

# Near-optimal method for highly smooth convex optimization

Sébastien Bubeck  
Microsoft Research

Qijia Jiang  
Stanford University

Yin Tat Lee \*  
University of Washington  
& Microsoft Research

Yuanzhi Li  
Stanford University

Aaron Sidford †  
Stanford University

June 25, 2019

## Abstract

We propose a near-optimal method for highly smooth convex optimization. More precisely, in the oracle model where one obtains the  $p^{\text{th}}$  order Taylor expansion of a function at the query point, we propose a method with rate of convergence  $\tilde{O}(1/k^{\frac{3p+1}{2}})$  after  $k$  queries to the oracle for any convex function whose  $p^{\text{th}}$  order derivative is Lipschitz.

## 1 Introduction

In this paper we generalize the important phenomenon of *acceleration* in smooth convex optimization [7, 6, 8] to higher orders of smoothness. We consider a  $p^{\text{th}}$ -order Taylor expansion oracle, that is given a query point  $x \in \mathbb{R}^d$  it returns a  $p^{\text{th}}$  order Taylor expansion of the objective function  $f$  at the point  $x$ :

$$f_p(y, x) = f(x) + \sum_{i=1}^p \frac{1}{i!} \nabla^i f(x) [y - x]^i.$$

We propose a new optimization method based on such oracle, see Algorithm 1, which we term *accelerated Taylor descent (ATD)*. We prove that it attains a nearly optimal rate of convergence under higher order smoothness (the matching lower bounds were recently proven in [1, 2]), namely after  $\tilde{O}(k)$  calls to the oracle it achieves error  $O(1/k^{\frac{3p+1}{2}})$ . This improves upon the  $O(1/k^{p+1})$  derived in [10] (both rates match for  $p = 1$ , i.e., the classical acceleration setting), and it matches the rate given in [5] for  $p = 2$ .

**Theorem 1.1.** *Let  $f$  denote a convex function whose  $p^{\text{th}}$  derivative is  $L_p$ -Lipschitz and let  $x^*$  denote a minimizer of  $f$ . Then ATD satisfies, with  $c_p = 2^{p-1}(p+1)^{\frac{3p+1}{2}}/(p-1)!$ ,*

$$f(y_k) - f(x^*) \leq \frac{c_p \cdot L_p \cdot \|x^*\|^{p+1}}{k^{\frac{3p+1}{2}}}. \quad (1)$$

Furthermore each iteration of ATD can be implemented in  $\tilde{O}(1)$  calls to a  $p^{\text{th}}$ -order Taylor expansion oracle. More precisely, given a precision  $\varepsilon > 0$ , at each iteration  $k$ , using at most

$$30p \log_2 p + \log_2 \left\lceil \frac{L_p \|x^*\|^{p+1}}{\varepsilon} \right\rceil$$

\*Research was supported in part by NSF Awards CCF-1740551, CCF-1749609, and DMS-1839116.

†Research was supported in part by NSF CAREER Award CCF-1844855.

calls to the  $p^{\text{th}}$ -order Taylor expansion oracle we find either a point  $y$  such that  $f(y) - f(x^*) \leq \varepsilon$ , or we find  $y_k$ .

Our method is largely inspired by [5], which focuses on  $p = 2$ , and we recall their framework in Section 2. We then specialize this framework to higher order smoothness in Section 3, where we derive and analyze ATD. A subtle point of ATD is that an iteration requires more than one call to the oracle due to the “line-search” [line 4, Algorithm 1]. We prove that  $\tilde{O}(1)$  calls suffice to implement an iteration in Section 4.

We note that the independent work [3], currently only available in Russian, derive a similar result to (1). From our understanding of their work it seems however that they do not work out the precise complexity of the binary search step (second part of the statement in Theorem 1.1, see also Section 4). Finally we note that yet another independent work [4] was posted on the arxiv a couple of days prior to us, with a similar result to Theorem 1.1. Interestingly it seems that their argument to control the complexity of the binary search is different (at least on the surface) from ours.

---

**Algorithm 1** Accelerated Taylor Descent

---

- 1: **Input:** convex function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $\nabla^p f$  is  $L_p$ -Lipschitz.
- 2: Set  $A_0 = 0, x_0 = y_0 = 0$
- 3: **for**  $k = 0$  **to**  $k = K - 1$  **do**
- 4:   Compute a pair  $\lambda_{k+1} > 0$  and  $y_{k+1} \in \mathbb{R}^d$  such that

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

where

$$y_{k+1} = \arg \min_y \left\{ f_p(y; \tilde{x}_k) + \frac{L_p}{p!} \|y - \tilde{x}_k\|^{p+1} \right\},$$

and

$$a_{k+1} = \frac{\lambda_{k+1} + \sqrt{\lambda_{k+1}^2 + 4\lambda_{k+1}A_k}}{2}, \quad A_{k+1} = A_k + a_{k+1}, \quad \text{and} \quad \tilde{x}_k = \frac{A_k}{A_{k+1}}y_k + \frac{a_{k+1}}{A_{k+1}}x_k.$$

- 5:   Update  $x_{k+1} := x_k - a_{k+1}\nabla f(y_{k+1})$
  - 6: **end for**
  - 7: **return**  $y_K$
- 

**Remark 1.2.** The definition of  $a_{k+1}$  was chosen such that  $\lambda_{k+1}A_{k+1} = a_{k+1}^2$ . To see this, note that  $a_{k+1}$  is a solution to  $a_{k+1}^2 - \lambda_{k+1}a_{k+1} - \lambda_{k+1}A_k = 0$ , which is equivalent as  $A_{k+1} = A_k + a_{k+1}$ .

## 2 Monteiro-Svaiter acceleration framework

Recall that Nesterov’s accelerated gradient descent [8, 9] produces a sequence of the form:

$$y_{k+1} = \tilde{x}_k - \lambda_{k+1}\nabla f(\tilde{x}_k), \tag{2}$$

for some step size  $\lambda_{k+1}$  and “momentum” point  $\tilde{x}_k$ . In this section we consider a variant proposed by Monteiro and Svaiter which replaces the gradient step by a form of “implicit gradient step”,

namely:

$$y_{k+1} \simeq \tilde{x}_k - \lambda_{k+1} \nabla f(y_{k+1}).$$

The rest of the section is merely a rewriting of [5], with the objective to motivate and prove the following result:

**Theorem 2.1.** *Let  $(y_k)_{k \geq 1}$  be a sequence of points in  $\mathbb{R}^d$  and  $(\lambda_k)_{k \geq 1}$  a sequence in  $\mathbb{R}_+$ . Define  $(a_k)_{k \geq 1}$  such that  $\lambda_k A_k = a_k^2$  where  $A_k = \sum_{i=1}^k a_i$ . Define also for any  $k \geq 0$ ,  $x_k = -\sum_{i=1}^k a_i \nabla f(y_i)$  (in particular  $x_0 = 0$ ) and  $\tilde{x}_k := \frac{a_{k+1}}{A_{k+1}} x_k + \frac{A_k}{A_{k+1}} y_k$ . Finally assume that*

$$\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1} \nabla f(y_{k+1}))\| \leq \|y_{k+1} - \tilde{x}_k\|. \quad (3)$$

Then one has for any  $x \in \mathbb{R}^d$ ,

$$f(y_k) - f(x) \leq \frac{2\|x\|^2}{\left(\sum_{i=1}^k \sqrt{\lambda_i}\right)^2}. \quad (4)$$

Furthermore if one has the following refined guarantee, for some  $\sigma \in [0, 1]$ ,

$$\|y_{k+1} - (\tilde{x}_k - \lambda_{k+1} \nabla f(y_{k+1}))\| \leq \sigma \cdot \|y_{k+1} - \tilde{x}_k\|, \quad (5)$$

then one also has

$$\sum_{i=1}^k \frac{A_i}{\lambda_i} \|y_i - \tilde{x}_{i-1}\|^2 \leq \frac{\|x^*\|^2}{1 - \sigma^2}. \quad (6)$$

To illustrate the power of Theorem 2.1, observe that for a  $L_1$ -smooth function (first-order smoothness) one has that Nesterov's accelerated gradient descent (2) directly satisfies (3) provided that  $\lambda_{k+1} = \frac{1}{L_1}$  (i.e., the classical step-size for smooth convex optimization). Using (4) this immediately shows that (2) has a rate of convergence of  $O(1/k^2)$

The key to higher-order acceleration will be to show that in fact one can take  $\lambda_k$  to be an increasing function of  $A_k$ , thanks to a careful use of (6). This will be done in Section 3.

We now embark on the road leading to Theorem 2.1.

## 2.1 Estimate sequence style analysis

Similarly to the original construction by Nemirovski [7, 6] (and taking inspiration from the conjugate gradient method) the starting point is to consider a linear combination of past gradients:  $x_k := -\sum_{i=1}^k a_i \nabla f(y_i)$ , where both the coefficients  $(a_i)$  and the query points  $(y_i)$  are yet to be defined. In the spirit of Nesterov's estimate sequence analysis, a key observation for such linear combination of gradients is that it minimizes an approximate lower bound on  $f$ :

**Lemma 2.2.** *Let  $\psi_0(x) = \frac{1}{2}\|x\|^2$  and define by induction  $\psi_k(x) = \psi_{k-1}(x) + a_k f_1(x, y_k)$ . Then  $x_k = -\sum_{i=1}^k a_i \nabla f(y_i)$  is the minimizer of  $\psi_k$ , and  $\psi_k(x) \leq A_k f(x) + \frac{1}{2}\|x\|^2$  where  $A_k = \sum_{i=1}^k a_i$ .*

The next idea is to produce a "control sequence"  $(z_k)_{k \geq 1}$  demonstrating that  $\psi_k$  is not too far below  $A_k f$ , which in turn would directly yield a convergence rate for  $z_k$  of order  $1/A_k$ :

**Lemma 2.3.** *Let  $(z_k)$  be a sequence such that*

$$\psi_k(x_k) - A_k f(z_k) \geq 0. \quad (7)$$

Then one has for any  $x$ ,

$$f(z_k) \leq f(x) + \frac{\|x\|^2}{2A_k}. \quad (8)$$

*Proof.* One has (recall Lemma 2.2):

$$A_k f(z_k) \leq \psi_k(x_k) \leq \psi_k(x) \leq A_k f(x) + \frac{1}{2} \|x\|^2.$$

□

## 2.2 A proof by induction

Our goal is now to come up with sequences  $(a_k, y_k, z_k)$  satisfying (7). The following lemma, resulting from elementary calculations, reveals a simple condition to obtain (7) from an induction argument:

**Lemma 2.4.** *One has for any  $x$ ,*

$$\begin{aligned} & \psi_{k+1}(x) - A_{k+1}f(y_{k+1}) - (\psi_k(x_k) - A_k f(z_k)) \\ & \geq A_{k+1} \nabla f(y_{k+1}) \cdot \left( \frac{a_{k+1}}{A_{k+1}} x + \frac{A_k}{A_{k+1}} z_k - y_{k+1} \right) + \frac{1}{2} \|x - x_k\|^2. \end{aligned}$$

*Proof.* First we note that (the first equality follows from the fact that the Hessian of  $\psi_k$  remains the identity for any  $k$ ):

$$\psi_k(x) = \psi_k(x_k) + \frac{1}{2} \|x - x_k\|^2, \text{ and } \psi_{k+1}(x) = \psi_k(x_k) + \frac{1}{2} \|x - x_k\|^2 + a_{k+1} f_1(x, y_{k+1}),$$

so that

$$\psi_{k+1}(x) - \psi_k(x_k) = a_{k+1} f_1(x, y_{k+1}) + \frac{1}{2} \|x - x_k\|^2. \quad (9)$$

Now we want to make appear the term  $A_{k+1}f(z_{k+1}) - A_k f(z_k)$  as a lower bound on the right hand side of (9) when evaluated at  $x = x_{k+1}$ . Using the inequality  $f_1(z_k, y_{k+1}) \leq f(z_k)$  we have:

$$\begin{aligned} a_{k+1} f_1(x, y_{k+1}) &= A_{k+1} f_1(x, y_{k+1}) - A_k f_1(x, y_{k+1}) \\ &= A_{k+1} f_1(x, y_{k+1}) - A_k \nabla f(y_{k+1}) \cdot (x - z_k) - A_k f_1(z_k, y_{k+1}) \\ &= A_{k+1} f_1 \left( x - \frac{A_k}{A_{k+1}} (x - z_k), y_{k+1} \right) - A_k f_1(z_k, y_{k+1}) \\ &\geq A_{k+1} f(y_{k+1}) - A_k f(z_k) + A_{k+1} \nabla f(y_{k+1}) \cdot \left( \frac{a_{k+1}}{A_{k+1}} x + \frac{A_k}{A_{k+1}} z_k - y_{k+1} \right), \end{aligned}$$

which concludes the proof. □

From Lemma 2.4 we see that it is natural to take for the control sequence  $z_k := y_k$ , so that:

$$\psi_{k+1}(x) - A_{k+1}f(y_{k+1}) - (\psi_k(x_k) - A_k f_k(y_k)) \quad (10)$$

$$\geq A_{k+1} \nabla f(y_{k+1}) \cdot \left( \frac{a_{k+1}}{A_{k+1}} x + \frac{A_k}{A_{k+1}} y_k - y_{k+1} \right) + \frac{1}{2} \|x - x_k\|^2. \quad (11)$$

We would like to pick the query point  $y_{k+1}$  so that (11) is nonnegative when evaluated at  $x = x_{k+1}$  (to satisfy (7)). One difficulty is that  $x_{k+1}$  itself depends on  $y_{k+1}$ , so in fact we will pick  $y_{k+1}$  so that the right side is nonnegative *for all*  $x$ . We write this as follows:

**Lemma 2.5.** Denoting  $\lambda_{k+1} := \frac{a_{k+1}^2}{A_{k+1}}$  and  $\tilde{x}_k := \frac{a_{k+1}}{A_{k+1}}x_k + \frac{A_k}{A_{k+1}}y_k$  one has:

$$\begin{aligned} & \psi_{k+1}(x_{k+1}) - A_{k+1}f(y_{k+1}) - (\psi_k(x_k) - A_kf(y_k)) \\ & \geq \frac{A_{k+1}}{2\lambda_{k+1}} \left( \|y_{k+1} - \tilde{x}_k\|^2 - \|y_{k+1} - (\tilde{x}_k - \lambda_{k+1}\nabla f(y_{k+1}))\|^2 \right). \end{aligned}$$

In particular, we have in light of (5)

$$\psi_k(x_k) - A_kf(y_k) \geq \frac{1 - \sigma^2}{2} \sum_{i=1}^k \frac{A_i}{\lambda_i} \|y_i - \tilde{x}_{i-1}\|^2.$$

*Proof.* We apply Lemma 2.4 with  $z_k = y_k$  and  $x = x_{k+1}$ , and note that (with  $\tilde{x} := \frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}y_k$ ):

$$\begin{aligned} & \nabla f(y_{k+1}) \cdot \left( \frac{a_{k+1}}{A_{k+1}}x + \frac{A_k}{A_{k+1}}y_k - y_{k+1} \right) + \frac{1}{2A_{k+1}} \|x - x_k\|^2 \\ & = \nabla f(y_{k+1}) \cdot (\tilde{x} - y_{k+1}) + \frac{1}{2A_{k+1}} \left\| \frac{A_{k+1}}{a_{k+1}} \left( \tilde{x} - \frac{A_k}{A_{k+1}}y_k \right) - x_k \right\|^2 \\ & = \nabla f(y_{k+1}) \cdot (\tilde{x} - y_{k+1}) + \frac{A_{k+1}}{2a_{k+1}^2} \left\| \tilde{x} - \left( \frac{a_{k+1}}{A_k}x_k + \frac{A_k}{A_{k+1}}y_k \right) \right\|^2. \end{aligned}$$

This yields:

$$\begin{aligned} & \psi_{k+1}(x_{k+1}) - A_{k+1}f(y_{k+1}) - (\psi_k(x_k) - A_kf(y_k)) \\ & \geq A_{k+1} \cdot \min_{x \in \mathbb{R}^d} \left\{ \nabla f(y_{k+1}) \cdot (x - y_{k+1}) + \frac{1}{2\lambda_{k+1}} \|x - \tilde{x}_k\|^2 \right\}. \end{aligned}$$

It only remains to compute the value of this minimum, which is an easy exercise.  $\square$

### 2.3 Proof of Theorem 2.1

For the first conclusion in Theorem 2.1, it suffices to combine Lemma 2.5 with Lemma 2.3, and to use the following observation:

**Lemma 2.6.** Let  $(\lambda_k)$  be a sequence of nonnegative numbers. Define  $(a_k)$  to be another sequence of nonnegative numbers such that  $\lambda_k A_k = a_k^2$ , where  $A_k = \sum_{i=1}^k a_i$ . In other words one has  $a_k = \frac{\lambda_k + \sqrt{\lambda_k^2 + 4\lambda_k A_{k-1}}}{2}$ . Furthermore one also has:

$$\sqrt{A_k} \geq \frac{1}{2} \sum_{i=1}^k \sqrt{\lambda_i}.$$

*Proof.* It suffices to observe that:

$$a_k = \frac{\lambda_k + \sqrt{\lambda_k^2 + 4\lambda_k A_{k-1}}}{2} \geq \frac{\lambda_k}{2} + \sqrt{\lambda_k A_{k-1}} \geq \left( \frac{\sqrt{\lambda_k}}{2} + \sqrt{A_{k-1}} \right)^2 - A_{k-1}.$$

$\square$

The second conclusion in Theorem 2.1 follows from Lemma 2.5 and Lemma 2.2.

### 3 Accelerated Taylor Descent

Nesterov's accelerated gradient descent (2) (with  $\lambda_k = 1/L_1$ ) can be rewritten as:

$$y_{k+1} = \arg \min_{y \in \mathbb{R}^d} f_1(y, \tilde{x}_k) + \frac{L_1}{2} \|y - \tilde{x}_k\|^2.$$

We naturally propose to use the following generalization for higher-order smoothness, which we term *accelerated Taylor descent (ATD)*:

$$y_{k+1} = \arg \min_{y \in \mathbb{R}^d} f_p(y, \tilde{x}_k) + \frac{L_p}{p!} \|y - \tilde{x}_k\|^{p+1}. \quad (12)$$

The term  $\|\cdot\|^{p+1}$  is added to ensure that the function being optimized is strictly convex. In Section 3.1 we first show that ATD satisfies (3) for a special value of  $\lambda_{k+1}$  defined in terms of  $y_{k+1}$ . We point out that there is an intricate issue here, in the sense that  $y_{k+1}$  depends on  $\lambda_{k+1}$  (through the definition of  $\tilde{x}_k$ ), and thus we will have to select the pair  $(y_{k+1}, \lambda_{k+1})$  simultaneously rather than sequentially. This is detailed in Section 3.2. Finally in Section 3.3 we use (6) with the special values of  $(\lambda_i)$  to derive the rate of convergence from Theorem 1.1.

#### 3.1 ATD and implicit gradient descent with large step size

The following lemma shows that minimizing the  $p^{\text{th}}$  order Taylor expansion (12) can be viewed as an implicit gradient step for some “large” step size:

**Lemma 3.1.** *Equation (5) holds true with  $\sigma = 1/2$  for (12), provided that one has:*

$$\frac{1}{2} \leq \lambda_{k+1} \frac{L_p \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}}{(p-1)!} \leq \frac{p}{p+1}. \quad (13)$$

*Proof.* Observe that the optimality condition gives:

$$\nabla_y f_p(y_{k+1}, \tilde{x}_k) + \frac{L_p \cdot (p+1)}{p!} (y_{k+1} - \tilde{x}_k) \|y_{k+1} - \tilde{x}_k\|^{p-1} = 0. \quad (14)$$

In particular we get:

$$y_{k+1} - (\tilde{x}_k - \lambda_{k+1} \nabla f(y_{k+1})) = \lambda_{k+1} \nabla f(y_{k+1}) - \frac{p!}{L_p \cdot (p+1) \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}} \nabla_y f_p(y_{k+1}, \tilde{x}_k).$$

By doing a Taylor expansion of the gradient function one obtains:

$$\|\nabla f(y) - \nabla_y f_p(y, x)\| \leq \frac{L_p}{p!} \|y - x\|^p,$$

so that we find:

$$\begin{aligned} & \|y_{k+1} - (\tilde{x}_k - \lambda_{k+1} \nabla f(y_{k+1}))\| \\ & \leq \lambda_{k+1} \frac{L_p}{p!} \|y_{k+1} - \tilde{x}_k\|^p + \left| \lambda_{k+1} - \frac{p!}{L_p \cdot (p+1) \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}} \right| \cdot \|\nabla_y f_p(y_{k+1}, \tilde{x}_k)\| \\ & \leq \|y_{k+1} - \tilde{x}_k\| \left( \lambda_{k+1} \frac{L_p}{p!} \|y_{k+1} - \tilde{x}_k\|^{p-1} + \left| \lambda_{k+1} \frac{L_p \cdot (p+1) \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}}{p!} - 1 \right| \right) \\ & = \|y_{k+1} - \tilde{x}_k\| \left( \frac{\eta}{p} + \left| \eta \cdot \frac{p+1}{p} - 1 \right| \right) \end{aligned}$$

where we used (14) in the second last equation and we let  $\eta := \lambda_{k+1} \frac{L_p \cdot \|y_{k+1} - \tilde{x}_k\|^{p-1}}{(p-1)!}$  in the last equation. The result follows from the assumption  $1/2 \leq \eta \leq p/(p+1)$  in (13).  $\square$

### 3.2 A continuity argument

We now claim that there exists a pair  $(y_{k+1}, \lambda_{k+1})$  that satisfies simultaneously (12) and (13). This is a direct consequence of the following lemma.

**Lemma 3.2.** *Let  $A \geq 0$ ,  $x, y \in \mathbb{R}^d$  such that  $f(x) \neq f(x^*)$ . Define the following functions:*

$$a(\lambda) = \frac{\lambda + \sqrt{\lambda^2 + 4\lambda A}}{2}, \quad x(\lambda) = \frac{a(\lambda)}{A + a(\lambda)}x + \frac{A}{A + a(\lambda)}y,$$

$$y(z) = \arg \min_{w \in \mathbb{R}^d} \left\{ f_p(w, z) + \frac{L_p}{p!} \|w - z\|^{p+1} \right\}, \quad g(\lambda) = \lambda \|y(x(\lambda)) - x(\lambda)\|^{p-1}.$$

Then we have  $g(\mathbb{R}_+) = \mathbb{R}_+$ .

*Proof.* First we claim that  $g(\lambda)$  is a continuous function of  $\lambda$ . The only non-trivial part of this statement is that  $y(z)$  is a continuous function of  $z$ . The latter statement follows easily from the strict convexity of the function being optimized, see also Section 4 for more details.

Next we claim that  $g(0) = 0$ , and furthermore since  $f(x) \neq f(x^*)$  we also have  $y(x) \neq x$  which in turns gives  $g(+\infty) = +\infty$ . This concludes the proof.  $\square$

### 3.3 Proof of (1) in Theorem 1.1

Recall from Lemma 2.3 that the rate of convergence of ATD is  $\|x^*\|^2/(2A_k)$ . We now finally give an estimate of  $A_k$ :

**Lemma 3.3.** *One has, with  $c_p = 2^{p-1}(p+1)^{\frac{3p+1}{2}}/(p-1)!$ ,*

$$A_k \geq \frac{1}{c_p \cdot L_p \cdot \|x^*\|^{p-1}} k^{\frac{3p+1}{2}}.$$

*Proof.* Using Lemma 3.1 (and in particular (13)) in (6) we obtain, with  $C_p = 8 \cdot \left(\frac{L_p}{(p-1)!}\right)^{\frac{2}{p-1}}$ ,

$$\sum_{i=1}^k \frac{A_i}{\lambda_i^{\frac{p+1}{p-1}}} \leq C_p \|x^*\|^2. \quad (15)$$

Now by reverse Hölder inequality, i.e.  $\|fg\|_1 \geq \|f\|_{\frac{1}{q}} \|g\|_{\frac{-1}{q-1}}$  for  $q \geq 1$ , and invoking this inequality with  $q = 1 + \frac{p-1}{2(p+1)} = \frac{3p+1}{2(p+1)}$  so that  $\frac{-1}{1-q} = -\frac{2(p+1)}{p-1}$ , we have

$$\sum_{j=1}^k \sqrt{\lambda_j} = \sum_{j=1}^k (A_j)^{\frac{p-1}{2(p+1)}} \left( \frac{A_j}{\lambda_j^{\frac{p+1}{p-1}}} \right)^{-\frac{p-1}{2(p+1)}} \geq \left( \sum_{j=1}^k A_j^{\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{2(p+1)}} \left( \sum_{j=1}^k \frac{A_j}{\lambda_j^{\frac{p+1}{p-1}}} \right)^{-\frac{p-1}{2(p+1)}}. \quad (16)$$

Combining (15) and (16) and using by Lemma 2.6 we have for all  $k \geq 1$  that

$$A_k \geq \frac{1}{4} \left( \sum_{j \in [k]} \sqrt{\lambda_j} \right)^2 \geq \frac{1}{4(C_p \|x^*\|^2)^{\frac{p-1}{p+1}}} \left( \sum_{j=1}^k A_j^{\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{p+1}}$$

Next we apply Lemma 3.4 (see below) with  $\alpha = \frac{p+1}{p-1}$ ,  $B_k = A_k^{\frac{p-1}{3p+1}}$  and  $c = \frac{1}{4^{\frac{p+1}{3p+1}} (C_p \|x^*\|^2)^{\frac{p-1}{3p+1}}}$ :

$$B_k \geq \left( \frac{2}{p+1} \cdot c \cdot k \right)^{\frac{p-1}{2}},$$

or in other words,  $A_k \geq \left( \frac{2}{p+1} \cdot c \cdot k \right)^{\frac{3p+1}{2}}$ , which concludes the proof.  $\square$

**Lemma 3.4.** *Given a non-decreasing positive sequence  $B_j$  such that  $B_k^\alpha \geq c \cdot \sum_{j=1}^k B_j$ . Then, we have that*

$$B_k \geq \left( \frac{\alpha-1}{\alpha} c \cdot k \right)^{\frac{1}{\alpha-1}}$$

*Proof.* We extend  $B_t = B_{\lceil t \rceil}$ . Note that

$$B_t^\alpha = B_{\lceil t \rceil}^\alpha \geq c \cdot \sum_{j=1}^{\lceil t \rceil} B_j \geq c \cdot \int_0^t B_s ds.$$

We can upper bound this integral inequality  $B_t \geq U_t$  where  $U_1 = B_1$  and

$$U_t^\alpha = c \cdot \int_0^t U_s ds.$$

Taking derivatives on both sides, we have

$$\alpha U_t^{\alpha-1} \frac{dU_t}{dt} = c \cdot U_t.$$

and hence  $\frac{dU_t^{\alpha-1}}{dt} = \frac{\alpha-1}{\alpha} c$ . Therefore, we have  $B_t \geq U_t = \left( \frac{\alpha-1}{\alpha} c \cdot (t-1) + B_1^{\alpha-1} \right)^{\frac{1}{\alpha-1}}$ . Finally, the result follows from  $B_1^{\alpha-1} \geq c$ .  $\square$

## 4 Complexity of the binary search step

In this section, we show how to find  $\lambda_{k+1}$  satisfying equation (13). For  $k=0$ , it is trivial since  $\tilde{x}_0 = 0$ . From now on, we fix some  $k > 0$ . To simplify the notation, we define  $\tilde{x}_\theta = (1-\theta)x_k + \theta y_k$ ,  $y_\theta = \arg \min_y F(y - \tilde{x}_\theta, \tilde{x}_\theta)$  with

$$F(z, x) = f_p(x+z, x) + \frac{L_p}{p!} \|z\|^{p+1},$$

and  $z_\theta = y_\theta - \tilde{x}_\theta$ . Note that the  $\lambda_{k+1}$  corresponding to  $\theta$  is given by  $\lambda_{k+1} = \frac{(1-\theta)^2}{\theta} A_k$ . Hence, our goal is to find  $\theta$  such that

$$\frac{1}{2} \leq \zeta(\theta) \leq \frac{p}{p+1} \quad \text{with} \quad \zeta(\theta) = \frac{(1-\theta)^2}{\theta} \frac{A_k \cdot L_p}{(p-1)!} \|z_\theta\|^{p-1}.$$

Note that  $\zeta(0) = +\infty$  and  $\zeta(1) = 0$ . Hence, we can use binary search to find  $\theta$  that is close to  $\theta^*$  such that  $\zeta(\theta^*) = \frac{7}{12}$  (or any value in  $(\frac{1}{2}, \frac{p}{p+1})$ ). The main difficulty is to show how close  $\theta$  need to be so that  $\zeta(\theta) \in [\frac{1}{2}, \frac{p}{p+1}]$ , or in other words to control the Lipschitz constant of  $\zeta(\theta)$ .

To bound the Lipschitz constant of  $\zeta(\theta)$ , we need to bound  $\|z_\theta\|$  and  $\|\frac{d}{d\theta} z_\theta\|$ . First, we give an upper bound on  $\|\frac{d}{d\theta} z_\theta\|$ .



**Lemma 4.1.** *We have:*

$$\left\| \frac{d}{d\theta} z_\theta \right\| \leq 5(p+1)^2 \cdot \|x^*\|.$$

*Proof.* To compute the derivative of  $z_\theta$ , we note by optimality condition that

$$\nabla_z F(z_\theta, \tilde{x}_\theta) = 0.$$

Taking derivatives with respect to  $\theta$  on both sides gives

$$\nabla_{zz}^2 F(z_\theta, \tilde{x}_\theta) \cdot \frac{d}{d\theta} z_\theta + \nabla_{zx}^2 F(z_\theta, \tilde{x}_\theta) \cdot \frac{d}{d\theta} \tilde{x}_\theta = 0.$$

Hence, we have

$$\frac{d}{d\theta} z_\theta = - (\nabla_{zz}^2 F(z_\theta, \tilde{x}_\theta))^{-1} \nabla_{zx}^2 F(z_\theta, \tilde{x}_\theta) \cdot (y_k - x_k). \quad (17)$$

To bound  $\frac{d}{d\theta} z_\theta$ , it suffices to compute  $\nabla_{zz}^2 F(z, x)$  and  $\nabla_{zx}^2 F(z, x)$ .

For  $\nabla_{zz}^2 F(z, x)$ , we have

$$\nabla_{zz}^2 F(z, x) = \nabla_{zz}^2 f_p(x+z, x) + \nabla^2 \left[ \frac{L_p}{p!} \|z\|^{p+1} \right].$$

By doing a Taylor expansion of the Hessian function, one obtains:

$$\|\nabla_{zz}^2 f_p(x+z, x) - \nabla^2 f(x+z)\| \leq \frac{L_p}{(p-1)!} \|z\|^{p-1}$$

and hence

$$\nabla_{zz}^2 F(z, x) \succeq \nabla^2 f(x+z) - \frac{L_p}{(p-1)!} \|z\|^{p-1} I + \frac{L_p(p+1)}{p!} \|z\|^{p-1} I \succeq \frac{L_p}{p!} \|z\|^{p-1} I$$

where we used that  $f$  is convex and

$$\nabla^2 [\|z\|^{p+1}] = (p+1) \|z\|^{p-1} \cdot I + (p+1)(p-1) \|z\|^{p-3} \cdot z z^\top. \quad (18)$$

For  $\nabla_{zx}^2 F(z, x)$ , we recall that  $F(z, x) = \sum_{i=0}^p \frac{1}{i!} D^i f(x) [z]^i + \frac{L_p}{p!} \|z\|^{p+1}$ , and hence

$$\begin{aligned} \nabla_{zx}^2 F(z, x) &= \sum_{i=1}^p \frac{1}{(i-1)!} D^{i+1} f(x) [z]^{i-1} \\ &= \nabla_{zz}^2 F(z, x) + \frac{1}{(p-1)!} D^{p+1} f(x) [z]^{p-1} - \nabla^2 \left[ \frac{L_p}{p!} \|z\|^{p+1} \right]. \end{aligned}$$

Therefore, we have

$$(\nabla_{zz}^2 F(z, x))^{-1} (\nabla_{zx}^2 F(z, x)) = I + (\nabla_{zz}^2 F(z, x))^{-1} \left( \frac{D^{p+1} f(x) [z]^{p-1}}{(p-1)!} - \nabla^2 \left[ \frac{L_p}{p!} \|z\|^{p+1} \right] \right).$$

and

$$\begin{aligned} \left\| (\nabla_{zz}^2 F(z, x))^{-1} (\nabla_{zx}^2 F(z, x)) \right\| &\leq 1 + \frac{p!}{L_p \|z\|^{p-1}} \left\| \frac{D^{p+1} f(x) [z]^{p-1}}{(p-1)!} - \nabla^2 \left[ \frac{L_p}{p!} \|z\|^{p+1} \right] \right\| \\ &\leq 1 + \frac{p!}{L_p \|z\|^{p-1}} \left( \frac{L_p}{(p-1)!} \|z\|^{p-1} + \frac{L_p \cdot (p+1)p}{p!} \cdot \|z\|^{p-1} \right) \\ &= (p+1)^2 \end{aligned}$$

where we used (18) and smoothness for the second inequality. Now, (17) and Lemma 4.7 below show

$$\left\| \frac{d}{d\theta} z_\theta \right\| \leq (p+1)^2 \cdot \|y_k - x_k\| \leq 5(p+1)^2 \cdot \|x^*\|.$$

□

**Lemma 4.2.** *We have that  $\|z_\theta\| \leq 12p^3 \|x^*\|$  for all  $0 \leq \theta \leq 1$ .*

*Proof.* By doing a Taylor expansion of the function  $f$ , one obtains:

$$f_p(\tilde{x}_\theta + z_\theta, \tilde{x}_\theta) \geq f(\tilde{x}_\theta + z_\theta) - \frac{L_p}{(p+1)!} \|z_\theta\|^{p+1}.$$

Hence, we have that

$$F(z_\theta, \tilde{x}_\theta) = f_p(\tilde{x}_\theta + z_\theta, \tilde{x}_\theta) + \frac{L_p}{p!} \|z_\theta\|^{p+1} \geq f(\tilde{x}_\theta + z_\theta) + \frac{L_p \cdot p}{(p+1)!} \|z_\theta\|^{p+1}. \quad (19)$$

Rearranging the term, we have that

$$\|z_\theta\|^{p+1} \leq \frac{(p+1)!}{L_p \cdot p} \cdot (F(z_\theta, \tilde{x}_\theta) - \min_x f(x)) \leq \frac{(p+1)!}{L_p \cdot p} \cdot (f(\tilde{x}_\theta) - \min_x f(x))$$

where we used that  $F(z_\theta, \tilde{x}_\theta) \leq F(0, \tilde{x}_\theta) = f(\tilde{x}_\theta)$ .

For  $\theta = 1$ , we have  $\tilde{x}_\theta = y_k$  and hence

$$\|z_1\|^{p+1} \leq \frac{(p+1)!}{L_p \cdot p} (f(y_k) - \min_x f(x)) \leq \frac{(p+1)!}{2p \cdot A_k \cdot L_p} \|x^*\|^2$$

where we used (8) at the end. Using Lemma 4.1 and Young's inequality, we have

$$\begin{aligned} \|z_\theta\| &\leq \left( \frac{(p+1)!}{2p \cdot A_k \cdot L_p} \right)^{\frac{1}{p+1}} \|x^*\|^{\frac{2}{p+1}} + 5(p+1)^2 \cdot \|x^*\| \\ &\leq \frac{2}{p+1} \|x^*\| + \frac{p-1}{p+1} \left( \frac{(p+1)!}{2p \cdot A_k \cdot L_p} \right)^{\frac{1}{p-1}} + 5(p+1)^2 \|x^*\|. \end{aligned}$$

Using  $A_k \geq \frac{k^{\frac{3p+1}{2}}}{c_p \cdot L_p \cdot \|x^*\|^{p-1}} \geq \frac{1}{c_p \cdot L_p \cdot \|x^*\|^{p-1}}$  and  $c_p = \frac{2^{p-1}(p+1)^{\frac{3p+1}{2}}}{(p-1)!}$ , we have

$$\|z_\theta\| \leq \left( \frac{2}{p+1} + \frac{p-1}{p+1} \left( \frac{(p+1)! \cdot c_p}{2p} \right)^{\frac{1}{p-1}} + 5(p+1)^2 \right) \|x^*\| \leq 12p^3 \|x^*\|.$$

□

Next, we have a lower bound of  $\|z_\theta\|$ . We also prove Lipschitzness of  $\theta \mapsto f(y_\theta)$ .

**Lemma 4.3.** *We have*

$$\|z_\theta\|^p \geq \frac{p!}{L_p \cdot (p+2) \cdot (12p^3 + 4) \|x^*\|} (f(y_\theta) - f(x^*)).$$

Furthermore  $\theta \mapsto f(y_\theta)$  is Lipschitz, with Lipschitz constant upper bounded by

$$L_p \cdot (12p^3 \|x^*\|)^{p+1}.$$

*Proof.* By the optimality of  $z_\theta$ , we have that

$$\nabla_z f_p(\tilde{x}_\theta + z_\theta, \tilde{x}_\theta) + \frac{L_p \cdot (p+1)}{p!} \|z_\theta\|^{p-1} z_\theta = 0.$$

By doing a Taylor expansion of the gradient function, one obtains:

$$\|\nabla_z f_p(\tilde{x}_\theta + z_\theta, \tilde{x}_\theta) - \nabla f(\tilde{x}_\theta + z_\theta)\| \leq \frac{L_p}{p!} \|z_\theta\|^p.$$

Hence, we have  $\|\nabla f(\tilde{x}_\theta + z_\theta)\| \leq \frac{L_p \cdot (p+2)}{p!} \|z_\theta\|^p$  and

$$f(y_\theta) = f(\tilde{x}_\theta + z_\theta) \leq f(x^*) + \frac{L_p \cdot (p+2)}{p!} \|z_\theta\|^p \|\tilde{x}_\theta + z_\theta - x^*\|.$$

Since  $\tilde{x}_\theta$  is convex combination of  $x_k$  and  $y_k$ , Lemma 4.7 shows that  $\|\tilde{x}_\theta - x^*\| \leq 4\|x^*\|$  and Lemma 4.2 shows that  $\|z_\theta\| \leq 12p^3\|x^*\|$ . Combining both, we have  $\|\tilde{x}_\theta + z_\theta - x^*\| \leq (12p^3 + 4)\|x^*\|$  and hence

$$f(y_\theta) - f(x^*) \leq \frac{L_p \cdot (p+2)}{p!} \|z_\theta\|^p \cdot (12p^3 + 4)\|x^*\|.$$

Rearranging gives the first inequality.

For the Lipschitz statement we note that, as above, we have:

$$\begin{aligned} f(y_\theta) - f(y_{\theta'}) &\leq \frac{L_p \cdot (p+2)}{p!} \|z_\theta\|^p \|y_\theta - y_{\theta'}\| \\ &\leq \frac{L_p \cdot (p+2)}{p!} \cdot (12p^3\|x^*\|)^p \cdot (\|\tilde{x}_\theta - \tilde{x}_{\theta'}\| + \|z_\theta - z_{\theta'}\|). \end{aligned}$$

Lemma 4.7 shows that  $\|\tilde{x}_\theta - \tilde{x}_{\theta'}\| = |\theta - \theta'| \cdot \|y_k - x_k\| \leq 5 \cdot \|x^*\| \cdot |\theta - \theta'|$ . Lemma 4.1 shows that  $\|z_\theta - z_{\theta'}\| \leq 5(p+1)^2\|x^*\| \cdot |\theta - \theta'|$ . Combining both, we have

$$f(y_\theta) - f(y_{\theta'}) \leq \frac{L_p \cdot (p+2)}{p!} \cdot (12p^3\|x^*\|)^p \cdot (5 + 5(p+1)^2)\|x^*\| \cdot |\theta - \theta'|.$$

□

We now give a bound on the Lipschitz constant  $\zeta(\theta)$ .

**Lemma 4.4.** *Denote*

$$\omega_p(\theta) = 4(12p^3)^{p+1} \cdot \left( 1 + A_k L_p \|x^*\|^{p-1} + \frac{L_p \|x^*\|^{p+1}}{\Delta(\theta)} \right),$$

and  $\Delta(\theta) = f(y_\theta) - f(x^*)$ . Then one has

$$\left| \frac{d}{d\theta} \log \zeta(\theta) \right| \leq \omega_p(\theta) \cdot \left( 1 + \frac{1}{\zeta(\theta)} + \zeta(\theta) \right).$$

*Proof.* Note that

$$\frac{d}{d\theta} \log \zeta(\theta) = -\frac{2}{1-\theta} - \frac{1}{\theta} + (p-1) \frac{z_\theta \cdot \frac{d}{d\theta} z_\theta}{\|z_\theta\|^2}.$$

Lemma 4.1 shows that

$$\left| \frac{d}{d\theta} \log \zeta(\theta) \right| \leq \frac{2}{1-\theta} + \frac{1}{\theta} + 5(p+1)^2(p-1) \frac{\|x^*\|}{\|z_\theta\|}.$$

The facts that

$$\frac{1}{1-\theta} \leq 1 + \frac{\theta}{(1-\theta)^2} = 1 + \frac{A_k \cdot L_p}{(p-1)! \cdot \zeta(\theta)} \|z_\theta\|^{p-1}$$

and that

$$\frac{1}{\theta} \leq 2 + \frac{(1-\theta)^2}{\theta} = 2 + \frac{(p-1)! \cdot \zeta(\theta)}{A_k \cdot L_p \cdot \|z_\theta\|^{p-1}},$$

yield:

$$\left| \frac{d}{d\theta} \log \zeta(\theta) \right| \leq 4 + \frac{2A_k \cdot L_p}{(p-1)! \cdot \zeta(\theta)} \|z_\theta\|^{p-1} + \frac{(p-1)! \cdot \zeta(\theta)}{A_k \cdot L_p \cdot \|z_\theta\|^{p-1}} + 5(p+1)^2(p-1) \frac{\|x^*\|}{\|z_\theta\|}.$$

It only remains to plug in Lemma 4.2 and Lemma 4.3 as follows: For the second term, we have

$$\frac{2A_k \cdot L_p}{(p-1)! \cdot \zeta(\theta)} \|z_\theta\|^{p-1} \leq \frac{2A_k \cdot L_p \cdot (12p^3 \|x^*\|)^{p-1}}{\zeta(\theta)}.$$

For the third term, we have

$$\begin{aligned} \frac{(p-1)! \cdot \zeta(\theta)}{A_k \cdot L_p \cdot \|z_\theta\|^{p-1}} &\leq \frac{(p-1)! \cdot 12p^3 \|x^*\|}{A_k \cdot L_p \cdot \|z_\theta\|^p} \cdot \zeta(\theta) \\ &\leq \frac{(p-1)! \cdot 12p^3 \|x^*\| L_p \cdot (p+2) \cdot (12p^3 + 4) \|x^*\|}{A_k \cdot L_p p! \cdot \Delta(\theta)} \cdot \zeta(\theta) \\ &\leq 4 \cdot \frac{(12p^3 \|x^*\|)^2}{A_k \cdot \Delta(\theta)} \cdot \zeta(\theta). \end{aligned}$$

Using  $A_k \geq \frac{k^{\frac{3p+1}{2}}}{c_p \cdot L_p \cdot \|x^*\|^{p-1}} \geq \frac{1}{c_p \cdot L_p \cdot \|x^*\|^{p-1}}$  and  $c_p = \frac{2^{p-1}(p+1)^{\frac{3p+1}{2}}}{(p-1)!}$ , we have

$$\begin{aligned} \frac{(p-1)! \cdot \zeta(\theta)}{A_k \cdot L_p \cdot \|z_\theta\|^{p-1}} &\leq \frac{2^{p+1}(p+1)^{\frac{3p+1}{2}} L_p \cdot \|x^*\|^{p-1} \cdot (12p^3 \|x^*\|)^2}{(p-1)! \Delta(\theta)} \cdot \zeta(\theta) \\ &\leq 2^{p+1}(p+1)^{\frac{3p+1}{2}} (12p^3)^2 \cdot \frac{L_p \cdot \|x^*\|^{p+1}}{\Delta(\theta)} \cdot \zeta(\theta) \\ &\leq 4 \cdot (12p^3)^{p+1} \cdot \frac{L_p \cdot \|x^*\|^{p+1}}{\Delta(\theta)} \cdot \zeta(\theta). \end{aligned}$$

For the last term, we have

$$\begin{aligned} 5(p+1)^2(p-1) \frac{\|x^*\|}{\|z_\theta\|} &\leq 5(p+1)^2(p-1) \frac{(12p^3 \|x^*\|)^{p-1} \|x^*\|}{\|z_\theta\|^p} \\ &\leq 5(p+1)^3 \cdot (12p^3 \|x^*\|)^{p-1} \cdot \frac{L_p \cdot (p+2) \cdot (12p^3 + 4) \|x^*\|^2}{p! \cdot \Delta(\theta)} \\ &\leq 4 \cdot (12p^3)^{p+1} \cdot \frac{L_p \cdot \|x^*\|^{p+1}}{\Delta(\theta)}. \end{aligned}$$

Combining all terms, we have the result

$$\left| \frac{d}{d\theta} \log \zeta(\theta) \right| \leq 4 + \frac{2A_k \cdot L_p \cdot (12p^3 \|x^*\|)^{p-1}}{\zeta(\theta)} + 4 \cdot (12p^3)^{p+1} \cdot \frac{L_p \cdot \|x^*\|^{p+1}}{\Delta(\theta)} \cdot (\zeta(\theta) + 1)$$

justifying the claimed upper bound.  $\square$

The next lemma is a straightforward calculus exercise which allows to us to analyze binary search with guarantees of the form given in Lemma 4.4.

**Lemma 4.5.** *Let  $g : [0, 1] \rightarrow \mathbb{R}_+$  and  $\theta^* \in [0, 1]$  such that  $g(\theta^*) = \frac{7}{12}$ . Let  $\omega \geq 0$  such that any  $\theta \in [0, 1]$  with  $|\theta - \theta^*| \leq \frac{1}{40\omega}$  satisfies*

$$\left| \frac{d}{d\theta} \log g(\theta) \right| \leq \omega \cdot \left( 1 + \frac{1}{g(\theta)} + g(\theta) \right).$$

Then one also has  $g(\theta) \in [\frac{1}{2}, \frac{2}{3}]$ .

*Proof.* Let  $h$  be the largest number such that  $|\theta - \theta^*| \leq h$  implies  $g(\theta) \in [\frac{1}{2}, \frac{2}{3}]$ . It suffices to show  $h \geq \frac{1}{40\omega}$ . Proceed by contradiction and suppose that  $h \leq \frac{1}{40\omega}$ . For any  $\theta$  such that  $|\theta - \theta^*| \leq h$ , by the assumption on  $g$  and  $h$ , we have

$$\left| \frac{d}{d\theta} g(\theta) \right| \leq \omega \cdot (g(\theta) + 1 + g^2(\theta)) \leq \omega \cdot \left( \frac{2}{3} + 1 + \left( \frac{2}{3} \right)^2 \right) = \frac{19}{9}\omega.$$

Hence, for any  $\theta$  such that  $|\theta - \theta^*| \leq h$ , we have  $|g(\theta) - g(\theta^*)| \leq h \cdot \frac{19}{9}\omega < \frac{1}{12}$ . Since  $g$  is continuous and  $g(\theta^*) = \frac{7}{12}$  this contradicts the assumption of  $h$  being the largest. Therefore  $|\theta - \theta^*| \leq \frac{1}{40\omega}$  implies that  $g(\theta) \in [\frac{1}{2}, \frac{2}{3}]$  as desired.  $\square$

Now, we can prove our main theorem of this section.

**Theorem 4.6.** *Let  $\varepsilon > 0$ . At iteration  $k$ , using at most  $30p \log_2 p + \log_2 \left\lceil \frac{L_p \|x^*\|^{p+1}}{\varepsilon} \right\rceil$  calls to the  $p^{\text{th}}$  order Taylor oracle we find either a point  $y$  such that  $f(y) - f(x^*) \leq \varepsilon$  or we find  $\lambda_{k+1}$  that satisfies (13).*

*Proof.* First note that we can assume  $A_k \leq \|x^*\|^2 / (2\varepsilon)$ , for otherwise  $f(y_k) - f(x^*) \leq \varepsilon$  by Lemma 2.3. Now using  $\log_2(1/\delta)$  binary search step on  $\zeta$ , let us find  $\theta$  such that  $|\theta - \theta^*| \leq \delta$  for some  $\theta^*$  with  $\zeta(\theta^*) = \frac{7}{12}$ .

If  $\Delta(\theta) \leq \varepsilon$  then we are done, so let us assume this is not the case. By the Lipschitz constant bound from Lemma 4.3, as well as choosing  $\delta$  smaller than  $\varepsilon/2$  divided by this Lipschitz constant, we obtain that  $\Delta(\theta') \geq \varepsilon/2$  for any  $\theta'$  such that  $|\theta - \theta'| \leq 2\delta$  (so in particular for any  $\theta'$  such that  $|\theta' - \theta^*| \leq \delta$ ). We now want to apply Lemma 4.5 to conclude that  $\zeta(\theta) \in [\frac{1}{2}, \frac{2}{3}]$ . For this we need to compute a value for  $\omega$  using Lemma 4.4 (and we will want  $\delta$  small enough so that  $\delta \leq \frac{1}{40\omega}$ ). One can easily verify that the following value of  $\omega$  works given the above:

$$\begin{aligned} \omega &\leq 4(12p^3)^{p+1} \cdot \left( 1 + A_k L_p \|x^*\|^{p-1} + \frac{L_p \|x^*\|^{p+1}}{\Delta(\theta)} \right) \\ &\leq 4(12p^3)^{p+1} \cdot \left( 1 + \frac{\|x^*\|^2}{2\varepsilon} L_p \|x^*\|^{p-1} + \frac{L_p \|x^*\|^{p+1}}{\varepsilon/2} \right) \\ &\leq 16 \cdot (12p^3)^{p+1} \cdot \left\lceil \frac{L_p \|x^*\|^{p+1}}{\varepsilon} \right\rceil. \end{aligned}$$

Hence we can choose

$$\frac{1}{\delta} = 640 \cdot (12p)^{3(p+1)} \cdot \left\lceil \frac{L_p \|x^*\|^{p+1}}{\varepsilon} \right\rceil \leq p^{30p} \cdot \left\lceil \frac{L_p \|x^*\|^{p+1}}{\varepsilon} \right\rceil$$

and binary search finishes in  $\log_2(1/\delta) = 30p \log_2 p + \log_2 \left\lceil \frac{L_p \|x^*\|^{p+1}}{\varepsilon} \right\rceil$  steps.  $\square$

Finally, we give the bound for  $\|x_k - x^*\|$  and  $\|y_k - x^*\|$ .

**Lemma 4.7.** *We have that  $\|x_k - x^*\| \leq \|x^*\|$  and  $\|y_k - x^*\| \leq 4\|x^*\|$  for all  $k$ .*

*Proof.* From Lemma 2.5 we have

$$\psi_{k+1}(x_{k+1}) - A_{k+1}f(y_{k+1}) \geq \sum_{i=1}^{k+1} \frac{A_i}{2\lambda_i} \left( (1 - \sigma^2) \|y_i - \tilde{x}_{i-1}\|^2 \right)$$

Since from Lemma 2.2

$$\psi_{k+1}(x_{k+1}) + \frac{1}{2} \|x^* - x_{k+1}\|^2 = \psi_{k+1}(x^*) \leq A_{k+1}f(x^*) + \frac{1}{2} \|x^*\|^2$$

altogether this gives

$$\sum_{i=1}^{k+1} \frac{A_i}{2\lambda_i} \left( (1 - \sigma^2) \|y_i - \tilde{x}_{i-1}\|^2 \right) \leq A_{k+1}(f^* - f(y_{k+1})) + \frac{1}{2} \|x^*\|^2 - \frac{1}{2} \|x^* - x_{k+1}\|^2$$

therefore we have that  $\|x_k - x^*\| \leq \|x^*\|$  for all  $k$ . Let  $D_k = \|y_k - x^*\|$ . Using  $\tilde{x}_k = \frac{A_k}{A_{k+1}}y_k + \frac{a_{k+1}}{A_{k+1}}x_k$ , we have

$$\|\tilde{x}_k - x^*\| \leq \frac{A_k}{A_{k+1}}D_k + \frac{a_{k+1}}{A_{k+1}}\|x^*\|.$$

Hence, we have  $D_{k+1} \leq \frac{A_k}{A_{k+1}}D_k + \frac{a_{k+1}}{A_{k+1}}\|x^*\| + \|y_{k+1} - \tilde{x}_k\|$ . Rescaling and summing over  $k$ , we have

$$\begin{aligned} D_{k+1} &\leq \|x^*\| + \|y_{k+1} - \tilde{x}_k\| + \frac{A_k}{A_{k+1}}\|y_k - \tilde{x}_{k-1}\| + \frac{A_{k-1}}{A_{k+1}}\|y_{k-1} - \tilde{x}_{k-2}\| + \dots \\ &\leq \|x^*\| + \frac{1}{A_{k+1}} \sum_{j=1}^{k+1} A_j \|y_j - \tilde{x}_{j-1}\| \\ &\leq \|x^*\| + \frac{\sqrt{\sum_{j=1}^{k+1} A_j \lambda_j}}{A_{k+1}} \sqrt{\sum_{j=1}^{k+1} \frac{A_j}{\lambda_j} \|y_j - \tilde{x}_{j-1}\|^2} \\ &\leq \|x^*\| + \frac{\sqrt{\sum_{j=1}^{k+1} \lambda_j}}{\sqrt{A_{k+1}}} \sqrt{\frac{\|x^*\|^2}{1 - \sigma^2}} \\ &\leq 4\|x^*\| \end{aligned}$$

where we used  $A_j$  is increasing and (6) in the second to last equation, and Lemma 2.6 and  $\sigma = \frac{1}{2}$  for the last.  $\square$

## References

- [1] N. Agarwal and E. Hazan. Lower bounds for higher-order convex optimization. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 774–792. PMLR, 2018.
- [2] Y. Arjevani, O. Shamir, and R. Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 2018.

- [3] A. Gasnikov, E. Gorbunov, D. Kovalev, A. Mohhamed, and E. Chernousova. The global rate of convergence for optimal tensor methods in smooth convex optimization. *Arxiv preprint arXiv:1809.00382*, 2018.
- [4] B. Jiang, H. Wang, and S. Zhang. An optimal high-order tensor method for convex optimization. *Arxiv preprint arXiv:1812.06557*, 2018.
- [5] R. D. C. Monteiro and B. F. 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.
- [6] A. Nemirovski. Orth-method for smooth convex optimization. *Izvestia AN SSSR, Ser. Tekhnicheskaya Kibernetika*, 2, 1982.
- [7] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.
- [8] Y. Nesterov. A method of solving a convex programming problem with convergence rate  $o(1/k^2)$ . *Soviet Mathematics Doklady*, 27(2):372–376, 1983.
- [9] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.
- [10] Y. Nesterov. Implementable tensor methods in unconstrained convex optimization. Core discussion papers, 2018. URL <https://ideas.repec.org/p/cor/louvco/2018005.html>.