

# INEXACT HIGH-ORDER PROXIMAL- POINT METHODS WITH AUXILIARY SEARCH PROCEDURE

Yurii Nesterov

REPRINT | 3251

## CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: [lidam-library@uclouvain.be](mailto:lidam-library@uclouvain.be)

<https://uclouvain.be/en/research-institutes/lidam/core/core-reprints.html>

## INEXACT HIGH-ORDER PROXIMAL-POINT METHODS WITH AUXILIARY SEARCH PROCEDURE\*

YURII NESTEROV†

**Abstract.** In this paper, we complement the framework of bilevel unconstrained minimization (BLUM) [Y. Nesterov, *Math. Program.*, to appear] by a new  $p$ th-order proximal-point method convergent as  $O(k^{-(3p+1)/2})$ , where  $k$  is the iteration counter. As compared with [Y. Nesterov, *Math. Program.*, to appear], we replace the auxiliary line search by a segment search. This allows us to bound its complexity by a logarithm of the desired accuracy. Each step in this search needs an approximate computation of a high-order proximal-point operator. Under the assumption on boundedness of  $(p+1)$ th derivative of the objective function, this can be done by one step of the  $p$ th-order augmented tensor method. In this way, for  $p = 2$ , we get a new second-order method with the rate of convergence  $O(k^{-7/2})$  and logarithmic complexity of the auxiliary search at each iteration. Another possibility is to compute the proximal-point operator by lower-order minimization methods. As an example, for  $p = 3$ , we consider the upper-level process convergent as  $O(k^{-5})$ . Assuming the boundedness of fourth derivative, an appropriate approximation of the proximal-point operator can be computed by a second-order method in logarithmic number of iterations. This combination gives a second-order scheme with much better complexity than the existing theoretical limits.

**Key words.** convex optimization, tensor methods, proximal-point operator, lower complexity bounds, optimal methods

**AMS subject classification.** 90C25

**DOI.** 10.1137/20M134705X

**1. Introduction. Motivation.** In this paper, we develop a new and efficient proximal-point method, which can be used at the upper level in the framework of bilevel unconstrained minimization (BLUM) [22]. In this approach, we need to decide first on the order of the upper-level procedure, and after that we choose a numerical scheme for an approximate computation of the high-order proximal-point operator. This structure allows a significant flexibility in applying the lower-order methods to the problem classes which were traditionally out of their scope. Consequently, the lower-order methods can now demonstrate a much higher rate of convergence breaking down the traditional theoretical limits.

In the first time, the possibility of using the second-order schemes for minimizing functions with a Lipschitz-continuous third derivative was demonstrated in [21]. The optimization method in this paper is a second-order implementation of the third-order scheme with the convergence rate  $O(k^{-4})$ , where  $k$  is the iteration counter. Note that the traditional upper bound for the rate of convergence of the second-order methods is  $O(k^{-7/2})$  [1, 2, 5]. Of course, the result in [21] does not contradict the theory since it is related only to a subclass of functions, which is usually considered as the natural problem class for the second-order methods. (These are functions with Lipschitz-continuous Hessians.) However, it destroys the traditional one-to-one correspondence between the problem classes and the orders of applicable minimization

\*Received by the editors June 22, 2020; accepted for publication (in revised form) May 31, 2021; published electronically November 8, 2021.

<https://doi.org/10.1137/20M134705X>

**Funding:** This paper has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 788368). It was also supported by Multidisciplinary Institute in Artificial intelligence MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

†Department of Mathematical Engineering, Universite catholique de Louvain, Louvain-La-Neuve, Belgium, 1348 (Yurii.Nesterov@uclouvain.be).

schemes. We cannot call a method optimal anymore without mentioning explicitly the corresponding problem class.

In the next paper [22], in order to explain the phenomena of [21], we proposed the framework BLUM based on the notion of a *high-order proximal-point operator*. The proximal approximation of function  $f(\cdot)$ , defined by

$$(1.1) \quad \varphi_\lambda(x) = \min_y \left\{ f(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}, \quad \lambda > 0,$$

was introduced by Moreau [14]. Later on, Martinet [13] proposed the first proximal-point method based on this operation. The importance of this construction for computational practice was questionable up to the developments of Rockafellar [25], who used the proximal-point iteration in the dual space for justifying the augmented Lagrangians. This dual scheme was accelerated by Güler [9], who introduced in this method some elements of the fast gradient method from [15]. Some attempts were made by Teboulle and Iusem, Svaiter, and Teboulle [26, 10] in studying proximal-point iteration with a nonquadratic kernel, concentrating mainly on strongly convex functions. However, for decades this idea was considered as a theoretical tool, which cannot be used in the efficient optimization algorithms.

In [22], it was suggested to regularize the objective function in (1.1) by a power of Euclidean norm  $\|\cdot\|^{p+1}$  with  $p \geq 1$ . We call the corresponding proximal step the  *$p$ th-order proximal-point operation*.<sup>1</sup> This terminology is justified by two facts. Firstly, it was shown that the corresponding simple proximal-point method converges as  $O(k^{-p})$ . The rate of convergence of the accelerated version of this method is  $O(k^{-(p+1)})$ . In both cases, we can use appropriate approximations of the proximal point.

Secondly, it appears that this approximation can be computed by one step of the  $p$ th-order tensor method provided that the  $p$ th derivative of the objective function is Lipschitz.

Another important contribution of [22] is the definition of the proximal-point operator with *line search*. Using this auxiliary operation, it is possible to achieve the rate of convergence of the order  $O(k^{-(3p+1)/2})$ . However, in [22] it was assumed that the proximal-point line search admits an exact computation. The main goal of this paper is to relax this assumption and justify an implementable version of this method.

**Contents.** In this paper, we study the methods for solving convex unconstrained optimization problem. It is organized as follows. In section 2, we introduce a  $p$ th-order proximal-point operator with *segment search* ( $p \geq 1$ ) and study its properties. We justify also the rate of convergence  $O(k^{-(3p+1)/2})$  of the accelerated method under the assumption that the corresponding proximal-point operators admit exact computation. At this stage, we do not need any assumptions on the objective function besides convexity. We discuss also the reason why the first-order proximal-point methods do not need auxiliary search.

In section 3, we define the acceptable approximations for the proximal-point operators and present a conceptual scheme of the accelerated method. We do not specify yet the segment search procedure, assuming only that it satisfies some stopping criterion. The rate of convergence of this scheme is also  $O(k^{-(3p+1)/2})$ .

In section 4, we describe the rules of a bisection procedure and show that its length is bounded by logarithmic factors. At each iteration of this scheme, we need to compute approximately a  $p$ th-order proximal point.

In section 5, we discuss different strategies for computing approximately the high-order proximal points. In section 5.1, we show that this can be done by one step of

<sup>1</sup>Thus, the standard construction (1.1) corresponds to the first-order proximal points.

the  $p$ th-order augmented tensor method. For that, we need to assume boundedness of the  $(p+1)$ th derivative of the objective function. In this way, we get the second-order methods for functions with a bounded third derivative with convergence rate  $O(k^{-7/2})$ .

Finally, in section 5.2 we present a second-order implementation of the third-order accelerated proximal-point method with segment search. It has the convergence rate  $O(k^{-5})$ . To the best of our knowledge, this is the fastest rate of convergence known so far for the second-order methods without assumptions on uniform convexity. The number of auxiliary computations of the Hessians at each iteration of this scheme is bounded by logarithmic factors.<sup>2</sup>

**Notation 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  $x \in \mathbb{E}$ .

Let us measure distances in  $\mathbb{E}$  and  $\mathbb{E}^*$  in a Euclidean norm. For that, using a self-adjoint positive-definite operator  $B : \mathbb{E} \rightarrow \mathbb{E}^*$  (notation  $B = B^* \succ 0$ ), we define

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

Sometimes, it will be convenient to treat  $x \in \mathbb{E}$  as a linear operator from  $\mathbb{R}$  to  $\mathbb{E}$  and  $x^*$  as a linear operator from  $\mathbb{E}^*$  to  $\mathbb{R}$ . In this case,  $xx^*$  is a linear operator from  $\mathbb{E}^*$  to  $\mathbb{E}$ , acting as follows:

$$(xx^*)g = \langle g, x \rangle x \in \mathbb{E}, \quad g \in \mathbb{E}^*.$$

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

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

We denote by  $\ell_{\bar{x}}(\cdot)$  the linear model of convex function  $f(\cdot)$  at point  $\bar{x} \in \mathbb{E}$ :

$$(1.2) \quad \ell_{\bar{x}}(x) \stackrel{\text{def}}{=} f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \leq f(x), \quad x \in \mathbb{E}.$$

Using the above norm, we can define the standard Euclidean proximal functions

$$d_{p+1}(x) = \frac{1}{p+1} \|x\|^{p+1}, \quad x \in \mathbb{E},$$

where  $p \geq 1$  is an integer parameter. These functions have the following derivatives:

$$(1.3) \quad \begin{aligned} \nabla d_{p+1}(x) &= \|x\|^{p-1} Bx, \quad x \in \mathbb{E}, \\ \nabla^2 d_{p+1}(x) &= \|x\|^{p-1} B + (p-1) \|x\|^{p-3} Bxx^*B \\ &\preceq p \|x\|^{p-1} B. \end{aligned}$$

Note that function  $d_{p+1}(\cdot)$  is uniformly convex (see, for example, Lemma 4.2.3 in [18]):

$$(1.4) \quad d_{p+1}(y) \geq d_{p+1}(x) + \langle \nabla d_{p+1}(x), y - x \rangle + \frac{1}{p+1} \left(\frac{1}{2}\right)^{p-1} \|y - x\|^{p+1}, \quad x, y \in \mathbb{E}.$$

In this paper, we also work with directional derivatives. For  $p \geq 1$ , denote by

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

<sup>2</sup>On the level of nonconvex optimization, an acceleration for the smaller problem classes was already observed in [6].

the directional derivative of order  $p \geq 1$  of the function  $f$  at  $x$  along directions  $h_i \in \mathbb{E}$ ,  $i = 1, \dots, p$ . Note that  $D^p f(x)[\cdot]$  is a *symmetric  $p$ -linear form*. Its *norm* is defined in a standard way:

$$(1.5) \quad \|D^p f(x)\| = \max_{h_1, \dots, h_p} \left\{ \left| D^p f(x)[h_1, \dots, h_p] \right| : \|h_i\| \leq 1, i = 1, \dots, p \right\}.$$

If all directions  $h_1, \dots, h_p$  are the same, we apply the notation

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

Note that, in general, we have (see, for example, Appendix 1 in [23])

$$(1.6) \quad \|D^p f(x)\| = \max_h \left\{ \left| D^p f(x)[h]^p \right| : \|h\| \leq 1 \right\}.$$

**2. Proximal-point methods with exact search.** Consider the following optimization problem:

$$(2.1) \quad \min_{x \in \mathbb{E}} f(x),$$

where  $f(\cdot)$  is a differentiable convex function. Denote by  $x^*$  one of its optimal solutions, and let  $f^* = f(x^*)$ .

All methods presented in this paper are based on the  *$p$ th-order proximal-point operators*, defined as follows:

$$(2.2) \quad \text{prox}_{f/H}^p(\bar{x}) = \arg \min_{x \in \mathbb{E}} \left\{ f_{\bar{x}, H}^p(x) \stackrel{\text{def}}{=} f(x) + Hd_{p+1}(x - \bar{x}) \right\},$$

where  $H > 0$  and the parameter  $p \geq 1$  is called the *order* of this operator.

In [22], it was shown that the rate of convergence of the high-order proximal-point methods can be increased by including in operation (2.2) a *line search*. In this paper, we will see that it is enough to use the proximal-point operators with *segment search*:

$$(2.3) \quad \text{SProx}_{f/H}^p(\bar{x}, \bar{u}) = \arg \min_{\substack{x \in \mathbb{E} \\ \tau \in [0, 1]}} \left\{ f(x) + Hd_{p+1}(x - \bar{x} - \tau \bar{u}) \right\} \in \mathbb{E} \times \mathbb{R}.$$

Denote by  $x(\bar{x}, \bar{u})$  and  $\tau(\bar{x}, \bar{u})$  the solutions of this minimization problem. Since

$$(2.4) \quad \min_{\substack{x \in \mathbb{E} \\ \tau \in [0, 1]}} \left\{ f(x) + Hd_{p+1}(x - \bar{x} - \tau \bar{u}) \right\} = \min_{\substack{y \in \mathbb{E} \\ \tau \in [0, 1]}} \left\{ f(y + \tau \bar{u}) + Hd_{p+1}(y - \bar{x}) \right\},$$

the point  $x_+ = x(\bar{x}, \bar{u})$  and the value  $\tau_+ = \tau(\bar{x}, \bar{u})$  satisfy the first-order optimality condition of the second problem in  $\tau$ ; that is,

$$(2.5) \quad (\tau - \tau_+) \langle \nabla f(x_+), \bar{u} \rangle \geq 0 \quad \forall \tau \in [0, 1].$$

On the other hand, the first-order optimality condition of (2.3) in  $x$  gives us the following equation:

$$\nabla f(x_+) + Hr_+^{p-1} B(x_+ - y_+) = 0,$$

where  $y_+ = \bar{x} + \tau_+ \bar{u}$  and  $r_+ = \|x_+ - y_+\|$ . Therefore,

$$(2.6) \quad \langle \nabla f(x_+), y_+ - x_+ \rangle = Hr_+^{p+1} = \left[ \frac{1}{H} \right]^{\frac{1}{p}} \|\nabla f(x_+)\|_*^{\frac{p+1}{p}}.$$

Let us assume that we can compute the exact value of operator  $\text{SProx}_{f/H}^p(\cdot, \cdot)$ . Then, it is possible to solve (2.1) by the following optimization scheme.

|       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (2.7) | <b>Exact <math>p</math>th-Order Proximal-Point Method With Segment Search</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|       | <p style="text-align: center;"><b>Initialization.</b> Choose <math>x_0 \in \mathbb{E}</math> and <math>H &gt; 0</math>. Define <math>A_0 = 0</math> and</p> $\psi_0(x) = \frac{1}{2}\ x - x_0\ ^2.$                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|       | <p><b>Iteration <math>k \geq 0</math>.</b></p> <ol style="list-style-type: none"> <li>1. Compute <math>v_k = \arg \min_{x \in \mathbb{E}} \psi_k(x)</math>. Set <math>u_k = v_k - x_k</math>.</li> <li>2. Compute <math>(x_{k+1}, \tau_k) = \text{SProx}_{f/H}^p(x_k, u_k)</math> and <math>g_k = \ \nabla f(x_{k+1})\ _*</math>.</li> <li>3. Compute <math>a_{k+1} &gt; 0</math> from the equation <math>\frac{a_{k+1}^2}{A_k + a_{k+1}} = \left[\frac{1}{H}\right]^{\frac{1}{p}} g_k^{\frac{1-p}{p}}</math>.</li> <li>4. Set <math>A_{k+1} = A_k + a_{k+1}</math>, and update <math>\psi_{k+1}(x) = \psi_k(x) + a_{k+1}\ell_{x_{k+1}}(x)</math>.</li> </ol> |

Using the same arguments as in [22], we can prove that the rate of convergence of this scheme is of the order  $O(k^{-(3p+1)/2})$  in terms of the functional residual. In order to see the impact of inexact computation of the proximal step presented in section 3, we give here the full proof of the corresponding statement. It is based on two lemmas on positive numbers justified in the appendix (section 6).

**THEOREM 2.1.** *Let the sequence  $\{x_k\}_{k \geq 0}$  be generated by method (2.7). Then for any  $k \geq 1$  we have*

$$(2.8) \quad f(x_k) - f^* \leq 2^p H R_0^{p+1} \left[1 + \frac{2(k-1)}{p+1}\right]^{-(3p+1)/2},$$

where  $R_0 = \|x_0 - x^*\|$ .

*Proof.* Note that

$$\psi_k^* \stackrel{\text{def}}{=} \min_{x \in \mathbb{E}} \psi_k(x) \stackrel{(1.2)}{\leq} \min_{x \in \mathbb{E}} [A_k f(x) + \frac{1}{2}\|x - x_0\|^2] \leq A_k f(x^*) + \frac{1}{2}\|x^* - x_0\|^2.$$

Let us prove now that for all  $k \geq 0$ , we have

$$(2.9) \quad A_k f(x_k) + B_k \leq \psi_k^*,$$

where  $B_k = \frac{1}{2} \left[\frac{1}{H}\right]^{\frac{1}{p}} \sum_{i=0}^{k-1} A_{i+1} g_i^{\frac{p+1}{p}}$ . For  $k = 0$ , we have  $A_0 = B_0 = \psi_0^* = 0$ . Hence, in this case (2.9) is valid. Assume it is valid for some  $k \geq 0$ . Then,

$$\begin{aligned} \psi_{k+1}(x) &= \psi_k(x) + a_{k+1}\ell_{x_{k+1}}(x) \\ &\stackrel{(2.9)}{\geq} A_k f(x_k) + B_k + \frac{1}{2}\|x - v_k\|^2 + a_{k+1}\ell_{x_{k+1}}(x) \\ &\geq A_k f(x_k) + B_k - \frac{1}{2}a_{k+1}^2 g_k^2 + a_{k+1}\ell_{x_{k+1}}(v_k). \end{aligned}$$

Denote  $y_k = x_k + \tau_k(v_k - x_k)$  and  $\hat{\tau}_k = \frac{a_{k+1}}{A_{k+1}} \in (0, 1]$ . Then

$$\begin{aligned}
 & A_k f(x_k) + a_{k+1} \ell_{x_{k+1}}(v_k) \\
 & \geq A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), A_k x_k + a_{k+1} v_k - A_{k+1} x_{k+1} \rangle \\
 & = A_{k+1} [f(x_{k+1}) + \langle \nabla f(x_{k+1}), (\hat{\tau}_k - \tau_k) u_k + y_k - x_{k+1} \rangle] \\
 (2.10) \quad & \stackrel{(2.5)}{\geq} A_{k+1} [f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle] \\
 & \stackrel{(2.6)}{=} A_{k+1} \left[ f(x_{k+1}) + \left[ \frac{1}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} \right].
 \end{aligned}$$

Thus, using the rule of step 3 in (2.7), we get

$$\psi_{k+1}^* \geq A_{k+1} f(x_{k+1}) + B_k - \frac{1}{2} a_{k+1}^2 g_k^2 + A_{k+1} \left[ \frac{1}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} = A_{k+1} f(x_{k+1}) + B_{k+1},$$

and (2.9) is proved.

From (2.9) and step 3 of method (2.7), we get three important consequences. Namely, for all  $k \geq 1$ , we have

$$\begin{aligned}
 (A) : \quad & f(x_k) - f^* \leq \frac{R_0^2}{2A_k}, \\
 (2.11) \quad (B) : \quad & \left[ \frac{1}{H} \right]^{\frac{1}{p}} \sum_{i=1}^k A_i g_{i-1}^{\frac{p+1}{p}} \leq R_0^2, \\
 (C) : \quad & A_k \left[ \frac{1}{H} \right]^{\frac{1}{p}} g_{k-1}^{\frac{1-p}{p}} = a_k^2.
 \end{aligned}$$

Let us obtain from this information a lower bound for the rate of growth of coefficients  $\{A_k\}_{k \geq 1}$ . This can be done in three steps.

First of all, note that in view of (2.11)<sub>C</sub>, we have

$$A_{k+1}^{1/2} - A_k^{1/2} = \frac{a_{k+1}}{A_{k+1}^{1/2} + A_k^{1/2}} \geq \frac{a_{k+1}}{2A_{k+1}^{1/2}} = \frac{1}{2} \left[ \frac{1}{H} \right]^{\frac{1}{2p}} g_k^{\frac{1-p}{2p}}.$$

Hence, denoting  $\xi_i = g_{i-1}^{\frac{p-1}{2p}}$ ,  $i \geq 1$ , we get  $A_k \geq \frac{1}{4} \left[ \frac{1}{H} \right]^{\frac{1}{p}} \left( \sum_{i=1}^k \frac{1}{\xi_i} \right)^2$ . On the other hand, in this notation, by (2.11)<sub>B</sub>, we have

$$\sum_{i=1}^k A_i \xi_i^\gamma \leq H^{\frac{1}{p}} R_0^2,$$

where  $\gamma = \frac{2(p+1)}{p-1}$ . Therefore, in view of Lemma 6.1, we have

$$\sum_{i=1}^k \frac{1}{\xi_i} \geq \left[ \frac{1}{H^{1/p} R_0^2} \right]^{1/\gamma} \left( \sum_{i=1}^k A_i^{\frac{1}{1+\gamma}} \right)^{\frac{1+\gamma}{\gamma}} = H^{-\frac{p-1}{2p(p+1)}} R_0^{-\frac{p-1}{p+1}} \left( \sum_{i=1}^k A_i^{\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{2(p+1)}}.$$

Thus, we get

$$A_k \geq \frac{1}{4} H^{-\frac{2}{p+1}} R_0^{-\frac{2(p-1)}{p+1}} \left( \sum_{i=1}^k A_i^{\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{p+1}}.$$



Applying now Lemma 6.2 with

$$\theta = \frac{1}{4}H^{-\frac{2}{p+1}}R_0^{-\frac{2(p-1)}{p+1}}, \quad \alpha = \frac{p-1}{3p+1}, \quad \beta = \frac{3p+1}{p+1},$$

we get  $A_k \geq (\frac{1}{4})^{\frac{p+1}{2}} \frac{1}{H} R_0^{1-p} [\frac{2k+p-1}{p+1}]^{\frac{3p+1}{2}}$ . It remains to use (2.11)<sub>A</sub>.  $\square$

*Remark 1.* In the proof of Theorem 2.1, it is easy to find the place where we need the segment search. This is the last inequality in the chain (2.10), where we apply the lower bound

$$\langle \nabla f(x_{k+1}), (\hat{\tau}_k - \tau_k)u_k \rangle \stackrel{(2.5)}{\geq} 0.$$

However, if  $p = 1$ , then we can achieve this goal in a simpler way. Indeed, in this case the new value of  $a_{k+1}$  at step 3 of the method (2.7) can be obtained *before* computing the point  $x_{k+1}$ . Consequently, the simple choice  $\tau_{k+1} = \hat{\tau}_{k+1}$  ensures validity of the above proof. This possibility is indeed used in fast gradient methods (e.g., [16]), where no line search is needed.

**3. Inexact proximal point computation.** Very often, the proximal-point operator (2.2) cannot be computed in a closed form. Instead, we are going to use an approximate solution of this problem obtained by an appropriate optimization procedure. Let us introduce the set of acceptable solutions to (2.2); that is,

$$(3.1) \quad \mathcal{A}_H^p(\bar{x}, \beta) = \left\{ x \in \mathbb{E} : \|\nabla f_{\bar{x}, H}^p(x)\|_* \leq \beta \|\nabla f(x)\|_* \right\},$$

where  $\beta \in [0, 1]$  is a tolerance parameter. Note that  $\text{prox}_{f/H}^p(\bar{x}) \in \mathcal{A}_H^p(\bar{x}, \beta)$ . However, since  $\nabla f_{\bar{x}, H}^p(\bar{x}) = \nabla f(\bar{x})$ , we see that  $\bar{x} \notin \mathcal{A}_H^p(\bar{x}, \beta)$  unless  $\bar{x} = x^*$ .

The main properties of the approximate proximal step are given in the following statement (see [22]).

LEMMA 3.1. *Let  $T \in \mathcal{A}_H^p(\bar{x}, \beta)$ . Then*

$$(3.2) \quad (1 - \beta)\|\nabla f(T)\|_* \leq H\|T - \bar{x}\|^p \leq (1 + \beta)\|\nabla f(T)\|_*,$$

$$(3.3) \quad \langle \nabla f(T), \bar{x} - T \rangle \geq \frac{H}{1+\beta}\|T - \bar{x}\|^{p+1}.$$

Moreover, if  $\beta \leq \frac{1}{p}$ , then

$$(3.4) \quad \langle \nabla f(T), \bar{x} - T \rangle \geq \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} \|\nabla f(T)\|_*^{\frac{p+1}{p}}.$$

We will need one more property of the approximate proximal step.

LEMMA 3.2. *Let  $T \in \mathcal{A}_H^p(\bar{x}, \beta)$  with  $\beta \in [0, \frac{1}{2})$ . Then*

$$(3.5) \quad \|T - x^*\|^2 \leq \frac{(1-\beta)^2}{1-2\beta}\|\bar{x} - x^*\|^2.$$

In particular, if  $\beta \leq \frac{3}{8}$ , then  $\|T - x^*\| \leq \frac{5}{4}\|\bar{x} - x^*\|$ .

*Proof.* Denote  $r = \|T - \bar{x}\|$ . Then

$$\begin{aligned} \|T - x^*\|^2 &= \|\bar{x} - x^*\|^2 + 2\langle B(T - \bar{x}), T - x^* \rangle - r^2 \\ &\stackrel{(1.3)}{=} \|\bar{x} - x^*\|^2 + \frac{2}{Hr^{p-1}}\langle \nabla f_{\bar{x}, H}^p(T) - \nabla f(T), T - x^* \rangle - r^2 \\ &\stackrel{(3.1)}{\leq} \|\bar{x} - x^*\|^2 + \frac{2\beta\|\nabla f(T)\|_*}{Hr^{p-1}}\|T - x^*\| - r^2 \\ &\stackrel{(3.2)}{\leq} \|\bar{x} - x^*\|^2 + \frac{2\beta r}{1-\beta}\|T - x^*\| - r^2. \end{aligned}$$

Maximizing the right-hand side of the last expression in  $r$ , we get (3.5). The remaining inequality follows from (3.5) since its right-hand side is monotone in  $\beta \in [0, \frac{1}{2}]$ .  $\square$

Note that the exact computation of operator  $\text{SProx}_{f/H}^p(\cdot, \cdot)$  is also impossible. Hence, in the remaining part of this section, we analyze an implementable version of method (2.7) based on approximate computations. Consider the following method.

|       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (3.6) | <b>Inexact <math>p</math>th-order Proximal-Point Method (Segment Search)</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|       | <p><b>Initialization.</b> Let <math>x_0 \in \mathbb{E}</math>, <math>\beta \in [0, \frac{3}{3p+2}]</math>, and <math>H &gt; 0</math>. Let <math>A_0 = 0</math> and</p> $\psi_0(x) = \frac{1}{2}\ x - x_0\ ^2.$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|       | <p><b>Iteration <math>k \geq 0</math>.</b></p> <ol style="list-style-type: none"> <li>1. Compute <math>v_k = \arg \min_{x \in \mathbb{E}} \psi_k(x)</math>. Define <math>u_k = v_k - x_k</math>.</li> <li>2. Compute <math>x_k^0 \in \mathcal{A}_H^p(x_k, \beta)</math>. <b>If</b> <math>\langle \nabla f(x_k^0), u_k \rangle \geq 0</math>, <b>then</b> define             <math display="block">\varphi_k(x) = \ell_{x_k^0}(x), \quad x_{k+1} = x_k^0, \quad g_k = \ \nabla f(x_k^0)\ _*.</math> </li> <li>3. <b>Else</b>, compute <math>x_k^1 \in \mathcal{A}_H^p(v_k, \beta)</math>. <b>If</b> <math>\langle \nabla f(x_k^1), u_k \rangle \leq 0</math>, <b>then</b> define             <math display="block">\varphi_k(x) = \ell_{x_k^1}(x), \quad x_{k+1} = x_k^1, \quad g_k = \ \nabla f(x_k^1)\ _*.</math> </li> <li>4. <b>Else</b>, find the values <math>0 \leq \tau_k^1 \leq \tau_k^2 \leq 1</math> with <math>T_k^1 \in \mathcal{A}_H^p(x_k + \tau_k^1 u_k, \beta)</math> and <math>T_k^2 \in \mathcal{A}_H^p(x_k + \tau_k^2 u_k, \beta)</math> satisfying the following conditions:             <math display="block">(a) \quad \beta_k^1 \leq 0 \leq \beta_k^2, \quad (b) \quad \alpha_k(\tau_k^1 - \tau_k^2)\beta_k^1 \leq \frac{1}{2} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}},</math>             where <math>\beta_k^1 = \langle \nabla f(T_k^1), u_k \rangle</math>, <math>\beta_k^2 = \langle \nabla f(T_k^2), u_k \rangle</math>, <math>\alpha_k = \frac{\beta_k^2}{\beta_k^2 - \beta_k^1} \in [0, 1]</math>, and             <math display="block">g_k = \left[ \alpha_k \ \nabla f(T_k^1)\ _*^{\frac{p+1}{p}} + (1 - \alpha_k) \ \nabla f(T_k^2)\ _*^{\frac{p+1}{p}} \right]^{\frac{p}{p+1}}.</math>             Set <math>\varphi_k(x) = \alpha_k \ell_{T_k^1}(x) + (1 - \alpha_k) \ell_{T_k^2}(x)</math> and <math>x_{k+1} = \alpha_k T_k^1 + (1 - \alpha_k) T_k^2</math>. </li> <li>5. Compute <math>a_{k+1} &gt; 0</math> from the equation <math>\frac{a_{k+1}^2}{A_k + a_{k+1}} = \frac{1}{2} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{1-p}{p}}</math>.</li> <li>6. Set <math>A_{k+1} = A_k + a_{k+1}</math>, and update <math>\psi_{k+1}(x) = \psi_k(x) + a_{k+1} \varphi_k(x)</math>.</li> </ol> |

In this scheme, steps 2, 3, and 4 compute an approximate value of the operator  $\text{SProx}_{f/H}^p(\cdot, \cdot)$ , replacing step 2 in method (2.7). Step 4 of method (3.6) is activated only if

$$\langle \nabla f(x_k^0), u_k \rangle \leq 0 \leq \langle \nabla f(x_k^1), u_k \rangle.$$

Thus, the condition (4a) of (3.6) is satisfied for  $\tau_k^1 = 0$  and  $\tau_k^2 = 1$ . Hence, these values can be used for starting the internal search procedure. Note that the coefficient  $\alpha_k$  at this step is chosen to ensure for the vector

$$\nabla \varphi_k \stackrel{\text{def}}{=} \alpha_k \nabla f(T_k^1) + (1 - \alpha_k) \nabla f(T_k^2)$$

the following equality:

$$(3.7) \quad \langle \nabla \varphi_k, u_k \rangle = \alpha_k \beta_k^1 + (1 - \alpha_k) \beta_k^2 = 0.$$

Thus, the estimating function  $\psi_k(\cdot)$  is updated by a convex combination of two linear models. The gradient of this combination is orthogonal to the search direction  $u_k$ .

For analyzing the rate of convergence of method (3.6), we need to prove several auxiliary results.

LEMMA 3.3. *For all  $k \geq 0$ , we have*

$$(3.8) \quad f(x_k) \geq \varphi_k(x_k) \geq f(x_{k+1}) + \frac{1}{2} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}.$$

*Proof.* The first inequality in (3.8) follows from the convexity of function  $f(\cdot)$ . In order to prove the second one, we need to consider separately the ways of constructing the function  $\varphi_k(\cdot)$  used at steps 2, 3, and 4.

1. Let  $\langle \nabla f(x_k^0), u_k \rangle \geq 0$ . Then

$$\varphi_k(x_k) = f(x_{k+1}) + \langle \nabla f(x_k^0), x_k - x_k^0 \rangle \stackrel{(3.4)}{\geq} f(x_{k+1}) + \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}.$$

Thus, the inequality (3.8) is valid.

2. Let  $\langle \nabla f(x_k^1), u_k \rangle \leq 0$ . Then

$$\begin{aligned} \varphi_k(x_k) &= f(x_{k+1}) + \langle \nabla f(x_k^1), x_k - x_k^1 \rangle = f(x_{k+1}) + \langle \nabla f(x_k^1), v_k - u_k - x_k^1 \rangle \\ &\geq f(x_{k+1}) + \langle \nabla f(x_k^1), v_k - x_k^1 \rangle \stackrel{(3.4)}{\geq} f(x_{k+1}) + \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}. \end{aligned}$$

Thus, in this case the inequality (3.8) is also valid.

3. Let us assume now that the function  $\varphi_k(\cdot)$  is formed in accordance to the rules of step 4. Denote  $y_k^1 = x_k + \tau_k^1 u_k$ ,  $y_k^2 = x_k + \tau_k^2 u_k$ , and  $\hat{y}_k = \frac{1}{2}(y_k^1 + y_k^2)$ . Then we have

$$\begin{aligned} \varphi_k(x_k) &= \alpha_k (f(T_k^1) + \langle \nabla f(T_k^1), x_k - T_k^1 \rangle) + (1 - \alpha_k) (f(T_k^2) + \langle \nabla f(T_k^2), x_k - T_k^2 \rangle) \\ &\stackrel{(3.7)}{\geq} f(x_{k+1}) + \alpha_k \langle \nabla f(T_k^1), \hat{y}_k - T_k^1 \rangle + (1 - \alpha_k) \langle \nabla f(T_k^2), \hat{y}_k - T_k^2 \rangle. \end{aligned}$$

Note that

$$\begin{aligned} \langle \nabla f(T_k^1), \hat{y}_k - T_k^1 \rangle &= \langle \nabla f(T_k^1), y_k^1 - \frac{1}{2}(y_k^1 - y_k^2) - T_k^1 \rangle \\ &\stackrel{(3.4)}{\geq} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} \|\nabla f(T_k^1)\|_*^{\frac{p+1}{p}} - \frac{1}{2}(\tau_k^1 - \tau_k^2) \beta_k^1, \\ \langle \nabla f(T_k^2), \hat{y}_k - T_k^2 \rangle &= \langle \nabla f(T_k^2), y_k^2 + \frac{1}{2}(y_k^1 - y_k^2) - T_k^2 \rangle \\ &\stackrel{(3.4)}{\geq} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} \|\nabla f(T_k^2)\|_*^{\frac{p+1}{p}} + \frac{1}{2}(\tau_k^1 - \tau_k^2) \beta_k^2. \end{aligned}$$

Hence,

$$\begin{aligned}\varphi_k(x_k) &\geq f(x_{k+1}) + \left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \frac{1}{2}\alpha_k(\tau_k^1 - \tau_k^2)\beta_k^1 + \frac{1}{2}(1-\alpha_k)(\tau_k^1 - \tau_k^2)\beta_k^2 \\ &= f(x_{k+1}) + \left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \alpha_k(\tau_k^1 - \tau_k^2)\beta_k^1 \stackrel{(4_b)}{\geq} f(x_{k+1}) + \frac{1}{2}\left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}}.\end{aligned}$$

Thus, inequality (3.8) is proved.  $\square$

COROLLARY 3.4. *For any  $k \geq 0$ , we have*

$$(3.9) \quad \sum_{i=k}^{\infty} g_i^{\frac{p+1}{p}} \leq 2 \left[\frac{H}{1-\beta}\right]^{\frac{1}{p}} (f(x_k) - f^*).$$

Denote  $D_0 = \max_{x \in \mathbb{E}} \{\|x - x_0\| : f(x) \leq f(x_0)\}$  and  $D_* = \max_{x \in \mathbb{E}} \{\|x - x^*\| : f(x) \leq f(x_0)\}$ . Clearly,

$$(3.10) \quad D_0 \leq 2D_*.$$

In algorithm (3.6), by the way of constructing the estimating functions  $\{\psi_k(\cdot)\}_{k \geq 0}$ , we have

$$(3.11) \quad \psi_k(x) \leq A_k f(x) + \frac{1}{2}\|x - x_0\|^2, \quad x \in \mathbb{E}, \quad k \geq 0.$$

LEMMA 3.5. *Let the sequence  $\{x_k\}_{k \geq 0}$  generated by method (3.6) be well defined. Then for all  $k \geq 0$  we have*

$$(3.12) \quad A_k f(x_k) + B_k \leq \psi_k^* = \min_{x \in \mathbb{E}} \psi_k(x),$$

where  $B_k = \frac{1}{4} \left[\frac{1-\beta}{H}\right]^{\frac{1}{p}} \sum_{i=0}^{k-1} A_{i+1} g_i^{\frac{p+1}{p}}$ . Moreover,  $\|v_k - x_k\| \leq D_0$  and

$$(3.13) \quad \|x_k - x^*\| \leq D_*, \quad \|v_k - x^*\| \leq D_*.$$

*Proof.* From inequality (3.12), we have

$$(3.14) \quad A_k f(x_k) + B_k + \frac{1}{2}\|x - v_k\|^2 \leq \psi_k(x) \stackrel{(3.11)}{\leq} A_k f(x) + \frac{1}{2}\|x - x_0\|^2, \quad x \in \mathbb{E}.$$

Taking in this inequality  $x = x_k$ , we get  $\|v_k - x_k\| \leq \|x_0 - x_k\|$ . On the other hand, the choice  $x = x_0$  ensures  $f(x_k) \leq f(x_0)$ . Hence,  $\|x_0 - x_k\| \leq D_0$ . Finally, the first inequality in (3.13) is valid since  $f(x_k) \leq f(x_0)$ , and the second one follows from (3.14) with  $x = x^*$ .

It remains to prove inequality (3.12). Let us do this by induction. For  $k = 0$  we have  $A_0 = B_0 = \psi_0^* = 0$ . Hence, the relation (3.12) is valid. Assume it is valid for some  $k \geq 0$ . Then

$$\begin{aligned}\psi_{k+1}(x) &= \psi_k(x) + a_{k+1}\varphi_k(x) \geq A_k f(x_k) + B_k + \frac{1}{2}\|x - v_k\|^2 + a_{k+1}\varphi_k(x) \\ &= A_k f(x_k) + B_k + \frac{1}{2}\|x - v_k\|^2 + a_{k+1}\varphi_k(x_k) + a_{k+1}\langle \nabla \varphi_k, x - x_k \rangle.\end{aligned}$$

Note that at the steps 2 and 4 of method (3.6), we have  $\langle \nabla \varphi_k, u_k \rangle \geq 0$ . Therefore,

$$\begin{aligned} & A_k f(x_k) + a_{k+1} \varphi_k(x_k) + a_{k+1} \langle \nabla \varphi_k, x - x_k \rangle \\ \stackrel{(3.8)}{\geq} & A_{k+1} f(x_{k+1}) + \frac{1}{2} A_{k+1} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1} \langle \nabla \varphi_k, x - x_k \rangle \\ \geq & A_{k+1} f(x_{k+1}) + \frac{1}{2} A_{k+1} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1} \langle \nabla \varphi_k, x - v_k \rangle. \end{aligned}$$

At the step 3 of method (3.6), we have  $\langle \nabla \varphi_k, u_k \rangle \leq 0$ . Hence,

$$\begin{aligned} & A_k f(x_k) + a_{k+1} \varphi_k(x_k) + a_{k+1} \langle \nabla \varphi_k, x - x_k \rangle \\ \geq & A_{k+1} f(x_{k+1}) + \langle \nabla \varphi_k, A_{k+1}(x_k - x_{k+1}) + a_{k+1}(x - x_k) \rangle \\ = & A_{k+1} f(x_{k+1}) + \langle \nabla \varphi_k, A_{k+1}(v_k - x_{k+1}) - A_k(v_k - x_k) + a_{k+1}(x - v_k) \rangle \\ \stackrel{(3.4)}{\geq} & A_{k+1} f(x_{k+1}) + A_{k+1} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1} \langle \nabla \varphi_k, x - v_k \rangle. \end{aligned}$$

Thus, in all cases,

$$\begin{aligned} & \psi_{k+1}(x) \\ \geq & A_{k+1} f(x_{k+1}) + B_k + \frac{1}{2} \|x - v_k\|^2 + \frac{1}{2} A_{k+1} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} + a_{k+1} \langle \nabla \varphi_k, x - v_k \rangle \\ \geq & A_{k+1} f(x_{k+1}) + B_k + \frac{1}{2} A_{k+1} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \frac{1}{2} a_{k+1}^2 g_k^2 \\ = & A_{k+1} f(x_{k+1}) + B_{k+1} + \frac{1}{4} A_{k+1} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{p+1}{p}} - \frac{1}{2} a_{k+1}^2 g_k^2 \\ = & A_{k+1} f(x_{k+1}) + B_{k+1}. \end{aligned}$$

□

COROLLARY 3.6. Let  $R_0 = \|x_0 - x^*\|$ . Then

$$(3.15) \quad \sum_{i=0}^{\infty} A_{i+1} g_i^{\frac{p+1}{p}} \leq 2 \left[ \frac{H}{1-\beta} \right]^{\frac{1}{p}} R_0^2.$$

*Proof.* Indeed, in taking in inequality (3.14)  $x = x^*$ , we have

$$A_k f(x_k) + B_k + \frac{1}{2} \|x^* - v_k\|^2 \leq A_k f^* + \frac{1}{2} R_0^2.$$

It remains to use the definition of  $B_k$ .

□

Now we can prove the main statement of this section.

THEOREM 3.7. Let the sequence  $\{x_k\}_{k \geq 0}$  generated by method (3.6) be well defined. Then for any  $k \geq 1$  we have

$$(3.16) \quad f(x_k) - f^* \leq \frac{4^p H R_0^{p+1}}{1-\beta} \left[ 1 + \frac{2(k-1)}{p+1} \right]^{-\frac{3p+1}{2}}.$$

*Proof.* In view of inequality (3.12), we have  $f(x_k) - f^* \leq \frac{R_0^2}{2A_k}$ . Hence, we only need to estimate from below the rate of growth of coefficients  $\{A_k\}_{k \geq 0}$ .

Using the same arguments as in the proof of Theorem 2.1, we get

$$A_{k+1}^{1/2} - A_k^{1/2} \geq \frac{a_{k+1}}{2A_{k+1}^{1/2}} = \sqrt{\frac{1}{8} \left[ \frac{1-\beta}{H} \right]^{\frac{1}{p}} g_k^{\frac{1-p}{p}}} \stackrel{\text{def}}{=} c_1 g_k^{\frac{1-p}{2p}}.$$

Denoting  $\xi_k = g_k^{\frac{p-1}{2p}}$ , we have  $\frac{1}{c_1} A_{t+1}^{1/2} \geq \sum_{k=0}^t \frac{1}{\xi_k}$  for any  $t \geq 0$ . On the other hand,

$$B_0 \stackrel{\text{def}}{=} 2 \left[ \frac{H}{1-\beta} \right]^{\frac{1}{p}} R_0^2 \stackrel{(3.15)}{\geq} \sum_{k=0}^t A_{k+1} g_k^{\frac{p-1}{p}} = \sum_{k=0}^t A_{k+1} \xi_k^{\frac{2(p+1)}{p-1}}.$$

Therefore, in view of Lemma 6.1, for  $\gamma = \frac{2(p+1)}{p-1}$  we have

$$\frac{1}{c_1} A_{t+1}^{1/2} \geq \frac{1}{B_0^{1/\gamma}} \left( \sum_{k=0}^t A_{k+1}^{\frac{1}{1+\gamma}} \right)^{\frac{1+\gamma}{\gamma}} = \frac{1}{B_0^{1/\gamma}} \left( \sum_{k=1}^{t+1} A_k^{\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{2(p+1)}}.$$

Thus, we get the following inequality:

$$(3.17) \quad A_t \geq \theta \left( \sum_{k=1}^t A_k^{\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{p+1}}, \quad t \geq 1,$$

where

$$\theta = \frac{c_1^2}{B_0^{2/\gamma}} = \left( \frac{1}{2} \right)^{\frac{4p+2}{p+1}} \left[ \frac{1-\beta}{H} \right]^{\frac{2}{p+1}} R_0^{-\frac{2(p-1)}{p+1}}.$$

Now we can use Lemma 6.2 with  $\alpha = \frac{p-1}{3p+1}$  and  $\beta = \frac{3p+1}{p+1}$ . Then  $1 - \alpha\beta = \frac{2}{p+1}$ , and by inequality (6.4), we get

$$(3.18) \quad A_k \geq \left( \frac{1}{2} \right)^{2p+1} \left[ \frac{1-\beta}{H} \right] R_0^{-(p-1)} \left[ \frac{p-1}{p+1} + \frac{2}{p+1} k \right]^{\frac{3p+1}{2}}. \quad \square$$

To conclude this section, let us look at the rate of convergence of method (3.6) in terms of the norm of the gradient. Note that the values  $\{g_k\}_{k \geq 0}$  generated by method (3.6) are the upper bounds for the norms of the objective function at certain points. Indeed, for step 2 and step 3 in method (3.6) we have  $g_k = \|\nabla f(x_{k+1})\|_*$ , and for step 4, it holds

$$\begin{aligned} g_k &= \left[ \alpha_k \|\nabla f(T_k^1)\|_*^{\frac{p+1}{p}} + (1 - \alpha_k) \|\nabla f(T_k^2)\|_*^{\frac{p+1}{p}} \right]^{\frac{p}{p+1}} \\ &\geq \min \left\{ \|\nabla f(T_k^1)\|_*, \|\nabla f(T_k^2)\|_* \right\}. \end{aligned}$$

THEOREM 3.8. *Let the sequence  $\{x_k\}_{k \geq 0}$  be generated by method (3.6). Then for any  $t = 2k$  with  $k \geq 1$  we have*

$$(3.19) \quad g_t^* \stackrel{\text{def}}{=} \min_{0 \leq i \leq t} g_i \leq 4^p(p+1)^{\frac{p(3p+1)}{2(p+1)}} \cdot \frac{HR_0^p}{1-\beta} \cdot \left[\frac{1}{t}\right]^{\frac{3}{2}p}.$$

*Proof.* Indeed, in view of inequality (3.16), we have

$$f(x_{k+1}) - f^* \leq \frac{4^p HR_0^{p+1}}{1-\beta} \left[1 + \frac{2k}{p+1}\right]^{-\frac{3p+1}{2}} \leq \frac{4^p HR_0^{p+1}}{1-\beta} \left[\frac{2k}{p+1}\right]^{-\frac{3p+1}{2}}.$$

On the other hand,

$$k(g_t^*)^{\frac{p+1}{p}} \leq \sum_{i=k+1}^t g_i^{\frac{p+1}{p}} \stackrel{(3.9)}{\leq} 2 \left[\frac{H}{1-\beta}\right]^{\frac{1}{p}} (f(x_{k+1}) - f^*).$$

Combining these two inequalities, we get (3.19).  $\square$

It is interesting that the rate of convergence (3.19) is the maximal rate compatible with the rate of growth (3.18) of coefficients  $\{A_k\}_{k \geq 0}$ . Indeed, if  $A_k = O(k^{(3p+1)/2})$ , then

$$a_{k+1} = O(k^{(3p-1)/2}), \quad \frac{a_{k+1}^2}{A_{k+1}} = O(k^{3(p-1)/2}).$$

On the other hand, by (3.19) we can guarantee that

$$g_k^{\frac{1-p}{p}} \approx O(k^{3(p-1)/2}),$$

which is compatible with the rule of step 5 in (3.6). Hence, a higher rate of decrease of the gradients should result in a higher rate of convergence of method (3.6).

**4. Complexity of auxiliary search.** In this section, we analyze a simple bisection strategy for implementing step 4 of method (3.6). Recall that at the beginning of this auxiliary process we have

$$\begin{aligned} \tau_k^1 &= 0, \quad \beta_k^1 = \langle \nabla f(x_k^0), u_k \rangle \leq 0, \\ \tau_k^2 &= 1, \quad \beta_k^2 = \langle \nabla f(x_k^1), u_k \rangle \geq 0. \end{aligned}$$

Since for  $p = 1$ , no auxiliary search is needed (see Remark 1); in this section we assume that  $p \geq 2$ . Thus, in method (3.6) we have  $\beta \leq \frac{3}{3p+2} \leq \frac{3}{8}$ , which allows us to use the second statement of Lemma 3.2.

The following bisection strategy has a single parameter: tolerance  $\beta \in [0, \frac{3}{3p+2}]$ .

**Initialization.** Set  $\tau_{k,0}^1 = 0$ ,  $\beta_{k,0}^1 = \beta_k^1$ ,  $T_{k,0}^1 = x_k^0$ ,  $\tau_{k,0}^2 = 1$ ,  $\beta_{k,0}^2 = \beta_k^2$ , and  $T_{k,0}^2 = x_k^1$ . Define  $\alpha_{k,0} = \frac{\beta_{k,0}^2}{\beta_{k,0}^2 - \beta_{k,0}^1} \in [0, 1]$ , and compute

$$g_{k,0} = \left[ \alpha_{k,0} \|\nabla f(x_k^0)\|_*^{\frac{p+1}{p}} + (1 - \alpha_{k,0}) \|\nabla f(x_k^1)\|_*^{\frac{p+1}{p}} \right]^{\frac{p}{p+1}}.$$

**Iteration**  $i \geq 0$ . **If**  $\alpha_{k,i}(\tau_{k,i}^1 - \tau_{k,i}^2)\beta_{k,i}^1 \leq \frac{1}{2}[\frac{1-\beta}{H}]^{\frac{1}{p}} g_{k,i}^{\frac{p+1}{p}}$ , **then** STOP.

**Else,**

(a) Compute  $\tau_{k,i}^+ = \frac{1}{2}(\tau_{k,i}^1 + \tau_{k,i}^2)$  and  $T_{k,i}^+ \in \mathcal{A}_H^p(x_k + \tau_{k,i}^+ u_k, \beta)$ .

(4.1) (b) Compute  $\beta_{k,i}^+ = \langle \nabla f(T_{k,i}^+), u_k \rangle$ . If  $\beta_{k,i}^+ \leq 0$ ,

then move the left bound  $\begin{cases} \tau_{k,i+1}^1 = \tau_{k,i}^+, \beta_{k,i+1}^1 = \beta_{k,i}^+, T_{k,i+1}^1 = T_{k,i}^+, \\ \tau_{k,i+1}^2 = \tau_{k,i}^2, \beta_{k,i+1}^2 = \beta_{k,i}^2, T_{k,i+1}^2 = T_{k,i}^2, \end{cases}$

else move the right bound  $\begin{cases} \tau_{k,i+1}^1 = \tau_{k,i}^1, \beta_{k,i+1}^1 = \beta_{k,i}^1, T_{k,i+1}^1 = T_{k,i}^1, \\ \tau_{k,i+1}^2 = \tau_{k,i}^+, \beta_{k,i+1}^2 = \beta_{k,i}^+, T_{k,i+1}^2 = T_{k,i}^+. \end{cases}$

(c). Define  $\alpha_{k,i+1} = \frac{\beta_{k,i+1}^2}{\beta_{k,i+1}^2 - \beta_{k,i+1}^1} \in [0, 1]$ , and compute

$$g_{k,i+1} = \left[ \alpha_{k,i+1} \|\nabla f(T_{k,i+1}^1)\|_*^{\frac{p+1}{p}} + (1 - \alpha_{k,i+1}) \|\nabla f(T_{k,i+1}^2)\|_*^{\frac{p+1}{p}} \right]^{\frac{p}{p+1}}.$$

In view of the updating rules in (4.1), for all  $i \geq 0$ , we have the following relations:

$$(4.2) \quad \tau_{k,i}^2 - \tau_{k,i}^1 = 2^{-i}, \quad \beta_{k,i}^1 \leq 0 \leq \beta_{k,i}^2.$$

Let us find an upper bound for the length of this process.

Note that

$$\begin{aligned} 0 &\leq -\alpha_{k,i} \beta_{k,i}^1 = \frac{-\beta_{k,i}^1 \beta_{k,i}^2}{\beta_{k,i}^2 - \beta_{k,i}^1} \leq \min\{-\beta_{k,i}^1, \beta_{k,i}^2\} \\ &\leq \|u_k\| \cdot \min\{\|\nabla f(T_{k,i}^1)\|_*, \|\nabla f(T_{k,i}^2)\|_*\}. \end{aligned}$$

Hence, in view of Lemma 3.5, the left-hand side of the stopping criterion in (4.1) does not exceed

$$D_0 \min\{\|\nabla f(T_{k,i}^1)\|_*, \|\nabla f(T_{k,i}^2)\|_*\} \cdot 2^{-i}.$$

On the other hand,  $g_{k,i} \geq \min\{\|\nabla f(T_{k,i}^1)\|_*, \|\nabla f(T_{k,i}^2)\|_*\}$ . Hence, the upper bound  $i_k$  for the length of the process (4.1) can be found from the following inequality:

$$(4.3) \quad \frac{HD_0^p}{1-\beta} \cdot 2^{p(1-i_k)} \leq \min\{\|\nabla f(T_{k,i}^1)\|_*, \|\nabla f(T_{k,i}^2)\|_*\}.$$



THEOREM 4.1. *If during the process (4.1), we have*

$$(4.4) \quad \min \left\{ f(T_{k,i}^1), f(T_{k,i}^2) \right\} \geq f^* + \epsilon, \quad i = 0, \dots, i_k,$$

where  $\epsilon > 0$  is the accuracy parameter, then

$$(4.5) \quad i_k \leq \left( 2 + \frac{1}{p} \log_2 \frac{5HD_*^{p+1}}{4(1-\beta)\epsilon} \right)_+,$$

where  $(\tau)_+ = \max\{\tau, 0\}$  for  $\tau \in \mathbb{R}$ .

*Proof.* In view of inequality (3.13), for any  $\tau \in [0, 1]$ , we have  $\|x_k + \tau u_k - x^*\| \leq D_*$ . Using now the second statement of Lemma 3.2, we get

$$\|T_{k,i}^1 - x^*\| \leq \frac{5}{4}D_*, \quad \|\cdot\|, \quad \|T_{k,i}^2 - x^*\| \leq \frac{5}{4}D_*.$$

Therefore,

$$\begin{aligned} \epsilon &\stackrel{(4.4)}{\leq} \min \left\{ f(T_{k,i}^1) - f^*, f(T_{k,i}^2) - f^* \right\} \\ &\leq \min \left\{ \langle \nabla f(T_{k,i}^1), T_{k,i}^1 - x^* \rangle, \langle \nabla f(T_{k,i}^2), T_{k,i}^2 - x^* \rangle \right\} \\ &\leq \frac{5}{4}D_* \min \left\{ \|\nabla f(T_{k,i}^1)\|_*, \|\nabla f(T_{k,i}^2)\|_* \right\}. \end{aligned}$$

This means that the stopping criterion in (4.1) will definitely be activated for  $i_k$  satisfying the inequality  $\frac{HD_*^p}{1-\beta} \cdot 2^{p(2-i_k)} \leq \frac{4\epsilon}{5D_*}$ , which gives us (4.5).  $\square$

**5. Practical implementations.** The results of Theorem 3.7 and Theorem 4.1 give us the upper bound

$$(5.1) \quad O \left( \left[ \frac{HD_*^{p+1}}{\epsilon} \right]^{\frac{2}{3p+1}} \ln \frac{HD_*^{p+1}}{\epsilon} \right), \quad p \geq 2,$$

for the number of computations of the inclusion

$$(5.2) \quad T \in \mathcal{A}_H^p(x, \beta), \quad x \in \mathbb{E},$$

which is sufficient for getting by method (3.6) with (4.1) an  $\epsilon$ -solution of (2.1) in terms of the function value. Note that up to this moment we did not assume anything about the objective function except its convexity. However, we need these assumptions for transforming the bound (5.1) into a bound for the number of calls of the oracle. We also need to specify a particular technique for solving the inclusion (5.2).

Thus, we naturally come to the framework of BLUM, which was introduced and discussed in section 4 of [22]. In this framework, at the upper level, we choose the order  $p \geq 1$  of the method (3.6), ensuring the complexity bound (5.1). On the lower level, we choose a numerical method for solving (5.2). In this way, we can construct many interesting combinations. In the remaining part of this section, we consider only two of them.

**5.1. Accelerated proximal-point methods with tensor steps.** This possibility was already considered in section 3 of [22] for the slower proximal-point methods with the convergence rate  $O(k^{-(p+1)})$ . Here we just apply these results to the faster schemes (3.6) with (4.1).

Denote  $M_{p+1}(f) = \sup_{x \in \mathbb{E}} \|D^{p+1}f(x)\|$ . For the objective function  $f(\cdot)$ , we assume that

$$(5.3) \quad M_{p+1}(f) < +\infty.$$

Define the standard *Taylor approximation* of  $f(\cdot)$  at point  $x \in \mathbb{E}$ :

$$\Omega_{x,p}(y) = f(x) + \sum_{k=1}^p \frac{1}{k!} D^k f(x) [y - x]^k.$$

We also need the *augmented Taylor polynomial*:

$$\hat{\Omega}_{x,p,M}(y) = \Omega_{x,p}(y) + \frac{M}{(p+1)!} \|y - x\|^{p+1}.$$

If  $M \geq M_{p+1}(f)$ , then this function provides us with a uniform upper bound for the objective. Moreover, if  $M \geq pM_{p+1}(f)$ , then function  $\hat{\Omega}_{x,p,M}(\cdot)$  is convex (see [19]). Therefore, we are able to compute the tensor step

$$(5.4) \quad T_{p,M}(x) = \arg \min_{y \in \mathbb{E}} \hat{\Omega}_{x,p,M}(y).$$

Let us allow some inexactness in the computation of this point. Namely, we assume that it is possible to compute a point  $T$  satisfying the following condition:

$$(5.5) \quad \|\nabla \hat{\Omega}_{x,p,M}(T)\|_* \leq \frac{\gamma}{1+\gamma} \|\nabla \Omega_{x,p}(T)\|_*,$$

where  $\gamma \in [0, \frac{\beta}{1+\beta})$  is the tolerance parameter.

Let us define now

$$(5.6) \quad M = \frac{1+\beta}{\beta(1-\gamma)-\gamma} M_{p+1}(f), \quad H = \frac{1}{p!} M.$$

Then in view of Lemma 3 in [22], we have

$$\|\nabla f(T) + H \nabla d_{p+1}(T - x)\|_* \leq \beta \|\nabla f(T)\|_*.$$

In other words, these values of the parameters ensure the inclusion  $T \in \mathcal{A}_H^p(x, \beta)$ .

Thus, a solution to the inclusion (5.2) can be computed in one step of the inexact tensor method provided that we are able to compute its appropriate approximation (5.5). Since the minimization problem (5.4) does not require new calls of the oracle of the objective function  $f(\cdot)$ , the bound (5.1) is valid for the number of calls of the  $p$ th-order oracle. Note that this bound coincides with the complexity bound for the recently developed tensor methods [4, 8, 11] (see also [7] for historical remarks and other contributions). However, the auxiliary problem to be solved at each iteration of our method is convex, which should give better practical results.

Let us mention two cases when the solution of (5.4) is relatively easy.

- $p = 2$ . Then (5.4) coincides with the auxiliary problem in cubic regularization of the Newton method [24, 17]. It can be solved very efficiently by an appropriate matrix factorization technique. In this case, we take in (5.5)  $\gamma = 0$ . The number of calls of the second-order oracle in our new method is bounded by (5.1) with  $p = 2$ , that is,

$$(5.7) \quad O\left(\left[\frac{M_3(f)D_*^3}{\epsilon}\right]^{\frac{2}{7}} \ln \frac{M_3(f)D_*^3}{\epsilon}\right).$$

- $p = 3$ . Then (5.4) can be solved by an auxiliary gradient method based on Bregman distances (see [21] for details). In this case, the computation of an approximate solution (5.2) needs  $O(\ln \frac{1}{\epsilon})$  computations of the gradient of the Taylor polynomial of degree tree. The number of calls of the third-order oracle in the upper-level method is bounded by (5.1) with  $p = 3$ , that is,

$$(5.8) \quad O\left(\left[\frac{M_4(f)D_*^4}{\epsilon}\right]^{\frac{1}{5}} \ln \frac{M_4(f)D_*^4}{\epsilon}\right).$$

As it was demonstrated in [21], by an appropriate approximation of the third-directional derivative by the differences of gradients, we can transform this type of methods into the second-order schemes. In this paper, we do not consider such a possibility since the finite differences are often numerically unstable. In any case, it seems that section 5.2 offers a better alternative.

To the best of our knowledge, the efficient technique for solving (5.4) with the order  $p \geq 4$  is not found yet.

**5.2. Second-order implementation of the third-order method.** Let us choose for the upper-level process  $p = 3$ . This means that the point  $T \in \mathcal{A}_H^3(\bar{x}, \beta)$  is computed as an approximate solution to the problem

$$(5.9) \quad \min_{x \in \mathbb{E}} \left\{ \varphi(x) \stackrel{\text{def}}{=} f(x) + Hd_4(x - \bar{x}) \right\}.$$

We are going to solve this problem by a *second-order method*.

Let us assume that  $M_4(f) < +\infty$ . Define

$$(5.10) \quad H = 3M_4(f), \quad \mu = \frac{1}{2}, \quad L = \frac{3}{2}, \quad \varkappa = \frac{1}{3}.$$

We also need the scaling function  $\rho(x) = \frac{1}{2} \langle \nabla^2 f(\bar{x})(x - \bar{x}), x - \bar{x} \rangle + Hd_4(x - \bar{x})$ . Then, in view of Theorem 3 in [22], function  $\varphi(\cdot)$  satisfies the following *relative nondegeneracy condition* [3, 12]:

$$(5.11) \quad \mu \nabla^2 \rho(x) \preceq \nabla \varphi(x) \preceq L \nabla^2 \rho(x), \quad x \in \mathbb{E}.$$

Hence, defining the Bregman distance  $\beta(x, y) = \rho(y) - \rho(x) - \langle \nabla \rho(x), y - x \rangle \geq 0$ , we get the following inequality:

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + L\beta(x, y), \quad x, y \in \mathbb{E}.$$

Consequently, we come to the following gradient method: Set  $x_0 = \bar{x}$ , and for  $k \geq 0$  iterate

$$(5.12) \quad x_{k+1} = \arg \min_{x \in \mathbb{E}} \left\{ \langle \nabla \varphi(x_k), x - x_k \rangle + L\beta(x_k, x) \right\}.$$

Note that in terms of function  $f(\cdot)$ , this is mostly a *first-order method*. The Hessian of the objective is computed only once in the beginning of the process in order to define the scaling function  $\rho(\cdot)$ .

In accordance to [12, 20], (5.12) has a linear rate of convergence:

$$(5.13) \quad \beta(x_{k+1}, \bar{x}^*) \leq (1 - \varkappa)\beta(x_k, \bar{x}_*), \quad k \geq 0,$$

where  $\bar{x}^*$  is the exact solution of (5.9). Thus,

$$(5.14) \quad \begin{aligned} \frac{H}{16} \|x_k - \bar{x}^*\|^4 &\stackrel{(1.4)}{\leq} \beta(x_k, \bar{x}^*) \leq (1 - \kappa)^k \beta(x_0, \bar{x}^*), \quad k \geq 0, \\ \beta(x_0, \bar{x}_*) &= \frac{1}{2} \langle \nabla^2 f(\bar{x})(\bar{x} - \bar{x}_*), \bar{x} - \bar{x}_* \rangle + \frac{H}{4} \|\bar{x} - \bar{x}_*\|^4. \end{aligned}$$

Using this fact, we need to find an upper bound for the number of iterations of this scheme which is sufficient for satisfying the stopping criterion

$$(5.15) \quad \|\nabla \varphi(x_k)\|_* \leq \beta \|\nabla f(x_k)\|_*.$$

For this, we need several auxiliary statements. Our main concern is the upper bound for the Hessians of functions  $f(\cdot)$  and  $\varphi(\cdot)$ . We start from a simple statement on second derivatives.

LEMMA 5.1. *For any  $x, y \in \mathbb{E}$ , we have*

$$(5.16) \quad \nabla^2 f(y) \preceq 2\nabla^2 f(x) + M_4(f) \|x - y\|^2 B.$$

*Proof.* Indeed, for any  $x, h \in \mathbb{E}$ , we have

$$\begin{aligned} \nabla^2 f(x+h) &\preceq \nabla^2 f(x) + D^3 f(x)[h] + \frac{1}{2} M_4(f) \|h\|^2 B, \\ 0 &\preceq \nabla^2 f(x-h) \preceq \nabla^2 f(x) - D^3 f(x)[h] + \frac{1}{2} M_4(f) \|h\|^2 B, \end{aligned}$$

and (5.16) follows.  $\square$

In our analysis, we need the following characteristic:

$$(5.17) \quad M_2 = \max_{x \in \mathbb{E}} \left\{ \|\nabla^2 f(x)\| : x \in B(x^*, D_*) \right\},$$

where  $B(x, r) = \{y \in \mathbb{E} : \|y - x\| \leq r\}$ . It can be used for bounding the norms of the gradients:

$$(5.18) \quad \|\nabla f(x)\| \leq M_2 D_*, \quad x \in B(x^*, D_*).$$

It also gives us a bound for the size of the level set of function  $\varphi(\cdot)$ .

LEMMA 5.2. *Let  $\bar{x} \in B(x^*, D_*)$  and  $x \in \mathcal{L}_\varphi = \{x \in \mathbb{E} : \varphi(x) \leq \varphi(\bar{x})\}$ . Then*

$$(5.19) \quad \|x - \bar{x}\| \leq \hat{R} \stackrel{\text{def}}{=} \left[ \frac{4M_2 D_*}{H} \right]^{1/3}.$$

*Proof.* Indeed, in this case,

$$\begin{aligned} \frac{H}{4} \|x - \bar{x}\|^4 &\leq f(\bar{x}) - f(x) \leq \langle \nabla f(\bar{x}), \bar{x} - x \rangle \leq \|\nabla f(\bar{x})\|_* \cdot \|\bar{x} - x\| \\ &\stackrel{(5.18)}{\leq} M_2 D_* \|\bar{x} - x\|. \end{aligned} \quad \square$$

It is also useful for bounding the Hessians of functions  $f(\cdot)$  and  $\varphi(\cdot)$ .

LEMMA 5.3. *Let  $\bar{x} \in B(x^*, D_*)$ . Then for any  $x \in \mathcal{L}_\varphi$  we have*

$$(5.20) \quad \begin{aligned} \|\nabla^2 f(x)\| &\leq M_f \stackrel{\text{def}}{=} 2M_2 + M_4(f) \hat{R}^2, \\ \|\nabla^2 \varphi(x)\| &\leq M_\varphi \stackrel{\text{def}}{=} L_f + 3H \hat{R}^2. \end{aligned}$$

*Proof.* Indeed, in view of the matrix inequality (5.16), we have

$$\nabla^2 f(x) \preceq 2\nabla^2 f(\bar{x}) + M_4(f)\|x - \bar{x}\|^2 B \stackrel{(5.19)}{\leq} (2M_2 + M_4(f)\hat{R}^2)B.$$

It remains to note that from (1.3)  $\nabla^2 d_4(x - \bar{x}) \preceq 3\|x - \bar{x}\|^2 B \preceq 3\hat{R}^2 B$ .  $\square$

The size of the constant  $M_2$  can be controlled in different ways. First of all, we can assume that  $M_2(f) < +\infty$ . Then we can take  $M_2 = M_2(f)$ . Another possibility is to use the relation (5.16). Then, for any  $x \in B(x^*, D_*)$ , we have

$$\nabla^2 f(x) \preceq 2\nabla^2 f(x^*) + M_4(f)D_*^2 B.$$

However, in any case this characteristic has a limited influence on the complexity bounds since it stays inside the logarithms.

Finally, we need one more fact on the properties of the high-order proximal-point operator (2.2) with  $p \geq 1$ .

LEMMA 5.4. *Let the second-order derivative of function  $f(\cdot)$  be bounded on some convex set  $\mathcal{L}$  by the constant  $M_f$ . If both points  $\bar{x} \in \mathcal{L}_\varphi$  and  $T = \text{prox}_{f/H}^p(\bar{x})$  belong to  $\mathcal{L}$ , then*

$$(5.21) \quad \|\nabla f(T)\|_* \geq \frac{H\|\nabla f(\bar{x})\|_*^p}{M_f^p + pH\|\nabla f(\bar{x})\|_*^{p-1}}.$$

*Proof.* Denote  $r = \|T - \bar{x}\|$  and  $\bar{g} = \|\nabla f(\bar{x})\|_*$ . Then

$$\|\nabla f(T)\|_* \geq \bar{g} - M_f r.$$

On the other hand, the first-order optimality condition for (2.2) reads as follows:

$$\nabla f(T) + H\nabla d_{p+1}(T - \bar{x}) = 0.$$

Hence, from (1.3)  $\|\nabla f(T)\|_* = Hr^p$ , and we get the following lower bound:

$$\|\nabla f(T)\|_* \geq \max\{\bar{g} - M_f r, Hr^p\}.$$

Since univariate function  $r^p$  is convex in  $r \geq 0$ , the minimum of the right-hand side of this inequality in  $r$  is above the following minimum:

$$g_* \stackrel{\text{def}}{=} \min_{r \geq 0} \max\{\bar{g} - M_f r, H[\hat{r}^p + p\hat{r}^{p-1}(r - \hat{r})]\} = \frac{H\hat{r}^{p-1}(p\bar{g} - (p-1)M_f \hat{r})}{M_f + pH\hat{r}^{p-1}},$$

where  $\hat{r}$  is an arbitrary positive number. Taking  $\hat{r} = \frac{\bar{g}}{M_f}$ , we get  $g_* = \frac{H\bar{g}^p}{M_f^p + pH\bar{g}^{p-1}}$ .  $\square$

Now we can prove our main result.

THEOREM 5.5. *Let  $\bar{x} \in B(x^*, D_*)$  and  $f(\bar{x}) - f^* \geq \epsilon > 0$ . Then the number of steps of method (5.12) sufficient for activating the stopping criterion (5.15) cannot exceed the value of the order*

$$(5.22) \quad O\left(\ln \frac{1}{\epsilon} + \ln D_* + \ln M_4(f) + \ln M_2\right).$$

*Proof.* Since  $\bar{x} \in B(x^*, D_*)$ , from the condition of the theorem we have

$$\epsilon \leq f(\bar{x}) - f^* \leq \langle \nabla f(\bar{x}), \bar{x} - x^* \rangle \leq \|\nabla f(\bar{x})\|_* D_*.$$

Therefore, in view of Lemma 5.4, we have

$$\|\nabla f(\bar{x}_*)\| \stackrel{(5.18)}{\geq} \frac{H\|\nabla f(\bar{x})\|_*^3}{M_f^3 + 3H(M_2 D_*)^2} \geq \frac{H\epsilon^3}{D_*^3(M_f^3 + 3H(M_2 D_*)^2)}.$$

Denote by  $\delta_k = (1 - \varkappa)^k$  the reduction factor in the convergence rate (5.14). Note that

$$\|x_k - \bar{x}^*\| \stackrel{(5.14)}{\leq} \delta_k^{1/4} \left[ \frac{8}{H} M_2 \hat{R}^2 + 4\hat{R}^4 \right]^{\frac{1}{4}}.$$

Since  $x_k \in \mathcal{L}_\varphi$ , we conclude that

$$\|\nabla f(x_k)\|_* \geq \|\nabla f(\bar{x}^*)\|_* - M_f \|x_k - \bar{x}^*\|.$$

On the other hand,  $\|\nabla \varphi(x_k)\|_* \leq M_\varphi \|x_k - \bar{x}^*\|$ . Therefore, the stopping criterion (5.15) is satisfied if

$$(M_\varphi + \beta M_f) \|x_k - \bar{x}^*\| \leq \beta \|\nabla f(\bar{x}^*)\|_*.$$

A sufficient condition for this is as follows:

$$\delta_k^{1/4} \left[ \frac{8}{H} M_2 \hat{R}^2 + 4\hat{R}^4 \right]^{\frac{1}{4}} (M_\varphi + \beta M_f) \leq \frac{\beta H \epsilon^3}{D_*^3(M_f^3 + 3H(M_2 D_*)^2)}.$$

Thus, the upper bound for the number of iterations in method (5.12) is of the order (5.22).  $\square$

Note that the main assumption of Theorem 5.5 is the inclusion  $\bar{x} \in B(x^*, D_*)$ . In view of (3.13), it is valid for all points  $\bar{x} = x_k + \tau(v_k - x_k)$ ,  $\tau \in [0, 1]$  used in the auxiliary search procedure (4.1).

Thus, we have proved that the third-order accelerated proximal-point method (3.6) and (4.1) can be implemented by the second-order method (5.12). This combination has the following performance characteristics:

- The rate of convergence of the upper-level method (3.6) (Theorem 3.7),  $O(k^{-5})$ ;
- The number of computations of Hessians per iteration of method (3.6) (Theorem 4.1),

$$N_H = O(\ln \frac{1}{\epsilon} + \ln D_* + \ln M_4(f));$$

- The number of computations of gradients per iteration of this method (Theorem 5.5),

$$N_G = O\left(N_H \times (\ln \frac{1}{\epsilon} + \ln D_* + \ln M_4(f) + \ln M_2)\right).$$

Note that the total time for computing the gradients is normally dominated by the time spent for the Hessians. Indeed, usually the Hessians are  $(\dim \mathbb{E})$  times more expensive than the gradients. At the same time, very often we have

$$\dim \mathbb{E} > \ln \frac{1}{\epsilon} + \ln D_* + \ln M_4(f) + \ln M_2.$$

**6. Appendix.** In this section, we present two useful statements on positive numbers.

LEMMA 6.1. *Let the positive numbers  $\{\xi_i\}_{i=1}^N$  satisfy the following condition:*

$$(6.1) \quad \sum_{i=1}^N w_i \xi_i^\gamma \leq B,$$

where  $\gamma \geq 1$  and all weights  $w_i$  are positive. Then

$$(6.2) \quad \sum_{i=1}^N \frac{1}{\xi_i} \geq \frac{1}{B^{1/\gamma}} \left( \sum_{i=1}^N w_i^{\frac{1}{1+\gamma}} \right)^{\frac{1+\gamma}{\gamma}}.$$

*Proof.* Indeed, minimizing the left-hand side of inequality (6.2) with the constraint (6.1), we have the following first-order optimality condition:

$$\frac{\lambda}{\xi_i^2} = w_i \xi_i^{\gamma-1}, \quad 1 \leq i \leq N,$$

where  $\lambda > 0$  is the Lagrange multiplier for the active constraint. Thus,

$$\xi_i = \left( \frac{\lambda}{w_i} \right)^{\frac{1}{\gamma+1}},$$

and the multiplier  $\lambda$  can be found from the equation

$$\sum_{i=1}^n w_i \left( \frac{\lambda}{w_i} \right)^{\frac{\gamma}{\gamma+1}} = B.$$

This is  $\lambda = \left( \frac{B}{S} \right)^{\frac{1+\gamma}{\gamma}}$ , where  $S = \sum_{i=1}^N w_i^{\frac{1}{1+\gamma}}$ . Hence, we get (6.2).  $\square$

**LEMMA 6.2.** *Let the sequence of positive numbers  $\{\eta_k\}_{k \geq 0}$  satisfy the following inequality:*

$$(6.3) \quad \eta_k \geq \theta \cdot \left( \sum_{i=1}^k \eta_i^\alpha \right)^\beta, \quad k \geq 1,$$

where  $\theta > 0$ ,  $\alpha \geq 0$ ,  $\beta \geq 0$ , and  $\alpha\beta < 1$ . Then

$$(6.4) \quad \eta_k \geq \theta^{\frac{1}{1-\alpha\beta}} \left[ \alpha\beta + (1-\alpha\beta)k \right]^{\frac{\beta}{1-\alpha\beta}}, \quad k \geq 1.$$

*Proof.* Denote  $C_k = \left( \sum_{i=1}^k \eta_i^\alpha \right)^{1-\alpha\beta}$ . Then, in view of (6.3), we have

$$C_{k+1}^{\frac{1}{1-\alpha\beta}} - C_k^{\frac{1}{1-\alpha\beta}} = \eta_{k+1}^\alpha \stackrel{(6.3)}{\geq} \theta^\alpha \cdot C_{k+1}^{\frac{\alpha\beta}{1-\alpha\beta}}.$$

Note that for  $k = 1$  we have from (6.3)  $C_1 = \eta_1^{\alpha(1-\alpha\beta)} \geq \theta^\alpha$ . On the other hand, since  $\frac{1}{1-\alpha\beta} > 1$ , function  $\tau^{\frac{1}{1-\alpha\beta}}$  is convex for  $\tau \geq 0$ . Therefore,

$$\begin{aligned} \theta^\alpha \cdot C_{k+1}^{\frac{\alpha\beta}{1-\alpha\beta}} &\leq C_{k+1}^{\frac{1}{1-\alpha\beta}} - C_k^{\frac{1}{1-\alpha\beta}} \leq \frac{1}{1-\alpha\beta} C_{k+1}^{\frac{1}{1-\alpha\beta}-1} (C_{k+1} - C_k) \\ &= \frac{1}{1-\alpha\beta} C_{k+1}^{\frac{\alpha\beta}{1-\alpha\beta}} (C_{k+1} - C_k). \end{aligned}$$

Hence,  $C_{k+1} - C_k \geq (1-\alpha\beta)\theta^\alpha$ , and we get inequality (6.4).  $\square$

**Acknowledgments.** We would like to thank four anonymous referees for the careful reading and useful suggestions.

## REFERENCES

- [1] N. AGARWAL AND E. HAZAN, *Lower Bounds for Higher-Order Convex Optimization*, preprint, arXiv:1710.10329v1, 2017, <https://arxiv.org/abs/1710.10329v1>.
- [2] Y. ARJEVANI, O. SHAMIR, AND R. SHIFF, *Oracle Complexity of Second-Order Methods for Smooth Convex Optimization*, preprint, arXiv:1705.07260v2, 2017, <https://arxiv.org/abs/1705.07260>.
- [3] H. H. BAUSCHKE, J. BOLTE, AND M. TEBoulLE, *A descent lemma beyond Lipschitz gradient continuity: First order methods revisited and applications*, Math. Oper. Res., 42 (2016), pp. 330–348.
- [4] S. BUBECK, Q. JIANG, Y. T. LEE, Y. LI, AND A. SIDFORD, *Near-Optimal Methods for Highly Smooth Convex Optimization*, preprint, arXiv:1812.0826v2, 2019, <https://arxiv.org/abs/1812.0826v2>.
- [5] E. G. BIRGIN, J. L. GARDENGHI, J. M. MARTINEZ, S. A. SANTOS, AND PH. L. TOINT, *Worstcase evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*, Math. Program., 163 (2017), pp. 359–368.
- [6] Y. CARMON, J. C. DUCHI, O. HINDER, AND A. SIDFORD, *Accelerated methods for nonconvex optimization*, SIAM J. Optim., 28 (2018), pp. 1751–1772.
- [7] A. GASNIKOV, P. DVURECHENSKY, E. GORBUNOV, E. VORONTSOVA, D. SELIKHANOVYCH, C. URIBE, B. JIANG, H. WANG, S. ZHANG, S. BUBECK, Q. JIANG, Y. T. LEE, Y. LI, AND A. SIDFORD, *Near optimal methods for minimizing convex functions with Lipschitz  $p$ th derivatives*, in Proceedings of the 32nd Conference on Learning Theory, PMLR, Phoenix, AZ, 2019, pp. 1392–1393.
- [8] A. GASNIKOV, E. GORBUNOV, D. KOVALEV, A. MOHHAMED, AND E. CHERNOUSOVA, *The Global Rate of Convergence for Optimal Tensor Methods in Smooth Convex Optimization*, preprint, arXiv:1809.00382, 2018, <https://arxiv.org/abs/1809.00382>.
- [9] O. GÜLER, *New proximal point algorithms for convex minimization*, SIAM J. Control Optim., 14 (1992), pp. 877–898.
- [10] A. N. IUSEM, B. F. SVAITER, AND M. TEBoulLE, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.
- [11] B. JIANG, H. WANG, AND S. ZANG, *An Optimal High-Order Tensor Method for Convex Optimization*, preprint, arXiv:1812.06557, 2018, <https://arxiv.org/abs/1812.06557>.
- [12] H. LU, R. FREUND, AND YU. NESTEROV, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. Optim., 28 (2018), pp. 333–354.
- [13] B. MARTINET, *Perturbation des méthodes d'optimisation. Application*, ESIAM. Math. Model. Numer. Anal., 12 (1978), pp. 153–171.
- [14] J. J. MOREAU, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
- [15] Y. NESTEROV, *A method for unconstrained convex minimization problem with the rate of convergence  $O(\frac{1}{k^2})$* , Sov. Math. Dokl., 269 (1983), pp. 543–547.
- [16] Y. NESTEROV, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
- [17] Y. NESTEROV, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program., 112 (2008), pp. 159–181.
- [18] Y. NESTEROV, *Lectures on Convex Optimization*, 2nd ed., Springer, Cham, Switzerland, 2018.
- [19] Y. NESTEROV, *Implementable tensor methods in unconstrained convex optimization*, Math. Program., 186 (2021), pp. 157–183, doi:10.1007/s10107-019-01449-1.
- [20] Y. NESTEROV, *Inexact basic tensor methods for some classes of convex optimization problems*, Optim. Methods Softw., 34 (2020), doi:10.1080/10556788.2020.1854252.
- [21] Y. NESTEROV, *Superfast second-order methods for unconstrained convex optimization*, J. Optim. Theory Appl., 191 (2021), pp. 1–30.
- [22] Y. NESTEROV, *Inexact accelerated high-order proximal-point methods*, Math. Program., to appear.
- [23] Y. NESTEROV AND A. NEMIROVSKII, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*, SIAM, Philadelphia, PA, 1994.
- [24] Y. NESTEROV AND B. POLYAK, *Cubic regularization of Newton's method and its global performance*, Math. Program., 108 (2006), pp. 177–205.
- [25] R. T. ROCKAFELLAR, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
- [26] M. TEBoulLE, *Entropic proximal mapping with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.