(x_{k+1}) \geq \langle F'(x_{k+1}), x_k - x_{k+1} \rangle \stackrel{(2.15)}{\geq} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}} \cdot \|F'(x_{k+1})\|_*^{\frac{p+1}{p}}.$$

Thus, at some moment the norm $\|F'(x_k)\|_*$ will be small enough to enter $\mathcal{G}$.

4 Global complexity bounds

Let us briefly discuss the global complexity bounds of the method (3.1), namely the number of iterations required for coming from an arbitrary initial point $x_0 \in \text{dom } h$ to the region $\mathcal{Q}$. First, note that for every step $T = T_H(x)$ of the method with parameter $H \geq pL_p$, we have

$$\begin{aligned} F(T) &\stackrel{(2.3)}{\leq} \Omega_p(f, x; T) + \frac{H}{(p+1)!} \|T - x\|^{p+1} + h(T) \\ &\stackrel{(2.6)}{=} \min_{y \in \mathbb{E}} \left\{ \Omega_p(f, x; y) + \frac{H}{(p+1)!} \|y - x\|^{p+1} + h(y) \right\} \\ &\stackrel{(2.3)}{\leq} \min_{y \in \mathbb{E}} \left\{ F(y) + \frac{H+L_p}{(p+1)!} \|y - x\|^{p+1} \right\}. \end{aligned}$$

Therefore,

$$F(T(x)) - F^* \leq \frac{H+L_p}{(p+1)!} \|x - x^*\|^{p+1}, \quad x \in \text{dom } h, \quad (4.1)$$

with $x^* \stackrel{\text{def}}{=} \arg \min_{y \in \mathbb{E}} F(y)$, which exists by our assumption. Denote by D the maximal radius of the initial level set of the objective, which we assume to be finite:

$$D \stackrel{\text{def}}{=} \sup_{x \in \text{dom } h} \left\{ \|x - x^*\| : F(x) \leq F(x_0) \right\} < +\infty.$$

Then, by monotonicity of the method (3.1) and by convexity we conclude

$$\frac{1}{D} (F(x_{k+1}) - F^*) \leq \frac{1}{D} \langle F'(x_{k+1}), x_{k+1} - x^* \rangle \leq \|F'(x_{k+1})\|_*. \quad (4.2)$$

In the general convex case, we can prove the global sublinear rate of convergence of the Tensor Method of the order $O(1/k^p)$ [13]. For completeness of presentation, let us prove an extension of this result onto the composite case.

Theorem 3 *For the method (3.1) with $H = pL_p$ we have*

$$F(x_k) - F^* \leq \frac{(p+1)(2p)^p}{p!} \cdot \frac{L_p D^{p+1}}{(k-1)^p}, \quad k \geq 2. \quad (4.3)$$

Proof:

Indeed, in view of (2.15) and (4.2), we have for every $k \geq 0$

$$\begin{aligned} F(x_k) - F(x_{k+1}) &\geq \langle F'(x_{k+1}), x_k - x_{k+1} \rangle \\ &\stackrel{(2.15)}{\geq} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}} \cdot \|F'(x_{k+1})\|_*^{\frac{p+1}{p}} \\ &\stackrel{(4.2)}{\geq} \left(\frac{p!}{(p+1)L_p D^{p+1}} \right)^{\frac{1}{p}} \cdot (F(x_{k+1}) - F^*)^{\frac{p+1}{p}}. \end{aligned}$$

Denoting $\delta_k = F(x_k) - F^*$ and $C = \left(\frac{p!}{(p+1)L_p D^{p+1}} \right)^{\frac{1}{p}}$, we obtain the following recurrence:

$$\delta_k - \delta_{k+1} \geq C \delta_{k+1}^{\frac{p+1}{p}}, \quad k \geq 0, \quad (4.4)$$

or for $\mu_k = C^p \delta_k \stackrel{(4.1)}{\leq} 1$, as follows:

$$\mu_k - \mu_{k+1} \geq \mu_{k+1}^{\frac{p+1}{p}}, \quad k \geq 0.$$

Then, Lemma 1.1 from [4] provides us with the following guarantee:

$$\mu_k \leq \left(\frac{p(1+\mu_1^{1/p})}{k-1} \right)^p \leq \left(\frac{2p}{k-1} \right)^p, \quad k \geq 2.$$

Therefore,

$$\delta_k = \frac{\mu_k}{C^p} \leq \left(\frac{2p}{C(k-1)} \right)^p = \frac{(p+1)(2p)^p}{p!} \cdot \frac{L_p D^{p+1}}{(k-1)^p}, \quad k \geq 2.$$

□

For a given degree $q \geq 2$ of uniform convexity with $\sigma_q > 0$, and for RCTM of order $p \geq q - 1$, let us denote by $\omega_{p,q}$ the following *condition number*:

$$\omega_{p,q} \stackrel{\text{def}}{=} \frac{p+1}{p!} \cdot \left(\frac{q-1}{q} \right)^{q-1} \cdot \frac{L_p D^{p-q+1}}{\sigma_q}.$$

Corollary 2 *In order to achieve the region $\mathcal{Q}$ it is enough to perform*

$$\left\lceil 2p \cdot \left(\frac{q^q}{(q-1)^{q-1}} \cdot \omega_{p,q}^{\frac{p+1}{p}} \right)^{\frac{1}{p-q+1}} \right\rceil + 2 \quad (4.5)$$

iterations of the method.

Proof:

Plugging (3.8) into (4.3). □

We can improve this estimate, knowing that the objective is globally uniformly convex (3.2). Then the linear rate of convergence arises at the first state, till the entering in the region $\mathcal{Q}$.

Theorem 4 *Let $\sigma_q > 0$ with $q \leq p + 1$. Then for the method (3.1) with $H = pL_p$, we have*

$$F(x_k) - F^* \leq \exp \left(-\frac{k}{1+\omega_{p,q}^{1/p}} \right) \cdot (F(x_0) - F^*), \quad k \geq 1. \quad (4.6)$$

Therefore, for a given $\varepsilon > 0$ to achieve $F(x_K) - F^ \leq \varepsilon$, it is enough to set*

$$K = \left\lceil (1 + \omega_{p,q}^{1/p}) \cdot \log \frac{F(x_0) - F^*}{\varepsilon} \right\rceil + 1. \quad (4.7)$$

Proof:

Indeed, for every $k \geq 0$

$$\begin{aligned}
F(x_k) - F(x_{k+1}) &\geq \langle F'(x_{k+1}), x_k - x_{k+1} \rangle \\
&\stackrel{(2.15)}{\geq} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}} \cdot \|F'(x_{k+1})\|_*^{\frac{p+1}{p}} \\
&= \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}} \cdot \|F'(x_{k+1})\|_*^{\frac{p-q+1}{p}} \cdot \|F'(x_{k+1})\|_*^{\frac{q}{p}} \\
&\stackrel{(4.2), (3.4)}{\geq} \left(\frac{p!}{p+1} \cdot \frac{\sigma_q}{L_p D^{p-q+1}} \right)^{\frac{1}{p}} \cdot \left(\frac{q}{q-1} \right)^{\frac{q-1}{p}} \cdot (F(x_{k+1}) - F^*) \\
&= \left(\frac{1}{\omega_{p,q}} \right)^{\frac{1}{p}} \cdot (F(x_{k+1}) - F^*).
\end{aligned}$$

Denoting $\delta_k = F(x_k) - F^*$, we obtain

$$\delta_{k+1} \leq \frac{\omega_{p,q}^{1/p}}{1 + \omega_{p,q}^{1/p}} \cdot \delta_k \leq \exp \left(-\frac{1}{1 + \omega_{p,q}^{1/p}} \right) \cdot \delta_k, \quad k \geq 1.$$

□

We see that, for RCTM with $p \geq 2$ minimizing the uniformly convex objective of degree $q \leq p+1$, the condition number $\omega_{p,q}^{1/p}$ is the main factor in the global complexity estimates (4.5) and (4.7). Since in general this number may be arbitrarily big, complexity estimate $\tilde{O}(\omega_{p,q}^{1/p})$ in (4.7) is much better than the estimate $O(\omega_{p,q}^{(p+1)/(p(p-q+1))})$ in (4.5) because of relation $\frac{p+1}{p-q+1} \geq 1$.

These global bounds can be improved, by using the *universal* [1, 6] and the *accelerated* [12, 5, 6, 3, 21] high-order schemes.

5 Application to proximal methods

Let us discuss now a general approach, which uses the local convergence of the methods for justifying the global performance of proximal iterations.

The proximal method [17] is one of the classical methods in theoretical optimization. Every step of the method for solving problem (2.1) is a minimization of the regularized objective:

$$x_{k+1} = \arg \min_{x \in \mathbb{E}} \left\{ a_{k+1} F(x) + \frac{1}{2} \|x - x_k\|^2 \right\}, \quad k \geq 0, \quad (5.1)$$

where $\{a_k\}_{k \geq 1}$ is a sequence of positive coefficients, related to the iteration counter.

Of course, in general, we can hope only to solve subproblem (5.1) inexactly. The questions of practical implementations and possible generalizations of the proximal method, are still in the area of intensive research (see, for example [7, 20, 19, 18]).

One simple observation on the subproblem (5.1) is that it is 1-strongly convex. Therefore, if we would be able to pick an initial point from the region of quadratic convergence (3.8) or (3.9), we could minimize it very quickly by RCTM of degree $p \geq 2$ up to

arbitrary accuracy. In this section, we are going to investigate this approach. For the resulting scheme, we will prove the global rate of convergence of the order $\tilde{O}(1/k^{\frac{p+1}{2}})$.

Denote by Φ_{k+1} the regularized objective from (5.1):

$$\Phi_{k+1}(x) \stackrel{\text{def}}{=} a_{k+1}F(x) + \frac{1}{2}\|x - x_k\|^2 = a_{k+1}f(x) + \frac{1}{2}\|x - x_k\|^2 + a_{k+1}h(x).$$

We fix a sequences of accuracies $\{\delta_k\}_{k \geq 1}$ and relax the assumption on exact minimization in (5.1). Now, at every step we need to find a point x_{k+1} and corresponding subgradient vector $g_{k+1} \in \partial\Phi_{k+1}(x_{k+1})$ with bounded norm:

$$\|g_{k+1}\|_* \leq \delta_{k+1}. \quad (5.2)$$

Denote

$$F'(x_{k+1}) \stackrel{\text{def}}{=} \frac{1}{a_{k+1}}(g_{k+1} - B(x_{k+1} - x_k)) \in \partial F(x_{k+1}).$$

The following global convergence result holds for the general proximal method with inexact minimization criterion (5.2).

Theorem 5 *Assume that there exist a minimum $x^* \in \text{dom } h$ of the problem (2.1). Then, for any $k \geq 1$, we have*

$$\sum_{i=1}^k a_i(F(x_i) - F^*) + \frac{1}{2} \sum_{i=1}^k a_i^2 \|F'(x_i)\|_*^2 + \frac{1}{2} \|x_k - x^*\|^2 \leq R_k(\delta), \quad (5.3)$$

where

$$R_k(\delta) \stackrel{\text{def}}{=} \frac{1}{2} \left(\|x_0 - x^*\| + \sum_{i=1}^k \delta_i \right)^2.$$

Proof:

First, let us prove that for all $k \geq 0$ and for every $x \in \text{dom } h$, we have

$$\begin{aligned} & \frac{1}{2} \|x_0 - x\|^2 + \sum_{i=1}^k a_i F(x) \\ & \geq \frac{1}{2} \|x_k - x\|^2 + \underbrace{\sum_{i=1}^k \left(a_i F(x_i) + \frac{a_i^2}{2} \|F'(x_i)\|_*^2 + \langle g_i, x - x_{i-1} \rangle - \frac{\delta_i^2}{2} \right)}_{\stackrel{\text{def}}{=} C_k(x)}. \end{aligned} \quad (5.4)$$

This is obviously true for $k = 0$. Let it hold for some $k \geq 0$. Consider the step number $k + 1$ of the inexact proximal method.

By condition (5.2), we have

$$\|a_{k+1}F'(x_{k+1}) + B(x_{k+1} - x_k)\|_*^2 \leq \delta_{k+1}^2.$$

Equivalently,

$$\langle a_{k+1}F'(x_{k+1}), x_k - x_{k+1} \rangle \geq \frac{a_{k+1}^2}{2} \|F'(x_{k+1})\|_*^2 + \frac{1}{2} \|x_{k+1} - x_k\|^2 - \frac{\delta_{k+1}^2}{2}. \quad (5.5)$$

Therefore, using the inductive assumption and strong convexity of $\Phi_{k+1}(\cdot)$, we conclude

$$\begin{aligned}
& \frac{1}{2}\|x_0 - x\|^2 + \sum_{i=1}^{k+1} a_i F(x) = \frac{1}{2}\|x_0 - x\|^2 + \sum_{i=1}^k a_i F(x) + a_{k+1} F(x) \\
& \stackrel{(5.4)}{\geq} \Phi_{k+1}(x) + C_k(x) \\
& \geq \Phi_{k+1}(x_{k+1}) + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{1}{2}\|x_{k+1} - x\|^2 + C_k(x) \\
& = a_{k+1} F(x_{k+1}) + \frac{1}{2}\|x_{k+1} - x_k\|^2 + \langle g_{k+1}, x_k - x_{k+1} \rangle \\
& \quad + \langle g_{k+1}, x - x_k \rangle + \frac{1}{2}\|x_{k+1} - x\|^2 + C_k(x) \\
& = a_{k+1} F(x_{k+1}) + \langle a_{k+1} F'(x_{k+1}), x_k - x_{k+1} \rangle - \frac{1}{2}\|x_{k+1} - x_k\|^2 \\
& \quad + \langle g_{k+1}, x - x_k \rangle + \frac{1}{2}\|x_{k+1} - x\|^2 + C_k(x) \\
& \stackrel{(5.5)}{\geq} a_{k+1} F(x_{k+1}) + \frac{a_{k+1}^2}{2} \|F'(x_{k+1})\|_*^2 - \frac{\delta_{k+1}^2}{2} \\
& \quad + \langle g_{k+1}, x - x_k \rangle + \frac{1}{2}\|x_{k+1} - x\|^2 + C_k(x) \\
& = \frac{1}{2}\|x_{k+1} - x\|^2 + C_{k+1}(x).
\end{aligned}$$

Thus, inequality (5.4) is valid for all $k \geq 0$.

Now, plugging $x \equiv x^*$ into (5.4), we have

$$\begin{aligned}
& \sum_{i=1}^k a_i (F(x_i) - F^*) + \frac{1}{2} \sum_{i=1}^k a_i^2 \|F'(x_i)\|_*^2 + \frac{1}{2} \|x_k - x^*\|^2 \\
& \leq \frac{1}{2} \|x_0 - x^*\|^2 + \frac{1}{2} \sum_{i=1}^k \delta_i^2 + \sum_{i=1}^k \langle g_i, x_{i-1} - x^* \rangle \tag{5.6} \\
& \stackrel{(5.2)}{\leq} \frac{1}{2} \|x_0 - x^*\|^2 + \frac{1}{2} \sum_{i=1}^k \delta_i^2 + \sum_{i=1}^k \delta_i \|x_{i-1} - x^*\| \stackrel{\text{def}}{=} \alpha_k.
\end{aligned}$$

In order to finish the proof, it is enough to show that $\alpha_k \leq R_k(\delta)$.

Indeed,

$$\begin{aligned}
\alpha_{k+1} &= \alpha_k + \frac{1}{2} \delta_{k+1}^2 + \delta_{k+1} \|x_k - x^*\| \\
&\stackrel{(5.6)}{\leq} \alpha_k + \frac{1}{2} \delta_{k+1}^2 + \delta_{k+1} \sqrt{2\alpha_k} \\
&= \left(\sqrt{\alpha_k} + \frac{1}{\sqrt{2}} \delta_{k+1} \right)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}\sqrt{\alpha_k} &\leq \sqrt{\alpha_{k-1}} + \frac{1}{\sqrt{2}}\delta_k \leq \dots \leq \sqrt{\alpha_0} + \frac{1}{\sqrt{2}} \sum_{i=1}^k \delta_i \\ &= \frac{1}{\sqrt{2}} \left(\|x_0 - x^*\| + \sum_{i=1}^k \delta_i \right) = \sqrt{R_k(\delta)}.\end{aligned}$$

□

Now, we are ready to use the result on the local superlinear convergence of RCTM in the norm of subgradient (Theorem 2), in order to minimize $\Phi_{k+1}(\cdot)$ at every step of inexact proximal method.

Note that

$$\partial\Phi_{k+1}(x) = a_{k+1}\partial F(x) + B(x - x_k),$$

and it is natural to start minimization process from the previous point x_k , for which $\partial\Phi_{k+1}(x_k) = a_{k+1}\partial F(x_k)$. Let us also notice, that the Lipschitz constant of the p th derivative ($p \geq 2$) of the smooth part of Φ_{k+1} is $a_{k+1}L_p$.

Using our previous notation, one step of RCTM can be written as follows:

$$T_H(\Phi_{k+1}, z) \stackrel{\text{def}}{=} \arg \min_{y \in \mathbb{E}} \left\{ a_{k+1}\psi_{z,p}(y) + \frac{H}{(p+1)!} \|y - z\|^{p+1} + a_{k+1}h(y) + \frac{1}{2} \|y - x_k\|^2 \right\},$$

where $H = a_{k+1}pL_p$. Then, a sufficient condition for $z = x_k$ to be in the region of superlinear convergence (3.9) is

$$a_{k+1} \|F'(x_k)\|_* \leq \left(\frac{p!}{a_{k+1}(p+1)L_p} \right)^{\frac{1}{p-1}},$$

or, equivalently

$$a_{k+1} \leq \left(\frac{1}{\|F'(x_k)\|_*} \right)^{\frac{p-1}{p}} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}}.$$

To be sure that x_k is strictly inside the region, we can pick:

$$\boxed{a_{k+1} = \left(\frac{1}{2\|F'(x_k)\|_*} \right)^{\frac{p-1}{p}} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{1}{p}}} \quad (5.7)$$

Note, that this rule requires fixing an initial subgradient $F'(x_0) \in \partial F(x_0)$, in order to choose a_1 .

Finally, we apply the following steps:

$$z_0 = x_k, \quad z_{t+1} = T_H(\Phi_{k+1}, z_t), \quad t \geq 0. \quad (5.8)$$

We can estimate the required number of these iterations as follows.

Lemma 3 *At every iteration $k \geq 0$ of the inexact proximal method, in order to achieve $\|\Phi'_{k+1}(z_t)\|_* \leq \delta_{k+1}$, it is enough to perform*

$$t_k = \left\lceil \frac{1}{\log_2 p} \cdot \log_2 \log_2 \left(\frac{2}{\delta_{k+1}} \max \left\{ \|x_0 - x^*\| + \sum_{i=1}^k \delta_i, \left(\frac{p! \|F'(x_0)\|_*}{(p+1)L_p 2^{p-1}} \right)^{\frac{1}{p}} \right\} \right) \right\rceil \quad (5.9)$$

steps of RCTM (5.8).

Proof:

Denote $\beta \equiv \left(\frac{a_{k+1}(p+1)L_p}{p!} \right)^{\frac{1}{p-1}} \stackrel{(5.7)}{=} \left(\frac{(p+1)L_p}{2 \cdot p! \cdot \|F'(x_k)\|_*} \right)^{\frac{1}{p}}$. Then, according to (3.7) we have

$$\begin{aligned} \beta \|\Phi'_{k+1}(z_t)\|_* &\leq (\beta \|\Phi'_{k+1}(z_t - 1)\|_*)^p \leq \dots \leq (\beta \|\Phi'_{k+1}(z_0)\|_*)^{p^t} \\ &= (\beta a_{k+1} \|F'(x_k)\|_*)^{p^t} = \left(a_{k+1}^{\frac{p}{p-1}} \left(\frac{(p+1)L_p}{p!} \right)^{\frac{1}{p-1}} \|F'(x_k)\|_* \right)^{p^t} \\ &\stackrel{(5.7)}{=} \left(\frac{1}{2} \right)^{p^t}. \end{aligned} \quad (5.10)$$

Therefore, for

$$t \geq \log_p \log_2 \left(\frac{1}{\beta \delta_{k+1}} \right) = \frac{1}{\log_2 p} \cdot \log_2 \log_2 \left(\frac{1}{\delta_{k+1}} \left(\frac{2 \cdot p! \cdot \|F'(x_k)\|_*}{(p+1)L_p} \right)^{\frac{1}{p}} \right), \quad (5.11)$$

it holds $\|\Phi'_{k+1}(z_t)\|_* \leq \delta_{k+1}$. To finish the proof, let us estimate $\|F'(x_k)\|_*$ from above. We have

$$\begin{aligned} 2^{\frac{3p-2}{p}} \left(\frac{(p+1)L_p}{p!} \right)^{\frac{2}{p}} R_k(\delta) &\stackrel{(5.3)}{\geq} 2^{\frac{2(p-1)}{p}} \left(\frac{(p+1)L_p}{p!} \right)^{\frac{2}{p}} \sum_{i=1}^k a_i^2 \|F'(x_i)\|_*^2 \\ &\stackrel{(5.7)}{=} \sum_{i=1}^k \|F'(x_{i-1})\|_*^{\frac{2(1-p)}{p}} \|F'(x_i)\|_*^2. \end{aligned} \quad (5.12)$$

Thus, for every $1 \leq i \leq k$ it holds

$$\|F'(x_i)\|_* \stackrel{(5.12)}{\leq} \|F'(x_{i-1})\|_*^q \cdot \mathcal{D}, \quad (5.13)$$

with $\mathcal{D} \equiv R_k^{1/2}(\delta) \left(\frac{(p+1)L_p}{p!} \right)^{\frac{1}{p}} 2^{\frac{3p-2}{2p}}$, and $q \equiv \frac{p-1}{p}$. Therefore,

$$\begin{aligned} \|F'(x_k)\|_* &\stackrel{(5.13)}{\leq} \|F'(x_0)\|_*^{q^k} \cdot \mathcal{D}^{1+q+q^2+\dots+q^{k-1}} = \|F'(x_0)\|_* \cdot \left(\|F'(x_0)\|_*^{q^{k-1}} \cdot \mathcal{D}^{\frac{1-q^k}{1-q}} \right) \\ &= \|F'(x_0)\|_* \cdot \left(\frac{\mathcal{D}^p}{\|F'(x_0)\|_*} \right)^{1-q^k} \leq \|F'(x_0)\|_* \cdot \max \left\{ \frac{\mathcal{D}^p}{\|F'(x_0)\|_*}, 1 \right\} \\ &= \max \left\{ \frac{(p+1)L_p 2^{p-1}}{p!} \left(\|x_0 - x^*\| + \sum_{i=1}^k \delta_i \right)^p, \|F'(x_0)\|_* \right\}. \end{aligned}$$

Substitution of this bound into (5.11) gives (5.9). $\square$

Let us prove now the rate of convergence for the outer iterations. This is a direct consequence of Theorem 5 and the choice (5.7) of the coefficients $\{a_k\}_{k \geq 1}$.

Lemma 4 *Let for a given $\varepsilon > 0$,*

$$F(x_k) - F^* \geq \varepsilon, \quad 1 \leq k \leq K. \quad (5.14)$$

Then for every $1 \leq k \leq K$, we have

$$F(\bar{x}_k) - F^* \leq \frac{L_p(\|x_0 - x^*\| + \sum_{i=1}^k \delta_i)^{p+1}}{k^{\frac{p+1}{2}}} \frac{(p+1)2^{p-2}V_k(\varepsilon)}{p!}, \quad (5.15)$$

where $\bar{x}_k \stackrel{\text{def}}{=} \frac{\sum_{i=1}^k a_i x_i}{\sum_{i=1}^k a_i}$, and $V_k(\varepsilon) \stackrel{\text{def}}{=} \left(\frac{\|F'(x_0)\|_* (\|x_0 - x^*\| + \sum_{i=1}^k \delta_i)}{\varepsilon} \right)^{\frac{p-1}{k}}$.

Proof:

Using the inequality between the arithmetic and geometric means, we obtain

$$\begin{aligned} R_k(\delta) &\stackrel{(5.3)}{\geq} \frac{1}{2} \sum_{i=1}^k a_i^2 \|F'(x_i)\|_*^2 \stackrel{(5.7)}{=} \frac{1}{8} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{2}{p-1}} \sum_{i=1}^k \frac{a_i^2}{a_{i+1}^{\frac{2p}{p-1}}} \\ &\geq \frac{k}{8} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{2}{p-1}} \left(\prod_{i=1}^k \frac{a_i^2}{a_{i+1}^{\frac{2p}{p-1}}} \right)^{\frac{1}{k}} \\ &= \frac{k}{8} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{2}{p-1}} \left(\frac{a_1}{a_{k+1}} \right)^{\frac{2p}{(p-1)k}} \left(\prod_{i=1}^k a_i \right)^{\frac{-2}{(p-1)k}} \\ &\geq \frac{k^{\frac{p+1}{2}}}{8} \left(\frac{p!}{(p+1)L_p} \right)^{\frac{2}{p-1}} \left(\frac{a_1}{a_{k+1}} \right)^{\frac{2p}{(p-1)k}} \left(\sum_{i=1}^k a_i \right)^{\frac{-2}{p-1}}. \end{aligned} \quad (5.16)$$

Therefore,

$$\begin{aligned} F(\bar{x}_k) - F^* &\stackrel{(5.3)}{\leq} \frac{R_k(\delta)}{\sum_{i=1}^k a_i} \stackrel{(5.16)}{\leq} \frac{R_k(\delta)^{\frac{p+1}{2}}}{k^{\frac{p+1}{2}}} \frac{(p+1)L_p}{p!} \left(\frac{a_{k+1}}{a_1} \right)^{\frac{p}{2}} 8^{\frac{p-1}{2}} \\ &= \frac{L_p(\|x_0 - x^*\| + \sum_{i=1}^k \delta_i)^{p+1}}{k^{\frac{p+1}{2}}} \frac{(p+1)2^{p-2}}{p!} \left(\frac{\|F'(x_0)\|_*}{\|F'(x_k)\|_*} \right)^{\frac{p-1}{k}}. \end{aligned}$$

By convexity, we have

$$\begin{aligned} \|F'(x_k)\|_* &\geq \frac{\langle F'(x_k), x_k - x^* \rangle}{\|x_k - x^*\|} \geq \frac{F(x_k) - F^*}{\|x_k - x^*\|} \\ &\stackrel{(5.14)}{\geq} \frac{\varepsilon}{\|x_k - x^*\|} \stackrel{(5.3)}{\geq} \frac{\varepsilon}{\|x_0 - x^*\| + \sum_{i=1}^k \delta_i}. \end{aligned}$$

Thus, $\left(\frac{\|F'(x_0)\|_*}{\|F'(x_k)\|_*} \right)^{\frac{p-1}{k}} \leq V_k(\varepsilon)$ and we obtain (5.15). $\square$

Remark 1 Note that $\left(\frac{1}{\varepsilon} \right)^{\frac{p-1}{k}} = \exp\left(\frac{p-1}{k} \ln \frac{1}{\varepsilon}\right)$. Therefore after $k = O\left(\ln \frac{1}{\varepsilon}\right)$ iterations, the factor $V_k(\varepsilon)$ is bounded by an absolute constant.

Since the local convergence of RCTM is very fast (5.9), we can choose the inner accuracies $\{\delta_i\}_{i \geq 1}$ small enough, to have the right hand side of (5.15) being of the order $\tilde{O}(1/k^{\frac{p+1}{2}})$. Let us present a precise statement.

Theorem 6 Let $\delta_k \equiv \frac{c}{k^s}$ for fixed absolute constants $c > 0$ and $s > 1$. Let for a given $\varepsilon > 0$, we have

$$F(x_k) - F^* \geq \varepsilon, \quad 1 \leq k \leq K.$$

Then, for every k such that $\ln \frac{\|F'(x_0)\|_* R}{\varepsilon} \leq k \leq K$, we get

$$F(\bar{x}_k) - F^* \leq \frac{L_p R^{p+1}}{k^{\frac{p+1}{2}}} \frac{(p+1)2^{p-2} \exp(p-1)}{p!}, \quad (5.17)$$

where

$$R \stackrel{\text{def}}{=} \|x_0 - x^*\| + \frac{cs}{s-1}.$$

The total number of oracle calls N_k during the first k iterations is bounded as follows:

$$N_k \leq k \cdot \left(1 + \frac{1}{\log_2 p} \log_2 \log_2 \frac{2Dk^s}{c}\right),$$

where

$$D \stackrel{\text{def}}{=} \max \left\{ R, \left(\frac{p! \|F'(x_0)\|_*}{(p+1)L_p 2^{p-1}} \right)^{\frac{1}{p}} \right\}.$$

Proof:

Indeed,

$$\begin{aligned} \sum_{i=1}^k \delta_i &= c \left(1 + \sum_{i=2}^k \frac{1}{i^s}\right) \leq c \left(1 + \int_1^k \frac{dx}{x^s}\right) = c \left(1 - \frac{1}{s-1} \int_1^k dx^{-(s-1)}\right) \\ &= c \left(1 - \frac{k^{-(s-1)}}{s-1} + \frac{1}{s-1}\right) \leq \frac{cs}{s-1}. \end{aligned}$$

Thus, we obtain (5.17) directly from the bound (5.15), and by the fact that

$$\begin{aligned} V_k(\varepsilon) &\equiv \left(\frac{\|F'(x_0)\|_* R}{\varepsilon} \right)^{\frac{p-1}{k}} = \exp \left(\frac{p-1}{k} \log \frac{\|F'(x_0)\|_* R}{\varepsilon} \right) \\ &\leq \exp(p-1), \end{aligned}$$

when $k \geq \ln \frac{\|F'(x_0)\|_* R}{\varepsilon}$.

Finally,

$$\begin{aligned} N_k &\stackrel{(5.9)}{\leq} \sum_{i=1}^k \left\lceil \frac{1}{\log_2 p} \log_2 \log_2 \frac{2D}{\delta_i} \right\rceil \leq k + \frac{1}{\log_2 p} \sum_{i=1}^k \log_2 \log_2 \frac{2Di^s}{c} \\ &\leq k + \frac{1}{\log_2 p} \sum_{i=1}^k \log_2 \log_2 \frac{2Dk^s}{c} = k \cdot \left(1 + \frac{1}{\log_2 p} \log_2 \log_2 \frac{2Dk^s}{c}\right). \end{aligned}$$

□

Note that we were able to justify the global performance of the scheme, using only the local convergence results for the inner method. It is interesting to compare our approach with the recent results on the path-following second-order methods [2].

We can drop the logarithmic components in the complexity bounds by using the *hybrid proximal methods* (see [11] and [10]), where at each iteration only one step of RCTM is performed. The resulting rate of convergence there is $O(1/k^{\frac{p+1}{2}})$, without any extra logarithmic factors. However, this rate is worse than the rate $O(1/k^p)$ provided by the Theorem 3 for the primal iterations of RCTM (3.1).

References

- [1] Nikita Doikov and Yurii Nesterov. Minimizing uniformly convex functions by cubic regularization of Newton method. *arXiv preprint arXiv:1905.02671*, 2019.
- [2] Pavel Dvurechensky and Yurii Nesterov. Global performance guarantees of second-order methods for unconstrained convex minimization. Technical report, CORE Discussion Paper, 2018.
- [3] 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 *Conference on Learning Theory*, pages 1374–1391, 2019.
- [4] Geovani N Grapiglia and Yurii Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIAM Journal on Optimization*, 27(1):478–506, 2017.
- [5] Geovani N Grapiglia and Yurii Nesterov. Accelerated regularized Newton methods for minimizing composite convex functions. *SIAM Journal on Optimization*, 29(1):77–99, 2019.
- [6] Geovani N Grapiglia and Yurii Nesterov. Tensor methods for minimizing functions with Hölder continuous higher-order derivatives. *arXiv preprint arXiv:1904.12559*, 2019.
- [7] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.
- [8] Leonid V Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.
- [9] Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.
- [10] M Marques Alves, Renato DC Monteiro, and Benar F Svaiter. Iteration-complexity of a Rockafellar’s proximal method of multipliers for convex programming based on second-order approximations. *Optimization*, pages 1–30, 2019.
- [11] Renato DC Monteiro and Benar F Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.
- [12] Yurii Nesterov. Accelerating the cubic regularization of Newton’s method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.
- [13] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Universite catholique de Louvain, Center for Operations Research and Econometrics (CORE)*, 2018.
- [14] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.
- [15] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

- [16] Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton’s method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.
- [17] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.
- [18] Saverio Salzo and Silvia Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex analysis*, 19(4):1167–1192, 2012.
- [19] Mark Schmidt, Nicolas L Roux, and Francis R Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.
- [20] Mikhail V Solodov and Benar F Svaiter. A unified framework for some inexact proximal point algorithms. *Numerical functional analysis and optimization*, 22(7-8):1013–1035, 2001.
- [21] Chaobing Song and Yi Ma. Towards unified acceleration of high-order algorithms under Hölder continuity and uniform convexity. *arXiv preprint arXiv:1906.00582*, 2019.