



High-order methods beyond the classical complexity bounds: inexact high-order proximal-point methods

Masoud Ahookhosh¹ · Yurii Nesterov²

Received: 24 October 2021 / Accepted: 21 November 2023
© The Author(s) 2024, corrected publication 2024

Abstract

We introduce a *Bi-level OPTimization* (BiOPT) framework for minimizing the sum of two convex functions, where one of them is smooth enough. The BiOPT framework offers three levels of freedom: (i) choosing the order p of the proximal term; (ii) designing an inexact p th-order proximal-point method in the upper level; (iii) solving the auxiliary problem with a lower-level non-Euclidean method in the lower level. We here regularize the objective by a $(p + 1)$ th-order proximal term (for arbitrary integer $p \geq 1$) and then develop the generic inexact high-order proximal-point scheme and its acceleration using the standard estimating sequence technique at the upper level. This follows at the lower level with solving the corresponding p th-order proximal auxiliary problem inexactly either by one iteration of the p th-order tensor method or by a lower-order non-Euclidean composite gradient scheme. Ultimately, it is shown that applying the accelerated inexact p th-order proximal-point method at the upper level and handling the auxiliary problem by the non-Euclidean composite gradient scheme lead to a $2q$ -order method with the convergence rate $\mathcal{O}(k^{-(p+1)})$ (for $q = \lfloor p/2 \rfloor$ and the iteration counter k), which can result to a superfast method for some specific class of problems.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 788368). The first author was partially supported by the Research Foundation Flanders (FWO) research project G081222N and UA BOF DocPRO4 project with ID 46929.

✉ Masoud Ahookhosh
masoud.ahookhosh@uantwerp.be

Yurii Nesterov
Yurii.Nesterov@uclouvain.be

¹ Department of Mathematics, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium

² Center for Operations Research and Econometrics (CORE) and Department of Mathematical Engineering (INMA), Catholic University of Louvain (UCL), 34voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium

Keywords Convex composite optimization · High-order proximal-point operator · Bi-level optimization framework · Lower complexity bounds · Optimal methods · Superfast methods

Mathematics Subject Classification 90C25 · 90C06 · 49J52 · 65K05 · 65Y20

1 Introduction

Motivation Central to the entire discipline of convex optimization is the concept of complexity analysis for evaluating the efficiency of a wide spectrum of algorithms dealing with such problems; see [21, 26]. For example, under the Lipschitz continuity of the gradient of the objective function, the fastest convergence rate for first-order methods is of $\mathcal{O}(k^{-2})$ for the iteration counter k ; cf. [22, 24]. Likewise, if the objective is twice differentiable with Lipschitz continuous Hessian, the best complexity for second-order methods is of $\mathcal{O}(k^{-7/2})$; see [7]. In the recent years, there is an increasing interest to applying high-order methods for both convex and nonconvex problems; see, e.g., [1, 7, 10, 12, 17]. If the objective is p -times differentiable with Lipschitz continuous p th derivatives, then the fastest convergence rate for p th-order methods is of $\mathcal{O}(k^{-(3p+1)/2})$; cf. [7].

In general, for convex problems, the classical setting involves a one-to-one correspondence between the methods and problem classes. In other words, there exists and unimprovable complexity bound for a class of methods applied to a class of problems under specific assumptions. In fact, under the Lipschitz continuity of the p th derivatives, the p th-order method is called *optimal* if it attains the convergence rate $\mathcal{O}(k^{-(3p+1)/2})$, and if a method attains a faster convergence rate (under stronger assumptions than the optimal methods), we call it *superfast*. For example, first-order methods with the convergence rate $\mathcal{O}(k^{-2})$ and second-order methods with the convergence rate $\mathcal{O}(k^{-7/2})$ are optimal under the Lipschitz continuity of the first and the second derivatives, respectively. Recently, in [30], a *superfast second-order method* with the convergence rate $\mathcal{O}(k^{-4})$ has been presented, which is faster than the classical lower bound $\mathcal{O}(k^{-7/2})$. The latter method consists of an implementation of a third-order tensor method where its auxiliary problem is handled by a Bregman gradient method requiring second-order oracles, i.e., this scheme is implemented as a second-order method. We note that this method assumes the Lipschitz continuity of third derivatives while the classical second-order methods apply to problems with Lipschitz continuous Hessian. This clearly explains that the convergence rate $\mathcal{O}(k^{-4})$ for this method is not a contradiction with classical complexity theory for second-order methods.

One of the classical methods for solving optimization problems is the *proximal-point* method that is given by

$$x_{k+1} = \arg \min_{x \in \mathbb{E}} \left\{ h(x) + \frac{1}{2\lambda} \|x - x_k\|^2 \right\}, \quad (1.1)$$

for the function $h(\cdot)$, a given point x_k , and $\lambda > 0$. The first appearance of this algorithm dated back to 1970 in the works of Martinet [19, 20], which is further studied by Rockafellar [32] when λ is replaced by a sequence of positive numbers $\{\lambda_k\}_{k \geq 0}$. Since its first presentation, this algorithm has been subject of great interest in both Euclidean and non-Euclidean settings, and many extensions has been proposed; for example see [5, 9, 11, 15, 16, 33].

Recently, Nesterov in [29] proposed a *bi-level unconstrained minimization* (BLUM) framework by defining a novel high-order proximal-point operator using a p th-order regularization term

$$\text{prox}_{h/H}^p(\bar{x}) = \arg \min_{x \in \mathbb{E}} \left\{ h(x) + \frac{H}{p+1} \|x - \bar{x}\|^{p+1} \right\},$$

see Sect. 2 for more details. This framework consists of two levels, where the upper level involves a scheme using the high-order proximal-point operator, and the lower-level is a scheme for solving the corresponding proximal-point minimization inexactly. Therefore, one has a freedom of choosing the order p of the proximal-point operator and can also choose a proper method to approximate the solution of the proximal-point auxiliary problem. Applying this framework to twice smooth unconstrained problems with $p = 3$, using an accelerated third-order method at the upper level, and solving the auxiliary problem by a Bregman gradient method lead to a second-order method with the convergence rate $\mathcal{O}(k^{-4})$. The main goals of this paper are to extend the results of [29] onto the composite case (i.e., for nonsmooth and constrained problems) and to provide a non-Euclidean method for solving the auxiliary p th-order proximal-point problem with an arbitrary $p \geq 1$.

1.1 Content

In this paper, we introduce a *Bi-level OPTimization* (BiOPT) framework that is an extension of the BLUM framework (see [29]) for the convex composite minimization. In our setting, the objective function is the sum of two convex functions, where one of them is smooth enough. At the first step, we regularize the objective function by a power of the Euclidean norm $\|\cdot\|^{p+1}$ with $p \geq 1$, following the same vein as (1.1). The resulted mapping is called *high-order proximal-point operator*, which is assumed to be minimized approximately at a reasonable cost. If the first function in our composite objective is smooth enough, in Sect. 2, we show that this auxiliary problem can be inexactly solved by one step of the p th-order tensor method (see Sect. 2.1). Afterwards, we show that the basic proximal-point method attains the convergence rate $\mathcal{O}(k^{-p})$ (see Sect. 2.2), while its accelerated counterpart obtains the convergence rate $\mathcal{O}(k^{-(p+1)})$ (see Sect. 2.3).

We next present our bi-level optimization framework in Sect. 3, which opens up entirely new ground for developing highly efficient algorithms for simple constrained and composite minimization problems. In the upper level, we can choose the order p of the proximal-point operator and apply both basic and accelerated proximal-point schemes using the estimation sequence technique. We then assume that the differentiable part of the proximal-point objective is smooth relative to some scaling

function (see [9, 18]) and then design a non-Euclidean composite gradient algorithm using a Bregman distance to solve this auxiliary problem inexactly. It is shown that the latter algorithm will be stopped after $\mathcal{O}(\log \frac{1}{\varepsilon})$ of iterations (for the accuracy parameter $\varepsilon > 0$) if the underlying cost function is relatively strongly convex. Hence, choosing a lower-order scaling function for the Bregman distance, there is a possibility to apply lower-order schemes for solving the auxiliary problem that will lead to lower-order methods in our convex composite setting.

Following our BiOPT framework, we finally pick a constant p for the p th-order proximal-point operator and apply the accelerated method to the composite problem at the upper level. Then, we introduce a high-order scaling function and show that the differentiable part of the proximal-point objective is L -smooth relative to this scaling function, for $L > 0$. We consequently apply the non-Euclidean composite gradient method to the auxiliary problem, which only needs the p th-order oracle for even p and the $(p - 1)$ th-order oracle for odd p . Therefore, we end up with a high-order method with the convergence rate of $\mathcal{O}(k^{-(p+1)})$ under some suitable assumptions. We emphasize while this convergence rate is faster than the classical lower bound $\mathcal{O}(k^{-(3p-2)/2})$ for $p = 3$, it is sub-optimal for other choices of p . However, we show that our method can overpass the classical optimal rates for some class of structured problems. We finally deliver some conclusion in Sect. 4.

1.2 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 the 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 $\bar{x} \in \mathbb{E}$ given by

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

Using the above norm, we can define the standard Euclidean prox-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:

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

Note that function $d_{p+1}(\cdot)$ is uniformly convex (see, for example, [26, Lemma 4.2.3]):

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

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

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

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

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

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, [31, Appendix 1])

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

In this paper, we work with functions from the problem classes \mathcal{F}_p , which are convex and p times continuously differentiable on \mathbb{E} . Denote by $M_p(f)$ its uniform upper bound for its p th derivative:

$$M_p(f) = \sup_{x \in \mathbb{E}} \|D^p f(x)\|. \quad (1.7)$$

2 Inexact high-order proximal-point methods

Let function $f : \mathbb{E} \rightarrow \mathbb{R}$ be closed convex and smooth enough and let $\psi : \mathbb{E} \rightarrow \mathbb{R}$ be a simple closed convex function such that $\text{dom } \psi \subseteq \text{int}(\text{dom } f)$. We now consider the

convex composite minimization problem

$$\min_{x \in \text{dom}\psi} \{F(x) = f(x) + \psi(x)\}, \tag{2.1}$$

where it is assumed that (2.1) has at least one optimal solution $x^* \in \text{dom}\psi$ and $F^* = F(x^*)$. This class of problems is general enough to encompass many practical problems from many application fields such as signal and image processing, machine learning, statistics, and so on. In particular, for the simple closed convex set $Q \subseteq \mathbb{E}$, the simple constrained problem

$$\begin{aligned} \min & f(x) \\ \text{s.t.} & x \in Q \end{aligned} \tag{2.2}$$

can be rewritten in the form (2.1), i.e.,

$$\min_{x \in \text{dom}\psi} f(x) + \delta_Q(x), \tag{2.3}$$

where $\delta_Q(\cdot)$ is the indicator function of the set Q given by

$$\delta_Q(x) = \begin{cases} 0 & \text{if } x \in Q, \\ +\infty & \text{if } x \notin Q. \end{cases}$$

Let us define the *p*th-order composite proximal-point operator

$$\text{prox}_{F/H}^p(\bar{x}) = \arg \min_{x \in \text{dom}\psi} \{f(x) + \psi(x) + Hd_{p+1}(x - \bar{x})\}, \tag{2.4}$$

for $H > 0$ and $p \geq 1$, which is an extension of the *p*th-order proximal-point operator given in [29]. Moreover, if $p = 1$, it reduces to the classical proximal operator

$$\text{prox}_{F/H}(\bar{x}) = \arg \min_{x \in \text{dom}\psi} \left\{ f(x) + \psi(x) + \frac{H}{2} \|x - \bar{x}\|^2 \right\}.$$

Our main objective is to investigate the global rate of convergence of high-order proximal-point methods in accelerated and non-accelerated forms, where we approximate the proximal-point operator (2.4) and study the complexity of such approximation. To this end, let us introduce the set of *acceptable solutions* of (2.4) by

$$\begin{aligned} \mathcal{A}_p^H(\bar{x}, \beta) &= \{(x, g) \in \text{dom}\psi \times \mathbb{E}^* : g \in \partial\psi(x), \|\nabla f_{\bar{x},p}^H(x) + g\|_* \\ &\leq \beta \|\nabla f(x) + g\|_*\}, \end{aligned} \tag{2.5}$$

where

$$f_{\bar{x},p}^H(x) = f(x) + Hd_{p+1}(x - \bar{x}), \tag{2.6}$$

where $\beta \in [0, 1)$ is the tolerance parameter. Note that if $\psi \equiv 0$, then the set $\mathcal{A}_p^H(\bar{x}, \beta)$ leads to *inexact acceptable solutions* for the problem (2.4), which was recently studied for smooth convex problems in [29]. Let us emphasize that extending the definition

of inexact acceptable solutions from [29] for nonsmooth functions is not a trivial task because not all subgradients $g \in \partial\psi(x)$ satisfy the inequality (2.5). In the more general setting of the composite minimization, we address this issue in Sect. 3.1 using a non-Euclidean composite gradient scheme that suggests which subgradient $g \in \partial\psi(x) \neq \emptyset$ can be explicitly used in (2.5).

Since function $F(\cdot)$ is convex and $d_{p+1}(\cdot)$ is uniformly convex, the minimization problem (2.4) has a unique solution that we assume to be computable at reasonable cost. Let us first see how the exact solution of (2.4) satisfies (2.5). The first-order optimality conditions for (2.4) ensure that

$$H\|T - \bar{x}\|^{p-1}B(\bar{x} - T) - \nabla f(T) \in \partial\psi(T).$$

Thus, for $g = H\|T - \bar{x}\|^{p-1}B(\bar{x} - T) - \nabla f(T)$, the inequality in (2.5) holds with any $\beta \in [0, 1)$, i.e., $(\text{prox}_{F/H}^p(\bar{x}), g) \in \mathcal{A}_p^H(\bar{x}, \beta)$. Furthermore, since $\nabla f_{\bar{x},p}^H(\bar{x}) = \nabla f(\bar{x})$, we have $(\bar{x}, g) \notin \mathcal{A}_p^H(\bar{x}, \beta)$ except if $\bar{x} = x^*$. In the next subsection, we show that an acceptable approximation of the operator (2.4) can be computed by applying one step of the p th-order tensor method (see [27]) satisfying (2.5), while a lower-order method will be presented in Sect. 3.1. Let us highlight here that we are not able to find an inexact solution in the sense of (2.5) for all points \bar{x} in a neighbourhood of the solution x^* ; however its exact solution always satisfies this inequality. We study this in the following example.

Example 2.1 Let us consider the minimization of function $f : \mathbb{R} \rightarrow \mathbb{R}$ given by $f(x) = x$ over the set $Q = \{x \in \mathbb{R} : x \geq 0\}$, where $x^* = 0$ is its unique solution. The indicator function of the set Q is given by $\psi : \mathbb{R} \rightarrow \mathbb{R}$ that is

$$\psi(x) = \delta_Q(x) = \begin{cases} 0 & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0, \end{cases}$$

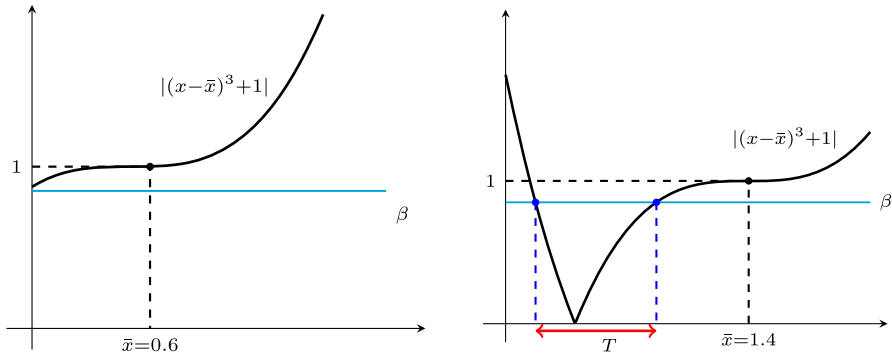
where its subdifferential is given by

$$\partial\psi(x) = \begin{cases} (-\infty, 0] & \text{if } x = 0, \\ \{0\} & \text{if } x > 0, \\ \emptyset & \text{if } x < 0. \end{cases}$$

Let us set $H = 1, B = 1, p = 3, \bar{x} \neq 0$. Hence, $f_{\bar{x},H}^3(x) = x + \frac{1}{4}|x - \bar{x}|^4$ that for $x \geq 0$ and $g \in \partial\psi(x)$ yield $\|\nabla f_{\bar{x},H}^3(x) + g\|_* = |1 + g + (x - \bar{x})^3|, \|\nabla f(x) + g\|_* = |1 + g|$. Therefore, for $\beta \in [0, 1)$, the inequality $\|\nabla f_{\bar{x},H}^3(T) + g\|_* \leq \beta\|\nabla f(T) + g\|_*$ leads to $|1 + g + (T - \bar{x})^3| \leq \beta|1 + g|$, i.e.,

$$\bar{x} - \sqrt[3]{1 + g + \beta|1 + g|} \leq T \leq \bar{x} - \sqrt[3]{1 + g - \beta|1 + g|}, \quad T \geq 0.$$

It is clear that there is no $T > 0$ (i.e., $g = 0$) such that the right-hand side inequality holds if we have $\bar{x} < \sqrt[3]{1 - \beta}$ [see Subfigure (a) of Fig. 1]. In this case, only the exact solution $T = 0$ of the auxiliary problem satisfies the inequality (2.5). Indeed, $(T, g) \in \mathcal{A}_p^H(\bar{x}, \beta)$ if we have



(a) No acceptable solutions for $\bar{x} = 0.6$. (b) Acceptable solutions for $\bar{x} = 1.4$.

Fig. 1 Subfigure (a) shows that for $\bar{x} < \sqrt[3]{1-\beta}$, only the exact solution of the auxiliary problem satisfies (2.5), and Subfigure (b) illustrates the set of solutions for $\bar{x} = 1.4$ and $\beta = 0.85$ satisfying $\bar{x} \geq \sqrt[3]{1+\beta}$

$$T \in \begin{cases} [\bar{x} - \sqrt[3]{1+g+\beta|1+g|}, \bar{x} - \sqrt[3]{1+g-\beta|1+g|}] & \text{if } \bar{x} - \sqrt[3]{1+g+\beta|1+g|} \geq 0, \\ [0, \bar{x} - \sqrt[3]{1+g-\beta|1+g|}] & \text{if } \bar{x} - \sqrt[3]{1+g+\beta|1+g|} < 0, \end{cases}$$

which we illustrate in Subfigure (b) of Fig. 1 for some special choices of β and \bar{x} .

We first present the following lemma, which is a direct consequence of the definition of acceptable solutions (2.5).

Lemma 2.2 (Properties of acceptable solutions) *Let $(T, g) \in \mathcal{A}_p^H(\bar{x}, \beta)$ for some $g \in \partial\psi(T)$. Then, we have*

$$(1 - \beta)\|\nabla f(T) + g\|_* \leq H\|T - \bar{x}\|^p \leq (1 + \beta)\|\nabla f(T) + g\|_*, \tag{2.7}$$

$$\langle \nabla f(T) + g, \bar{x} - T \rangle \geq \frac{H}{1+\beta}\|T - \bar{x}\|^{p+1}. \tag{2.8}$$

If additionally $\beta \leq \frac{1}{p}$, then

$$\langle \nabla f(T) + g, \bar{x} - T \rangle \geq \left(\frac{1-\beta}{H}\right)^{1/p} \|\nabla f(T) + g\|_*^{\frac{p+1}{p}}. \tag{2.9}$$

Proof From (2.5) and the reverse triangle inequality, we obtain

$$\begin{aligned} |H\|T - \bar{x}\|^p - \|\nabla f(T) + g\|_*| &\leq \|\nabla f(T) + H\|T - \bar{x}\|^{p-1}B(T - \bar{x}) \\ &\quad + g\|_* \leq \beta\|\nabla f(T) + g\|_*, \end{aligned}$$

i.e., the inequality (2.7) holds. Squaring both sides of the inequality in (2.5), we come to

$$\begin{aligned} \|\nabla f(T) + g\|_*^2 + 2H\|T - \bar{x}\|^{p-1} \langle \nabla f(T) + g, B(T - \bar{x}) \rangle \\ + H^2\|T - \bar{x}\|^{2p} \leq \beta^2\|\nabla f(x) + g\|_*^2, \end{aligned}$$

leading to

$$\begin{aligned}
 \langle \nabla f(T) + g, B(\bar{x} - T) \rangle &\geq \frac{1-\beta^2}{2H\|T-\bar{x}\|^{p-1}} \|\nabla f(T) + g\|_*^2 + \frac{H}{2} \|T - \bar{x}\|^{p+1} \\
 &\geq \frac{H(1-\beta^2)}{2(1+\beta)^2} \|T - \bar{x}\|^{p+1} + \frac{H}{2} \|T - \bar{x}\|^{p+1} \\
 &= \frac{H}{(1+\beta)} \|T - \bar{x}\|^{p+1},
 \end{aligned}
 \tag{2.10}$$

giving (2.8). Let us consider the function $\zeta : \mathbb{R}_+ \rightarrow \mathbb{R}$ with $\zeta(r) = \frac{1-\beta^2}{2Hr^{p-1}} \|\nabla f(T) + g\|_*^2 + \frac{H}{2} r^{p+1}$, which is the right-hand side of the inequality (2.10) with $r = \|T - \bar{x}\|$. From the inequality (2.7), we obtain $r \geq \hat{r} = \left(\frac{1-\beta}{H} \|\nabla f(T) + g\|_*\right)^{1/p}$. Taking the derivative of ζ at \hat{r} and $\beta \leq \frac{1}{p}$, we get

$$\begin{aligned}
 \zeta'(\hat{r}) &= \left(\frac{(1-p)(1+\beta)}{2} + \frac{(p+1)(1-\beta)}{2}\right) \|\nabla f(T) + g\|_* \\
 &= (1 - \beta p) \|\nabla f(T) + g\|_* \geq 0.
 \end{aligned}$$

Together with (2.10), this implies (2.9). □

2.1 Solving (2.4) with p th-order tensor methods

In this section, we assume that $f(\cdot)$ is p th-order differentiable with $M_{p+1}(f) < +\infty$ and show that an acceptable solution satisfying the inequality (2.5) can be obtained by applying one step of the tensor method given in [27].

The Taylor expansion of the function $f(\cdot)$ at $x \in \mathbb{E}$ is denoted by

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

and it holds that

$$\|\nabla f(y) - \nabla \Omega_{x,p}(y)\|_* \leq \frac{M_{p+1}(f)}{p!} \|y - x\|^p. \tag{2.11}$$

For $M > 0$, let us define the *augmented Taylor approximation* as

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

Note that if $M \geq M_{p+1}(f)$, then $F(y) \leq \widehat{\Omega}_{x,p}(y) + \psi(y)$, which is a uniform upper bound for $F(\cdot)$. In the case $M \geq pM_{p+1}(f)$, the function $\widehat{\Omega}_{x,p}(y) + \psi(y)$ is convex, as confirmed by [27, Theorem 1], which implies that one will be able to minimize the problem (2.1) by the *tensor step*, i.e.,

$$T_{p,M}^{f,g}(x) = \arg \min_{y \in \text{dom}\psi} \widehat{\Omega}_{x,p}(y) + \psi(y). \tag{2.12}$$

We next show that an approximate solution of (2.12) can be employed as an acceptable solution of the proximal-point operator (2.4) by the inexact p th-order tensor method proposed in [14, 27].

Lemma 2.3 (Acceptable solutions by the tensor method (2.12)) *Let $(1 - \gamma)M > M_{p+1}(f)$ and the approximate solution T of (2.12) satisfies*

$$\|\nabla \widehat{\Omega}_{x,p}(T) + g\|_* \leq \frac{\gamma}{1+\gamma} \|\nabla \Omega_{x,p}(T) + g\|_*, \quad (2.13)$$

for some $g \in \partial\psi(T)$ and $\gamma \in \left[0, \frac{\beta}{1+\beta}\right)$. Then, for point $T = T_{p,M}^{f,g}(x)$, it holds

$$\|\nabla f(T) + \frac{M}{p!} \nabla d_{p+1}(T - x) + g\|_* \leq \frac{M_{p+1}(f) + \gamma M}{(1-\gamma)M - M_{p+1}(f)} \|\nabla f(T) + g\|_*. \quad (2.14)$$

Proof It follows from (2.13) that

$$\frac{\gamma}{1+\gamma} \|\nabla \Omega_{x,p}(T) + g\|_* \geq \|\nabla \widehat{\Omega}_{x,p}(T) + g\|_* \geq \|\nabla \Omega_{x,p}(T) + g\|_* - \frac{M}{p!} \|T - x\|^p,$$

which consequently implies

$$\|\nabla \Omega_{x,p}(T) + g\|_* \leq (1 + \gamma) \frac{M}{p!} \|T - x\|^p,$$

for some $g \in \partial\psi(T)$. Together with (1.3), (2.11), and (2.13), this yields

$$\begin{aligned} \frac{M_{p+1}(f)}{p!} \|T - x\|^p &\geq \|\nabla f(T) - \nabla \Omega_{x,p}(T)\|_* \\ &= \|\nabla f(T) - \nabla \widehat{\Omega}_{x,p}(T) + \frac{M}{p!} \nabla d_{p+1}(T - x)\|_* \\ &= \|\nabla f(T) + \frac{M}{p!} \nabla d_{p+1}(T - x) + g - (\nabla \widehat{\Omega}_{x,p}(T) + g)\|_* \\ &\geq \|\nabla f(T) + \frac{M}{p!} \nabla d_{p+1}(T - x) + g\|_* - \|\nabla \widehat{\Omega}_{x,p}(T) + g\|_* \\ &\geq \|\nabla f(T) + \frac{M}{p!} \nabla d_{p+1}(T - x) + g\|_* - \frac{\gamma M}{p!} \|T - x\|^p \\ &\geq \|\frac{M}{p!} \nabla d_{p+1}(T - x)\|_* - \|\nabla f(T) + g\|_* - \frac{\gamma M}{p!} \|T - x\|^p, \end{aligned}$$

implying $\frac{1}{p!} \|T - x\|^p \leq \frac{1}{(1-\gamma)M - M_{p+1}(f)} \|\nabla f(T) + g\|_*$. This and the inequality

$$\|\nabla f(T) + \frac{M}{p!} \nabla d_{p+1}(T - x) + g\|_* \leq \frac{M_{p+1}(f) + \gamma M}{p!} \|T - x\|^p,$$

obtained in the above chain lead to the desired result (2.14). \square

We note that setting $M = \frac{1+\beta}{\beta(1-\gamma)-\gamma} M_{p+1}(f)$ and $H = \frac{M}{p!}$, the inequality (2.14) can be rewritten in the form

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

which implies $(T, g) \in \mathcal{A}_p^H(x, \beta)$. In order to illustrate the results of Lemma 2.3, we study the following one-dimensional example.

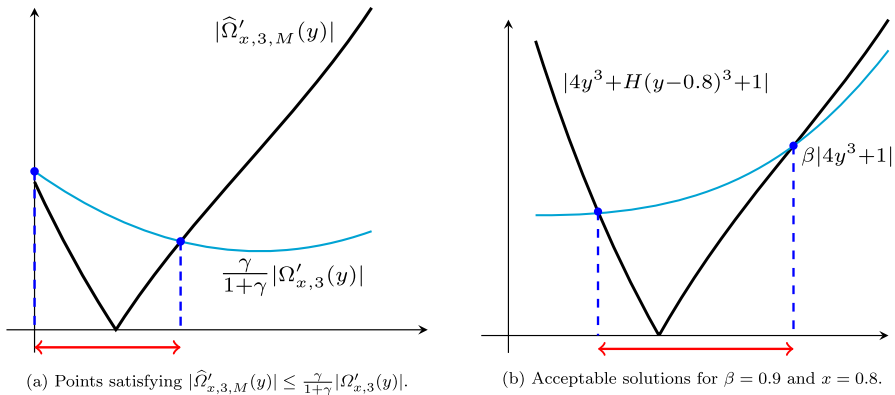


Fig. 2 Subfigure (a) stands for the set of points y satisfying the inequality $|\widehat{\Omega}'_{x,3,M}(y)| \leq \frac{\gamma}{1+\gamma} |\Omega'_{x,3}(y)|$ with $x = 0.8$, $\gamma = 8/19$, and $\beta = 0.9$, and Subfigure (b) illustrates the set of acceptable solutions for $x = 0.8$ and $\beta = 0.9$

Example 2.4 Let us consider the minimization of the one-dimensional function $F : \mathbb{R} \rightarrow \mathbb{R}$ given by $F(x) = x^4 + |x|$, where $x^* = 0$ is its unique solution. In the setting of the problem (2.1), we have $f(x) = x^4$ and $\psi(x) = |x|$. Let us set $p = 3$, i.e., we have $M_4(f) = 24$ and

$$\begin{aligned} \Omega_{x,3}(y) &= x^4 + 4x^3(y-x) + 6x^2(y-x)^2 + 4x(y-x)^3, \\ \widehat{\Omega}_{x,3,M}(y) &= \Omega_{x,3}(y) + \frac{M}{24}(y-x)^4, \end{aligned}$$

where $M = 1.9M_4(f)$. Thus,

$$\Omega'_{x,3}(y) = 4x^3 + 12x^2(y-x) + 12x(y-x)^2, \quad \widehat{\Omega}'_{x,3,M}(y) = \Omega'_{x,3}(y) + \frac{M}{6}(y-x)^3.$$

Setting $\gamma = \frac{8}{19} \in [0, \frac{9}{19}]$ and $x = 0.8$, we illustrate the feasible area of $|\widehat{\Omega}'_{x,3,M}(y)| \leq \frac{\gamma}{1+\gamma} |\Omega'_{x,3}(y)|$ and acceptable solutions in Subfigures (a) and (b) of Fig. 2, respectively. We note that with our choice of γ and M , we have $(1-\gamma)M > M_4(f)$, which implies that all assumptions of Lemma 2.3 are valid.

In Sect. 3, we further extend our discussion concerning the computation of an acceptable solution $\mathcal{A}_p^H(\bar{x}, \beta)$ for the p th-order proximal-point problem (2.4) by the lower-level methods.

2.2 Inexact high-order proximal-point method

In this section, we introduce our inexact high-order proximal-point method for the composite minimization (2.1) and verify its rate of convergence.

We now consider our first inexact high-order proximal-point scheme that generates a sequence of iterations satisfying

$$(T_k, g_{k+1}) \in \mathcal{A}_p^H(x_k, \beta), \tag{2.15}$$

for $g_{k+1} \in \partial\psi(T_k)$ which we summarize in Algorithm 1.

Algorithm 1: Inexact High-Order Proximal-Point Algorithm	
Input: $x_0 \in \text{dom}\psi$, $\beta \in [0, 1/p]$, $H > 0$, $\varepsilon > 0$, $k = 0$;	
1	begin
2	while $F(x_k) - F^* > \varepsilon$ do
3	Find $(T_k, g_{k+1}) \in \mathcal{A}_p^H(x_k, \beta)$ with $g_{k+1} \in \partial\psi(T_k)$ and set $x_{k+1} = T_k$; $k = k + 1$;
4	end
5	end

In order to verify the convergence rate of Algorithm 1, we need the next lemma, which was proved in [28, Lemma 11].

Lemma 2.5 [28, Lemma 11] *Let $\{\xi_k\}_{k \geq 0}$ be a sequence of positive numbers satisfying*

$$\xi_k - \xi_{k+1} \geq \xi_{k+1}^{1+\alpha}, \quad k \geq 0, \tag{2.16}$$

for $\alpha \in (0, 1]$. Then, for $k \geq 0$, the following holds

$$\xi_k \leq \frac{\xi_0}{\left(1 + \frac{\alpha k}{1+\alpha} \log(1+\xi_0^\alpha)\right)^{1/\alpha}} \leq \left(\left(1 + \frac{1}{\alpha}\right) (1 + \xi_0^\alpha) \cdot \frac{1}{k}\right)^{1/\alpha}. \tag{2.17}$$

Let us investigate the rate of convergence of Algorithm 1. Let us first define the radius of the initial level set of the function ψ in (2.1) as $D_0 = \max_{x \in \text{dom}\psi} \{\|x - x^*\| : F(x) \leq F(x_0)\} < +\infty$.

Theorem 2.6 (Convergence rate of Algorithm 1) *Let the sequence $\{x_k\}_{k \geq 0}$ be generated by the inexact high-order proximal-point method (2.15) with $\beta \in [0, 1/p]$. Then, for $k \geq 0$, we have*

$$F(x_k) - F^* \leq \frac{1}{2} \left(\frac{1}{1-\beta} H D_0^{p+1} + F(x_0) - F^*\right) \left(\frac{2p+2}{k}\right)^p. \tag{2.18}$$

Proof From the convexity of $\psi(\cdot)$ and (2.9), we obtain

$$\begin{aligned} F(x_k) - F(x_{k+1}) &\geq \langle \nabla f(x_{k+1}) + g_{k+1}, x_{k+1} - x_k \rangle \\ &\geq \left(\frac{1-\beta}{H}\right)^{1/p} \|\nabla f(x_{k+1}) + g_{k+1}\|_*^{\frac{p+1}{p}}, \end{aligned}$$

with $g_{k+1} \in \partial\psi(x_{k+1})$ and $(x_{k+1}, g_{k+1}) \in \mathcal{A}_p^H(x_k, \beta)$. By Cauchy–Schwartz inequality, we get

$$\begin{aligned} F(x_{k+1}) - F^* &\leq \langle \nabla f(x_{k+1}) + g_{k+1}, x_{k+1} - x^* \rangle \\ &\leq \|\nabla f(x_{k+1}) + g_{k+1}\|_* \|x_{k+1} - x^*\| \\ &\leq D_0 \|\nabla f(x_{k+1}) + g_{k+1}\|_*. \end{aligned}$$

It follows from the last two inequalities, that

$$F(x_k) - F(x_{k+1}) \geq \left(\frac{1-\beta}{HD_0^{p+1}}\right)^{1/p} (F(x_{k+1}) - F^*)^{\frac{p+1}{p}}.$$

Setting $\xi_k = \frac{1-\beta}{HD_0^{p+1}}(F(x_k) - F^*)$ and $\alpha = 1/p$, we see that the condition (2.16) is satisfied for all $k \geq 0$. Therefore, from Lemma 2.5, we have

$$\xi_k \leq \left(\left(1 + \frac{1}{\alpha}\right) (1 + \xi_0^\alpha) \cdot \frac{1}{k} \right)^\alpha \leq \left(1 + \frac{1}{\alpha}\right)^\alpha 2^{\frac{1-\alpha}{\alpha}} (1 + \xi_0) \left(\frac{1}{k}\right)^\alpha,$$

adjusting (2.18). □

2.3 Accelerated inexact high-order proximal-point method

In this section, we accelerate the scheme (2.15) by applying a variant of the *standard estimating sequences technique*, which has been used as a standard tool for accelerating first- and second-order methods; see, e.g., [2, 8, 22–26].

Let $\{A_k\}_{k \geq 0}$ be a sequence of positive numbers generated by $A_{k+1} = A_k + a_{k+1}$ for $a_k > 0$. The idea of the estimating sequences techniques is to generate a sequence of estimating functions $\{\Psi_k(x)\}_{k \geq 0}$ of $F(\cdot)$ in such a way that, at each iteration $k \geq 0$, the inequality

$$A_k F(x_k) \leq \Psi_k^* \equiv \min_{x \in \text{dom}\psi} \Psi_k(x), \quad k \geq 0 \tag{2.19}$$

holds true. Let us set $c_p = \left(\frac{1-\beta}{H}\right)^{1/p}$. Following [29, 30], we set

$$A_k = \left(\frac{c_p}{2}\right)^p \left(\frac{k}{p+1}\right)^{p+1}, \quad a_{k+1} = A_{k+1} - A_k, \quad k \geq 0. \tag{2.20}$$

For $x_0, y_k \in \mathbb{E}$ and $(T_k, g_{k+1}) \in \mathcal{A}_p^H(y_k, \beta)$, let us define the *estimating sequence*

$$\Psi_{k+1}(x) = \begin{cases} d_{p+1}(x - x_0) & \text{if } k = 0, \\ \Psi_k(x) + a_{k+1}[\ell_{T_k}(x) + \psi(x)] & \text{if } k \geq 1. \end{cases} \tag{2.21}$$

Lemma 2.7 *Let the sequence $\{\Psi_k(x)\}_{k \geq 0}$ be generated by (2.21) and $v_k = \arg \min_{x \in \mathbb{E}} \Psi_k(x)$. Then, it holds that*

$$\begin{aligned} A_k F(x) + d_{p+1}(x - x_0) &\geq \Psi_k(x) \\ &\geq \Psi_k^* + \frac{1}{p+1} \left(\frac{1}{2}\right)^{p-1} \|x - v_k\|^{p+1}, \quad \forall x \in \text{dom}\psi, \quad k \geq 0. \end{aligned} \tag{2.22}$$

Proof The proof is given by induction on k . For $k = 0$, $\Psi_0 = d_{p+1}(x - x_0)$ and so (2.22) holds. We now assume that (2.22) holds for k and show it for $k + 1$. Then, it

follows from (2.21) and the subgradient inequality that

$$\begin{aligned}\Psi_{k+1}(x) &= \Psi_k(x) + a_{k+1}[\ell_{x_{k+1}}(x) + \psi(x)] \\ &\leq A_k F(x) + d_{p+1}(x - x_0) + a_{k+1}[\ell_{x_{k+1}}(x) + \psi(x)] \\ &\leq A_k F(x) + d_{p+1}(x - x_0) + a_{k+1} F(x),\end{aligned}$$

leading to (2.22) for $k + 1$. The right-hand side inequality in (2.22) is a direct consequence of the definition of $\Psi_k(\cdot)$ and (1.4). \square

We next present an accelerated version of the scheme (2.5).

Algorithm 2: Accelerated Inexact High-Order Proximal-Point Algorithm

<p>Input: $x_0 \in \text{dom}\psi$, $\beta \in [0, 1/p]$, $H > 0$, $y_0 = v_0 = x_0$, $\Psi_0(x) = d_{p+1}(x - x_0)$, $\varepsilon > 0$, $k = 0$;</p> <p>1 begin</p> <p>2 while $F(x_k) - F^* > \varepsilon$ do</p> <p>3 Compute $v_k = \arg \min_{x \in \mathbb{E}} \Psi_k(x)$ and compute A_{k+1} and a_{k+1} by (2.20);</p> <p>4 Set $y_k = \frac{A_k}{A_{k+1}}x_k + \frac{a_{k+1}}{A_{k+1}}v_k$ and compute $(T_k, g_{k+1}) \in \mathcal{A}_p^H(y_k, \beta)$;</p> <p>5 Find x_{k+1} such that $F(x_{k+1}) \leq F(T_k)$;</p> <p>6 Update $\Psi_{k+1}(x)$ by (2.21) and set $k = k + 1$;</p> <p>7 end</p> <p>8 end</p>
--

In the subsequent result, we investigate the convergence rate of the sequence generated by the accelerated inexact high-order proximal-point method (Algorithm 2).

Theorem 2.8 (Convergence rate of Algorithm 2) *Let the sequence $\{x_k\}_{k \geq 0}$ be generated by Algorithm 2 with $\beta \in [0, 1/p]$. Then, the following statements hold:*

- (i) for all $k \geq 0$, the inequality (2.19) holds;
- (ii) for all $k \geq 0$,

$$F(x_k) - F^* \leq \frac{H}{2(1-\beta)} d_{p+1}(x_0 - x^*) \left(\frac{2p+2}{k} \right)^{p+1}; \quad (2.23)$$

- (iii) $d_{p+1}(v_k - x^*) \leq 2^{p-1} d_{p+1}(x_0 - x^*)$, for all $k \geq 0$.

Proof We first show by induction that (2.19) holds. Since $A_0 = 0$ and $\Psi_0 = d_{p+1}(x - x_0)$, it clearly holds for $k = 0$. We now assume that inequality (2.19) holds for $k \geq 0$

and prove it for $k + 1$. From (2.22), the induction assumption $\Psi_k^* \geq A_k F(x_k)$, and the subgradient inequality, we obtain

$$\begin{aligned} \Psi_{k+1}^* &= \min_{x \in \text{dom}\psi} \{ \Psi_k(x) + a_{k+1}[\ell_{T_k}(x) + \psi(x)] \} \\ &\geq \min_{x \in \text{dom}\psi} \left\{ \Psi_k^* + \sigma_p \|x - v_k\|^{p+1} + a_{k+1}[\ell_{T_k}(x) + \psi(x)] \right\} \\ &\geq \min_{x \in \text{dom}\psi} \left\{ A_k F(x_k) + a_{k+1}[\ell_{T_k}(x) + \psi(x)] + \sigma_p \|x - v_k\|^{p+1} \right\} \\ &\geq \min_{x \in \text{dom}\psi} \left\{ A_k F(x_k) + a_{k+1}[f(T_k) + \langle \nabla f(T_k) + g_{k+1}, x - T_k \rangle + \psi(T_k)] \right. \\ &\quad \left. + \sigma_p \|x - v_k\|^{p+1} \right\} \\ &= \min_{x \in \text{dom}\psi} \left\{ A_{k+1} F(T_k) + \langle \nabla f(T_k) + g_{k+1}, a_{k+1}(x - T_k) + A_k(x_k - T_k) \rangle \right. \\ &\quad \left. + \sigma_p \|x - v_k\|^{p+1} \right\} \\ &\geq \min_{x \in \text{dom}\psi} \left\{ A_{k+1} F(T_k) + \langle \nabla f(T_k) + g_{k+1}, a_{k+1}(x - v_k) + A_{k+1}(y_k - T_k) \rangle \right. \\ &\quad \left. + \sigma_p \|x - v_k\|^{p+1} \right\}, \end{aligned}$$

with $\sigma_p = \frac{1}{p+1} \left(\frac{1}{2}\right)^{p-1}$. For all $x \in \text{dom}\psi$, we have

$$\begin{aligned} &a_{k+1} \langle \nabla f(T_k) + g_{k+1}, x - v_k \rangle + \frac{1}{p+1} \left(\frac{1}{2}\right)^{p-1} \|x - v_k\|^{p+1} \\ &\geq -\frac{p}{p+1} 2^{\frac{p-1}{p}} (a_{k+1} \|\nabla f(T_k) + g_{k+1}\|_*)^{\frac{p+1}{p}}. \end{aligned}$$

It follows from (2.9) and $(T_k, g_{k+1}) \in \mathcal{A}_p^H(y_k, \beta)$ that

$$\langle \nabla f(T_k) + g_{k+1}, y_k - T_k \rangle \geq c_p \|\nabla f(T_k) + g_{k+1}\|_*^{\frac{p+1}{p}}.$$

Combining the last three inequalities yields

$$\begin{aligned} \Psi_{k+1}^* &\geq A_{k+1} F(T_k) + c_p A_{k+1} \|\nabla f(T_k) + g_{k+1}\|_*^{\frac{p+1}{p}} \\ &\quad - \frac{p}{p+1} 2^{\frac{p-1}{p}} (a_{k+1} \|\nabla f(T_k) + g_{k+1}\|_*)^{\frac{p+1}{p}} \\ &= A_{k+1} F(T_k) + \left(c_p A_{k+1} - \frac{p}{p+1} 2^{\frac{p-1}{p}} a_{k+1}^{\frac{p+1}{p}} \right) \|\nabla f(T_k) + g_{k+1}\|_*^{\frac{p+1}{p}} \\ &\geq A_{k+1} F(T_k) + \left(c_p A_{k+1} - 2a_{k+1}^{\frac{p+1}{p}} \right) \|\nabla f(T_k) + g_{k+1}\|_*^{\frac{p+1}{p}}. \end{aligned} \tag{2.24}$$

On the other hand, from (2.20), it can be deduced

$$\begin{aligned} \frac{(A_{k+1} - A_k)^{\frac{p+1}{p}}}{A_{k+1}} &= \frac{c_p \left(\left(\frac{k+1}{p+1} \right)^{p+1} - \left(\frac{k}{p+1} \right)^{p+1} \right)^{\frac{p+1}{p}}}{2 \left(\frac{k+1}{p+1} \right)^{p+1}} \\ &= \frac{c_p}{2} \left(\frac{k+1}{p+1} - \frac{k}{p+1} \left(1 - \frac{1}{k+1} \right)^p \right)^{\frac{p+1}{p}} \leq \frac{c_p}{2}, \end{aligned}$$

leading to

$$a_k^{\frac{p+1}{p}} \leq \frac{c_p}{2} A_{k+1}, \quad k \geq 0.$$

Together with (2.24) and $f(T_k) \geq F(x_{k+1})$, this ensures $\Psi_{k+1}^* \geq A_{k+1}F(x_{k+1})$, i.e., the assertion (i) holds. Invoking the inequalities (2.19) and (2.22), we come to

$$F(x_k) - F^* \leq \frac{1}{A_k} d_{p+1}(x_0 - x^*) = \left(\frac{2}{c_p} \right)^p \left(\frac{p+1}{k} \right)^{p+1} d_{p+1}(x_0 - x^*),$$

adjusting the inequality (2.23).

It follows from (2.19), (2.22), $F(x_k) - F^* \geq 0$, and $x = x^*$ that

$$\begin{aligned} d_{p+1}(x_0 - x^*) &\geq -A_k F^* + \Psi_k^* + \left(\frac{1}{2} \right)^{p-1} d_{p+1}(v_k - x^*) \\ &\geq (F(x_k) - F^*) + \left(\frac{1}{2} \right)^{p-1} d_{p+1}(v_k - x^*), \end{aligned}$$

which leads to the assertion (iii). □

3 BiOPT: Bi-level OPTimization framework

As we have seen in the previous sections, solving the convex composite problem (2.1) by an inexact high-order proximal-point method involves two steps: (i) choosing a p th-order proximal-point method as an upper-level scheme; (ii) choosing a lower-level method for computing a point $(T, g) \in \mathcal{A}_p^H(\bar{x}, \beta)$. This gives us two degrees of freedom in the strategy of finding a solution to the problem (2.1), which is why we call this framework *Bi-level OPTimization* (BiOPT). At the upper level, we do not need to impose any assumption on the objective $F(\cdot)$ apart from its convexity. At the lower-level method, we need some additional assumption on this objective function. Moreover, in the BiOPT setting, the complexity of a scheme leans on the complexity of both upper- and lower-level methods.

On the basis of the results of Sect. 2.1, the auxiliary problem (2.4) can be solved by applying one step of the p th-order tensor method. This demands the computation of i th ($i = 1, \dots, p$) directional derivatives of function $f(\cdot)$ and the condition (2.13), which might not be practical in general. Therefore, we could try to apply a lower-order method to the auxiliary problem (2.4), which leads to an efficient implementation of the BiOPT framework. This is the main motivation of the following sections.

3.1 Non-Euclidean composite gradient method

Let us assume that k is a fixed iteration of either Algorithm 1 or Algorithm 2, and we need to compute an *acceptable solution* z_k of (2.4) satisfying (2.5). To do so, we introduce a non-Euclidean composite gradient method and analyze the convergence properties of the sequence $\{z_i\}_{i \geq 0}$ generated by this scheme, which satisfies in the limit inequality (2.5). Our main tool for such developments is the *relative smoothness condition* (see [9, 18] for more details and examples).

Notice that an acceptable solution of the auxiliary problem (2.4) requires that the function $\varphi_k : \mathbb{E} \rightarrow \mathbb{R}$ given by

$$\varphi_k(z) = f_{y_k, p}^H(z) + \psi(z), \quad \forall k \geq 0, z \in \text{dom} \psi \tag{3.1}$$

be minimized approximately, delivering a point $y_k \in \text{dom} \psi$, satisfying the inequality (2.5). We here define $z_k^* = \arg \min_{z \in \text{dom} \psi} \varphi_k(z)$. Let us consider a simple example in which $f : \mathbb{R} \rightarrow \mathbb{R}$ with $f \equiv 0$ and $y_k = 0$. Then, the function $f_{0, H}^2 : \mathbb{R} \rightarrow \mathbb{R}$ defined as $f_{0, H}^2(z) = \frac{1}{3}|z|^3$ with $\nabla f_{0, H}^2(z) = |z|z$, which is not Lipschitz continuous. This shows that one cannot expect the Lipschitz smoothness of $f_{y_k, p}^H(\cdot)$ for $p \geq 2$. However, it can be shown that this function belongs to a wider class of functions called *relatively smooth*, which we describe next.

Let function $\rho : \mathbb{E} \rightarrow \mathbb{R}$ be closed, convex, and differentiable. We call it a *scaling function*. Now, the non-symmetric *Bregman distance* function $\beta_\rho : \mathbb{E} \times \mathbb{E} \rightarrow \mathbb{R}$ with respect to ρ is given by

$$\beta_\rho(x, y) = \rho(y) - \rho(x) - \langle \nabla \rho(x), y - x \rangle. \tag{3.2}$$

For $x, y, z \in \mathbb{E}$, it is easy to see (e.g., the proof of Lemma 3 in [28]) that

$$\beta_\rho(x, z) - \beta_\rho(y, z) + \beta_\rho(y, x) = \langle \nabla \rho(y) - \nabla \rho(x), z - x \rangle. \tag{3.3}$$

For a convex function $h : \mathbb{E} \rightarrow \mathbb{R}$, we say that $h(\cdot)$ is *L-smooth relative to $\rho(\cdot)$* if there exists a constant $L > 0$ such that $(L\rho - h)(\cdot)$ is convex, and we call it *μ -strongly convex relative to $\rho(\cdot)$* if there exists $\mu > 0$ such that $(h - \mu\rho)(\cdot)$ is convex; cf. [9, 18]. The constant $\kappa = \mu/L$ is called the *condition number* of $h(\cdot)$ relative to the scaling function $\rho(\cdot)$.

In the following lemma, we characterize the latter two conditions.

Lemma 3.1 [18, Proposition 1.1] *The following assertions are equivalent:*

- (i) $h(\cdot)$ is *L-smooth* and *μ -strongly convex relative to the scaling function $\rho(\cdot)$* ;
- (ii) $\mu\beta_\rho(x, y) \leq h(x) - h(y) - \langle \nabla h(y), x - y \rangle \leq L\beta_\rho(x, y)$;
- (iii) $\mu \langle \nabla \rho(y) - \nabla \rho(x), y - x \rangle \leq \langle \nabla h(y) - \nabla h(x), y - x \rangle \leq L \langle \nabla \rho(y) - \nabla \rho(x), y - x \rangle$;
- (iv) $\mu \nabla^2 \rho(x) \preceq \nabla^2 h(x) \preceq L \nabla^2 \rho(x)$.

Let us introduce the following assumptions on the minimization problem (3.1):

- (H1) $\rho(\cdot)$ is uniformly convex of degree $p + 1$ and parameter $\sigma > 0$, i.e., $\beta_\rho(x, y) \geq \frac{\sigma}{p+1} \|y - x\|^{p+1}$;
- (H2) there exist constants $\mu, L > 0$ such that the function $f_{y_k, p}^H(\cdot)$ is L -smooth and μ -strongly convex relative to the scaling function $\rho(\cdot)$.

Note that in (H2) we could introduce the parameters $\mu_{f_{y_k, p}^H}, L_{f_{y_k, p}^H} > 0$; however, for sake of simplicity we use $\mu, L > 0$. In this subsection, for the sake of generality, we assume the existence of the scaling function $\rho(\cdot)$ such that the conditions (H1)–(H2) hold; however, in Sect. 3.2 we introduce a specific scaling function satisfying (H1)–(H2).

We are in position now to develop a non-Euclidean composite gradient scheme for minimizing (3.1) based on the assumptions (H1)–(H2). For given $y_k, z_i \in \text{dom} \psi$ and $H, L > 0$, we introduce the non-Euclidean composite gradient scheme

$$z_{i+1} = \arg \min_{z \in \mathbb{E}} \left\{ \left\langle \nabla f_{y_k, p}^H(z_i), z - z_i \right\rangle + \psi(z) + 2L\beta_\rho(z_i, z) \right\}, \quad z_0 = y_k, \quad (3.4)$$

which is a first-order method and the point z_k^* denotes the optimal solution of (3.4). Note that the first-order optimality conditions for (3.4) leads to the following variational principle

$$\left\langle \nabla f_{y_k, p}^H(z_i) + 2L(\nabla \rho(z_{i+1}) - \nabla \rho(z_i)), z - z_{i+1} \right\rangle + \psi(z) \geq \psi(z_{i+1}). \quad (3.5)$$

For the sequence $\{z_i\}_{i \geq 0}$ generated by the scheme (3.4), we next show the monotonicity of the sequence $\{\varphi_k(z_i)\}_{i \geq 0}$.

Lemma 3.2 (Non-Euclidean composite gradient inequalities) *Let $\{z_i\}_{i \geq 0}$ be generated by the scheme (3.4). Then, for $z_0 = y_k$, it holds that*

$$\varphi_k(z_{i+1}) \leq \varphi_k(z_i) - L\beta_\rho(z_i, z_{i+1}). \quad (3.6)$$

Moreover, we have

$$\beta_\rho(z_{i+1}, z) \leq \vartheta^{i+1} \beta_\rho(y_k, z) + \frac{1}{2L} \left(\frac{1 - \vartheta^{i+1}}{1 - \vartheta} \right) (\varphi_k(z) - \varphi_k(z_{i+1})), \quad (3.7)$$

where $\vartheta = 1 - \frac{\kappa}{2}$.

Proof Since z_{i+1} is a solution of (3.4), it holds

$$\left\langle \nabla f_{y_k, p}^H(z_i), z_{i+1} - z_i \right\rangle + \psi(z_{i+1}) + 2L\beta_\rho(z_i, z_{i+1}) \leq \psi(z_i).$$

Together with the L -smoothness of $f_{y_k, p}^H(\cdot)$ relative to $\rho(\cdot)$, this implies

$$\begin{aligned} f_{y_k, p}^H(z_{i+1}) &\leq f_{y_k, p}^H(z_i) + \left\langle \nabla f_{y_k, p}^H(z_i), z_{i+1} - z_i \right\rangle + L\beta_\rho(z_i, z_{i+1}) \\ &\leq f_{y_k, p}^H(z_i) + \psi(z_i) - \psi(z_{i+1}) - L\beta_\rho(z_i, z_{i+1}), \end{aligned}$$

giving (3.6).

Setting $x = z_{i+1}$ and $y = z_i$ in the three point identity (3.3), applying the inequality (3.5), and using Lemma 3.1(b) it can be concluded that

$$\begin{aligned}
 \beta_\rho(z_{i+1}, z) - \beta_\rho(z_i, z) &= \langle \nabla\rho(z_i) - \nabla\rho(z_{i+1}), z - z_{i+1} \rangle - \beta_\rho(z_i, z_{i+1}) \\
 &\leq \frac{1}{2L} \left[\left\langle \nabla f_{y_k, p}^H(z_i), z - z_{i+1} \right\rangle + \psi(z) - \psi(z_{i+1}) \right] \\
 &\quad - \beta_\rho(z_i, z_{i+1}) \\
 &= \frac{1}{2L} \left[f_{y_k, p}^H(z_i) + \left\langle \nabla f_{y_k, p}^H(z_i), z - z_i \right\rangle + \psi(z) \right] \\
 &\quad - \frac{1}{2L} \left[f_{y_k, p}^H(z_i) + \left\langle \nabla f_{y_k, p}^H(z_i), z_{i+1} - z_i \right\rangle + \psi(z_{i+1}) \right] \\
 &\quad - \beta_\rho(z_i, z_{i+1}) \\
 &\leq \frac{1}{2L} \left[f_{y_k, p}^H(z_i) + \left\langle \nabla f_{y_k, p}^H(z_i), z - z_i \right\rangle + \psi(z) - \varphi_k(z_{i+1}) \right] \\
 &\leq \frac{1}{2L} \left[\varphi_k(z) - \varphi_k(z_{i+1}) - \mu\beta_\rho(z_i, z) \right]. \tag{3.8}
 \end{aligned}$$

Accordingly, since $\varphi_k(z_{i+1}) \leq \varphi_k(z_i)$ for $i \in \mathbb{N}$, we get

$$\begin{aligned}
 \beta_\rho(z_{i+1}, z) &\leq \vartheta\beta_\rho(z_i, z) + \frac{1}{2L} (\varphi_k(z) - \varphi_k(z_{i+1})) \\
 &\leq \vartheta^2\beta_\rho(z_{i-1}, z) + \frac{1}{2L} (1 + \vartheta) (\varphi_k(z) - \varphi_k(z_{i+1})) \\
 &\leq \dots \leq \vartheta^{i+1}\beta_\rho(y_k, z) + \frac{1}{2L} \left(\sum_{j=0}^i \vartheta^j \right) (\varphi_k(z) - \varphi_k(z_{i+1})),
 \end{aligned}$$

justifying the inequality (3.7). □

In summary, we come to the following non-Euclidean composite gradient algorithm.

Algorithm 3: Non-Euclidean Composite Gradient Algorithm

Input: $z_0 = y_k \in \text{dom}\psi$, $\beta \in [0, 1/p]$, $L > 0$, $i = 0$;

```

1 begin
2   repeat
3     Compute  $z_{i+1}$  by (3.4);
4     Set  $g_{i+1} = L(\nabla\rho(z_i) - \nabla\rho(z_{i+1})) - \nabla f_{y_k, p}^H(z_i) \in \partial\psi(z_{i+1})$  and  $i = i + 1$ ;
5   until  $\|\nabla f_{y_k, H}^p(z_{i+1}) + g_{i+1}\|_* \leq \beta\|\nabla f(z_{i+1}) + g_{i+1}\|_*$ 
6    $i_k^* = i$ ;
7 end

```

Output: $T_k = z_{i_k^*}$ and $g_{k+1} = L(\nabla\rho(z_{i_k^*-1}) - \nabla\rho(z_{i_k^*})) - \nabla f_{y_k, p}^H(z_{i_k^*-1}) \in \partial\psi(T_k)$.

We now assume that the auxiliary problem (3.4) can be solved exactly. For the sequence $\{z_i\}_{i \geq 0}$ given by (3.4), we will stop the scheme as soon as $\|\nabla f_{y_k, H}^p(z_{i+1}) + g_{i+1}\|_* \leq \beta\|\nabla f(z_{i+1}) + g_{i+1}\|_*$ holds, and then we set $z_k = z_{i+1}$. In the remainder of this section, we show that this stopping criterion holds for i large enough.

Setting $z = y_k$ in the inequality (3.7), it follows the $(p + 1)$ -uniform convexity of $\rho(\cdot)$ with parameter $\sigma > 0$ and $\vartheta = 1 - \frac{\kappa}{2} \in (0, 1)$ that

$$\begin{aligned} \|y_k - z_{i+1}\|^{p+1} &\leq \frac{p+1}{\sigma} \beta_\rho(z_{i+1}, y_k) \leq \frac{p+1}{2\sigma L} \left(\frac{1 - \vartheta^{i+1}}{1 - \vartheta} \right) (\varphi_k(y_k) - \varphi_k(z_{i+1})) \\ &\leq \frac{p+1}{2\sigma L} \left(\frac{1 - \vartheta^{i+1}}{1 - \vartheta} \right) (\varphi_k(y_k) - \inf \varphi_k) \leq \frac{p+1}{\sigma\mu} (F(y_k) - F^*) \\ &< +\infty, \quad \forall i \in \mathbb{N}, \end{aligned}$$

Let us define the bounded convex set

$$\begin{aligned} \mathcal{L}_k(y_k, \Delta_k) &= \{z \in \mathbb{E} : \|y_k - z\| \leq \Delta_k, \varphi_k(z) \leq \varphi_k(y_k)\}, \\ \Delta_k &= \left(\frac{p+1}{\sigma\mu} (F(y_k) - F^*) \right)^{1/(p+1)}, \end{aligned} \tag{3.9}$$

i.e., $\{z_i\}_{i \geq 0} \subseteq \mathcal{L}_k(y_k, \Delta_k)$.

The next results shows that the sequence $\{\text{dist}(0, \partial\varphi_k(z_i))\}_{i \geq 0}$ vanishes, for $\{z_i\}_{i \geq 0}$ generated by Algorithm 3. For doing so, we also require that

(H3) $\|\nabla^2 \rho(\cdot)\| \leq \bar{L}$ on the set $\mathcal{L}_k(y_k, \Delta_k)$ with $\bar{L} > 0$.

Lemma 3.3 (Subsequential convergence) *Let $\{z_i\}_{i \geq 0}$ be generated by Algorithm 3. If (H1)–(H3) hold, then*

$$\varphi_k(z_i) - \varphi_k(z_{i+1}) \geq C \|\mathcal{G}_{i+1}\|_*^{p+1}, \quad C = \frac{L\sigma}{(p+1)(L-\mu)^{p+1}\bar{L}^{p+1}}, \tag{3.10}$$

where

$$\begin{aligned} \mathcal{G}_{i+1} &= \nabla f_{y_k, p}^H(z_{i+1}) + g_{i+1}, \\ g_{i+1} &= L(\nabla \rho(z_i) - \nabla \rho(z_{i+1})) - \nabla f_{y_k, p}^H(z_i) \in \partial\psi(z_{i+1}). \end{aligned} \tag{3.11}$$

This consequently implies

$$\lim_{i \rightarrow +\infty} \text{dist}(0, \partial\varphi_k(z_{i+1})) = 0. \tag{3.12}$$

Proof Writing the first-order optimality conditions for the minimization problem (3.4), there exists $g_{i+1} \in \partial\psi(z_{i+1})$ such that

$$\nabla f_{y_k, p}^H(z_i) + g_{i+1} + L(\nabla \rho(z_{i+1}) - \nabla \rho(z_i)) = 0,$$

leading to

$$g_{i+1} = L(\nabla \rho(z_i) - \nabla \rho(z_{i+1})) - \nabla f_{y_k, p}^H(z_i).$$

In light of the convexity of $f_{y_k,p}^H(\cdot)$ and $\psi(\cdot)$, we obtain $\partial\varphi_k(\cdot) = \nabla f_{y_k,p}^H(\cdot) + \partial\psi(\cdot)$, i.e.,

$$\mathcal{G}_{i+1} = \nabla f_{y_k,p}^H(z_{i+1}) + g_{i+1} \in \partial\varphi_k(z_{i+1}).$$

On the bounded set $\mathcal{L}_k(y_k, \Delta_k)$, it holds that

$$\begin{aligned} -\mathcal{G}_{i+1} &= L(\nabla\rho(z_{i+1}) - \nabla\rho(z_i)) - (\nabla f_{y_k,p}^H(z_{i+1}) - \nabla f_{y_k,p}^H(z_i)) \\ &= \int_0^1 [(L\nabla^2\rho - \nabla^2 f_{y_k,p}^H)(z_i + \tau(z_{i+1} - z_i))](z_{i+1} - z_i) d\tau. \end{aligned}$$

Let us define

$$\begin{aligned} B &= \left(L\nabla^2\rho(z)\right)^{-\frac{1}{2}} \nabla^2 f_{y_k,p}^H(z) \left(L\nabla^2\rho(z)\right)^{-\frac{1}{2}}, \\ z &= z_i + \tau(z_{i+1} - z_i), \quad \tau \in [0, 1], \end{aligned}$$

which clearly implies

$$\begin{aligned} \left(L\nabla^2\rho(z) - \nabla^2 f_{y_k,p}^H(z)\right)^2 &= \left(\left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} (I - B) \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}}\right)^2 \\ &= \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} (I - B) \left(L\nabla^2\rho(z)\right) (I - B) \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}}. \end{aligned}$$

Together with (H2), (H3), and Lemma 3.1(iv), this leads to

$$\begin{aligned} &\|(L\nabla^2\rho(z) - \nabla^2 f_{y_k,p}^H(z))h\|^2 \\ &= \left\langle [L\nabla^2\rho(z) - \nabla^2 f_{y_k,p}^H(z)]^2 h, h \right\rangle \\ &= \left\langle \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} (I - B) \left(L\nabla^2\rho(z)\right) (I - B) \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} h, h \right\rangle \\ &= \left\langle (I - B) \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} h, \left(L\nabla^2\rho(z)\right) (I - B) \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} h \right\rangle \\ &\leq L\bar{L} \left\langle (I - B)^2 \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} h, \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} h \right\rangle \\ &\leq L\bar{L} \left(1 - \frac{\mu}{L}\right)^2 \left\langle \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} h, \left(L\nabla^2\rho(z)\right)^{\frac{1}{2}} h \right\rangle \\ &\leq L^2\bar{L}^2 \left(1 - \frac{\mu}{L}\right)^2 \|h\|^2. \end{aligned} \tag{3.13}$$

This and (H1) yield

$$\begin{aligned} \|\mathcal{G}_{i+1}\|_* &\leq \|(L\nabla^2\rho - \nabla^2 f_{y_k,p}^H)(z)\| \|z_{i+1} - z_i\| \\ &\leq (L - \mu)\bar{L}\|z_{i+1} - z_i\| \leq (L - \mu)\bar{L} \left(\frac{p+1}{\sigma}\beta_\rho(z_i, z_{i+1})\right)^{1/(p+1)}. \end{aligned}$$

Thus, it can be concluded from (3.6) that

$$\varphi_k(z_i) - \varphi_k(z_{i+1}) \geq L\beta_\rho(z_i, z_{i+1}) \geq \frac{L\sigma}{(p+1)(L-\mu)^{p+1}\bar{L}^{p+1}} \|\mathcal{G}_{i+1}\|_*^{p+1},$$

giving (3.10). Thus, $C \sum \|\mathcal{G}_{i+1}\|_*^{p+1} \leq \varphi_k(y_k) - \inf \varphi_k \leq F(y_k) - F^* < +\infty$, i.e., $\lim_{i \rightarrow \infty} \|\mathcal{G}_{i+1}\| = 0$. Together with the inequality $\text{dist}(0, \partial\varphi_k(z_{i+1})) \leq \|\mathcal{G}_{i+1}\|$, this implies (3.12). \square

We now show the well-definedness and complexity of Algorithm 3 in the subsequent result.

Theorem 3.4 (Well-definedness of Algorithm 3) *Let us assume that all conditions of Lemma 3.3 hold, let $\{z_i\}_{i \geq 0}$ be a sequence generated by Algorithm 3, and let*

$$F(z_i) - F(x^*) \geq \varepsilon, \quad \forall i \geq 0, \tag{3.14}$$

where x^* is a minimizer of $F(\cdot)$ and $\varepsilon > 0$ is the accuracy parameter. Moreover, assume that there exists a constant $D > 0$ such that $\|z_i - x^*\| \leq D$ for all $i \geq 0$. Then, for the subgradients

$$\begin{aligned} \mathcal{G}_k^* &= \nabla f_{y_k,p}^H(z_{i_k}^*) + g_{k+1} \in \partial\varphi_k(z_{i_k}^*), \quad g_{k+1} \\ &= L(\nabla\rho(z_{i_k^*-1}^*) - \nabla\rho(z_{i_k^*}^*)) - \nabla f_{y_k,p}^H(z_{i_k^*-1}^*) \in \partial\psi(z_{i_k^*}^*), \end{aligned}$$

and $z_{i_k^*} \in \text{dom}\psi$, the maximum number of iterations i_k^* needed to guarantee the inequality

$$\|\mathcal{G}_{i_k^*}^*\|_* \leq \beta \|\nabla f(z_{i_k^*}^*) + g_{k+1}\|_* \tag{3.15}$$

satisfies

$$i_k^* \leq 1 + \frac{2(p+1)}{\kappa} \log \left(\frac{D}{\beta} \left(\frac{2L}{C} \beta_\rho(y_k, z_k^*) \right)^{1/(p+1)} \frac{1}{\varepsilon} \right), \tag{3.16}$$

where C is defined in (3.10) and $\varepsilon > 0$ is the accuracy parameter.

Proof Combining the subgradient and Cauchy–Schwartz inequalities with $\|z_i - x^*\| \leq D$, it can be deduced that

$$\|\nabla f(z_i) + g_i\|_* \geq \frac{F(z_i) - F^*}{\|z_i - x^*\|} \geq \frac{F(z_i) - F^*}{D} \geq \frac{\varepsilon}{D}, \tag{3.17}$$

for any $g_i \in \partial\psi(z_i)$. From (3.10), there exists $C > 0$ such that

$$\varphi_k(z_{i_k^*-1}) - \varphi_k(z_k^*) \geq \varphi_k(z_{i_k^*-1}) - \varphi_k(z_{i_k^*}) \geq C \|\mathcal{G}_{i_k^*}\|_*^{p+1},$$

for $\mathcal{G}_{i_k^*} = \nabla f_{y_k,p}^H(z_{i_k^*}) + g_{k+1} \in \partial\varphi_k(z_{i_k^*})$ and $g_{k+1} \in \partial\psi(z_{i_k^*})$. Together with (3.7), this implies

$$\begin{aligned} \|\mathcal{G}_{i_k^*}\|_* &\leq C^{-1/(p+1)} \left(\varphi_k(z_{i_k^*-1}) - \varphi_k(z_k^*) \right)^{1/(p+1)} \\ &\leq \left(\frac{2L}{C} \right)^{1/(p+1)} \left(\left(1 - \frac{\kappa}{2} \right)^{i_k^*-1} \beta_\rho(y_k, z_k^*) - \beta_\rho(z_{i_k^*-1}, z_k^*) \right)^{1/(p+1)} \\ &\leq \left(\frac{2L}{C} \beta_\rho(y_k, z_k^*) \right)^{1/(p+1)} \left(1 - \frac{\kappa}{2} \right)^{(i_k^*-1)/(p+1)}. \end{aligned}$$

Since $1 - \frac{\kappa}{2} \in (0, 1)$, for large enough i_k^* ,

we have $\frac{\beta\varepsilon}{D} \leq \left(\frac{2L}{C} \beta_\rho(y_k, z_k^*) \right)^{1/(p+1)} \left(1 - \frac{\kappa}{2} \right)^{(i_k^*-1)/(p+1)}$, i.e., the bound (3.16) is valid by (3.17) with $i = i_k^*$. □

3.2 Bi-level high-order methods

In the BiOPT framework, we here consider Algorithm 2 using the p th-order proximal-point operator in the upper-level, and in the lower-level we solve the auxiliary problem by the high-order non-Euclidean composite gradient method described in Algorithm 3. As such, our proposed algorithm only needs the p th-order oracle for even p and the $(p - 1)$ th-order oracle for odd p , which attains the complexity of $\mathcal{O}(\varepsilon^{-1/(p+1)})$.

In the remainder of this section, we set $p \geq 2$ and $q = \lfloor p/2 \rfloor$. For $\xi > 1$, let us define the function $\rho_{y_k,p}^H : \mathbb{E} \rightarrow \mathbb{R}$ given by

$$\rho_{y_k,p}^H(x) = \sum_{j=1}^q \frac{2}{(2j)!} D^{2j} f(y_k)[x - y_k]^{2j} + \frac{3H}{2} d_{p+1}(x - y_k), \tag{3.18}$$

which we will remarkably show to be uniformly convex with degree $p + 1$ and parameter $\sigma_{\rho_{y_k,p}^H} = 2^{1-p}H$. For $p = 3$, the scaling function $\rho_k(z) = \frac{1}{2} \langle \nabla^2 f(y_k)(z - y_k), z - y_k \rangle + 3M_4(f)d_4(z - y_k)$ has been suggested in [29], which is slightly different than ours for $p = 3$. Owing to this foundation, we can show that the function $f_{y_k,p}^H(\cdot)$ is L -smooth relative to the scaling function $\rho_{y_k,p}^H(\cdot)$, which paws the way toward algorithmic developments. We begin next with showing the uniform convexity of $\rho_{y_k,p}^H(\cdot)$. To this end, we need the p th-order Taylor expansion of the function $f(\cdot)$ around $y \in \text{dom} f$ given by

$$f(x) = \Omega_{y,p}(x) + \frac{1}{p!} \int_0^1 (1 - \xi)^p D^{p+1} f(y + \xi(x - y))[x - y]^{p+1} d\xi, \tag{3.19}$$

for $x \in \text{dom } f$ and $\Omega_{y,p}(x) = f(y) + \sum_{k=1}^p \frac{1}{k!} D^k f(y)[x-y]^k$. It is not hard to show that

$$\nabla^2 f(x) \leq \nabla^2 \Omega_{y,p}(x) + \frac{M_{p+1}(f)}{(p-1)!} \|x-y\|^{p-1} B, \quad (3.20)$$

see [27, Theorem 1].

Theorem 3.5 (Uniform convexity and smoothness of $\rho_{y_k,p}^H(\cdot)$) *For any $x - y_k \in \mathbb{E}$ and $\xi > 1$, if $p \geq 2$ and $q = \lfloor p/2 \rfloor$, then*

$$-\mathcal{M}_{y_k,p}(x) \leq \sum_{j=1}^q \frac{1}{(2j-1)!} D^{2j+1} f(y_k)[x-y_k]^{2j-1} \leq \mathcal{M}_{y_k,p}(x), \quad (3.21)$$

where

$$\mathcal{M}_{y_k,p}(x) = \sum_{j=1}^q \frac{1}{(2j-2)!} D^{2j} f(y_k)[x-y_k]^{2j-2} + \frac{M_{p+1}(f)}{(p-1)!} \|x-y_k\|^{p-1} B.$$

Moreover, for

$$H = \frac{4M_{p+1}(f)}{(p-1)!}, \quad (3.22)$$

$\rho_{y_k,p}^H(\cdot)$ given in (3.18) is uniformly convex with degree $p+1$ and parameter $\sigma_{\rho_{y_k,p}^H} = 2^{1-p}H$.

Proof Let us fix arbitrary directions $u, h = x - y_k \in \mathbb{E}$. Setting $y = y_k$, it follows from (3.20) that

$$\begin{aligned} 0 &\leq \left\langle \nabla^2 f(x)u, u \right\rangle \leq \left\langle \nabla^2 \Omega_{y_k,p}(x)u, u \right\rangle + \frac{M_{p+1}(f)}{(p-1)!} \|h\|^{p-1} \|u\|^2 \\ &= \left\langle \sum_{i=2}^p \frac{1}{(i-2)!} D^i f(y_k)[h]^{i-2}u, u \right\rangle + \frac{M_{p+1}(f)}{(p-1)!} \|h\|^{p-1} \|u\|^2. \end{aligned}$$

Hence, splitting the sum into the odd and even terms, we come to

$$\begin{aligned} &-\left\langle \sum_{j=1}^q \frac{1}{(2j-2)!} D^{2j} f(y_k)[h]^{2j-2}u, u \right\rangle - \frac{M_{p+1}(f)}{(p-1)!} \|h\|^{p-1} \|u\|^2 \\ &\leq \left\langle \sum_{j=1}^q \frac{1}{(2j-1)!} D^{2j+1} f(y_k)[h]^{2j-1}u, u \right\rangle, \end{aligned}$$

leading to the left hand side of (3.21). Replacing h by $-h$, it holds that

$$\begin{aligned} \left\langle \sum_{j=1}^q \frac{1}{(2j-1)!} D^{2j+1} f(y_k)[h]^{2j-1}u, u \right\rangle &\leq \left\langle \sum_{j=1}^q \frac{1}{(2j-2)!} D^{2j} f(y_k)[h]^{2j-2}u, u \right\rangle \\ &\quad + \frac{M_{p+1}(f)}{(p-1)!} \|h\|^{p-1} \|u\|^2, \end{aligned}$$

giving the right-hand side of (3.21).

From the p th-order Taylor expansion of the function f at y_k , (3.20), (3.21), and (1.3), we obtain

$$\begin{aligned}
 0 \leq \nabla^2 f(x) &\leq \sum_{i=2}^p \frac{1}{(i-2)!} D^i f(y_k)[h]^{i-2} + \frac{1}{(p-1)!} M_{p+1}(f) \|h\|^{p-1} B \\
 &\leq \sum_{j=1}^q \frac{2}{(2j-2)!} D^{2j} f(y_k)[h]^{2j-2} + \frac{2}{(p-1)!} M_{p+1}(f) \|h\|^{p-1} B \\
 &\leq \sum_{j=1}^q \frac{2}{(2j-2)!} D^{2j} f(y_k)[h]^{2j-2} + \frac{2}{(p-1)!} M_{p+1}(f) \nabla^2 d_{p+1}(h) \\
 &\leq \sum_{j=1}^q \frac{2}{(2j-2)!} D^{2j} f(y_k)[h]^{2j-2} + \frac{H}{2} \nabla^2 d_{p+1}(h), \tag{3.23}
 \end{aligned}$$

which implies the convexity of the term $\sum_{j=1}^q \frac{2}{(2j)!} D^{2j} f(y_k)[h]^{2j} + \frac{H}{2} d_{p+1}(h)$. Together with the uniform convexity of $d_{p+1}(\cdot)$ with degree $p + 1$ and parameter 2^{1-p} , this implies the uniform convexity of $\rho_{y_k, H}(\cdot)$ with degree $p + 1$ and parameter $\sigma_{\rho_{y_k, p}^H} = 2^{1-p} H$. \square

Theorem 3.5 clearly implies that the assumption (H1) is satisfied for the scaling function $\rho_{y_k, p}^H(\cdot)$ (3.18). We next show that $f_{y_k, H}^p(\cdot)$ is smooth and strongly convex relative this scaling function for $p = 3$, which is inspired by [27].

Theorem 3.6 (Relative smoothness and strong convexity of $f_{y_k, H}^p(\cdot)$ for $p = 3$) *For $p = 3$ and $\xi > 1$, it holds*

$$\begin{aligned}
 \max \left\{ \frac{1}{2} \left(1 + \frac{1}{\xi} \right), \frac{5+\xi}{6} \right\} \nabla^2 \rho_{y_k, H}(x) &\succeq \nabla^2 f_{y_k, H}^3(x) \\
 &\succeq \min \left\{ \frac{1}{2} \left(1 - \frac{1}{\xi} \right), \frac{5-\xi}{6} \right\} \nabla^2 \rho_{y_k, H}(x), \tag{3.24}
 \end{aligned}$$

In particular, for $\xi = 2$, function $f_{y_k, H}^3 : \mathbb{E} \rightarrow \mathbb{R}$ is $\frac{3}{4}$ -smooth and $\frac{1}{4}$ -strongly convex relative to $\rho_{y_k, p}^H(\cdot)$ defined in (3.18).

Proof Following the proof of Theorem 3.5 for $p = 3$, it is not hard to show that (3.21) is satisfied with

$$\mathcal{M}_{y_k, 3}(x) = \frac{1}{\xi} \left\langle \nabla^2 f(y_k)h, h \right\rangle + \frac{\xi M_4(f)}{2} \|h\|^2 B, \quad h = x - y_k, \quad \xi > 1.$$

From the third-order Taylor expansion of the function f at y_k , (3.20), (3.21), and (1.3), we obtain

$$\begin{aligned} \nabla^2 f(x) &\leq \sum_{k=2}^3 \frac{1}{(k-2)!} D^k f(y_k)[h]^{k-2} + \frac{1}{2} M_4(f) \|h\|^2 B \\ &\leq \left(1 + \frac{1}{\xi}\right) \nabla^2 f(y_k) + \frac{(1+\xi)}{2} M_4(f) \|h\|^2 B \\ &\leq \left(1 + \frac{1}{\xi}\right) \nabla^2 f(y_k) + \frac{(1+\xi)}{2} M_4(f) \nabla^2 d_4(h). \end{aligned}$$

In light of $f_{y_k, H}^3(\cdot) = f(\cdot) + Hd_4(\cdot - y_k)$, $H = 2M_4(f)$, and (3.22), we can write

$$\begin{aligned} \nabla^2 f_{y_k, H}^3(x) &\leq \left(1 + \frac{1}{\xi}\right) \nabla^2 f(y_k) + \left[2M_4(f) + \frac{(1+\xi)}{2} M_4(f)\right] \nabla^2 d_4(x - y_k) \\ &\leq \max\left\{\frac{1}{2}\left(1 + \frac{1}{\xi}\right), \frac{5+\xi}{6}\right\} \nabla^2 \rho_{y_k, H}(x). \end{aligned}$$

On the other hand, applying the third-order Taylor expansion (3.19) and (3.21) yield

$$\begin{aligned} \nabla^2 f(x) &\geq \sum_{k=2}^3 \frac{1}{(k-2)!} D^k f(y_k) - \frac{1}{2} M_4(f) \|x - y_k\|^2 B \\ &\geq \left(1 - \frac{1}{\xi}\right) \nabla^2 f(y_k) - \frac{(1-\xi)}{2} M_4(f) \|h\|^2 B \\ &\geq \left(1 - \frac{1}{\xi}\right) \nabla^2 f(y_k) - \frac{(1-\xi)}{2} M_4(f) \nabla^2 d_4(h). \end{aligned}$$

We therefore have

$$\begin{aligned} \nabla^2 f_{y_k, H}^3(x) &\geq \left(1 - \frac{1}{\xi}\right) \nabla^2 f(y_k) + \left[2M_4(f) + \frac{(1-\xi)}{2} M_4(f)\right] \nabla^2 d_4(x - y_k) \\ &\geq \min\left\{\frac{1}{2}\left(1 - \frac{1}{\xi}\right), \frac{5-\xi}{6}\right\} \nabla^2 \rho_{y_k, H}(x), \end{aligned}$$

giving (3.24). Setting $\xi = 2$ in (3.24) and invoking Lemma 3.1(iv), the claims are valid. □

While Theorem 3.6 establishes the smoothness and strong convexity of $f_{y_k, H}^p(\cdot)$ relative to $\rho_{y_k, p}^H(\cdot)$ for $p = 3$, the relative strong convexity is not true for other $p \geq 2$. As such, in the subsequent result for arbitrary $p \geq 2$, we show that $f_{y_k, H}^p(\cdot)$ is 1-smooth relative to $\rho_{y_k, p}^H(\cdot)$ and $f_{y_k, H}^p(\cdot)$ is uniformly convex with degree $p + 1$ and parameter $\sigma_{f_{y_k, H}^p} = 2^{1-p} H$.

Theorem 3.7 (Relative smoothness of $f_{y_k, H}^p(\cdot)$) *Let $H \geq M_{p+1}(f)$ and let $p \geq 2$ and $q = \lfloor p/2 \rfloor$. Then, the function $f_{y_k, H}^p : \mathbb{E} \rightarrow \mathbb{R}$ is 1-smooth relative to $\rho_{y_k, p}^H(\cdot)$ defined in (3.18). Moreover, the function $f_{y_k, H}^p(\cdot)$ is uniformly convex with degree $p + 1$ and parameter $\sigma_{f_{y_k, H}^p} = 2^{1-p} H$.*

Proof In light of $f_{y_k,p}^H(\cdot) = f(\cdot) + Hd_{p+1}(\cdot - y_k)$ and (3.23), we can write

$$\nabla^2 f_{y_k,p}^H(x) \leq \sum_{j=1}^q \frac{2}{(2j-2)!} D^{2j} f(y_k)[h]^{2j-2} + \frac{3H}{2} \nabla^2 d_{p+1}(h) = \nabla^2 \rho_{y_k,H}(x),$$

implying 1-smooth of $f_{y_k,H}^p(\cdot)$ relative to $\rho_{y_k,p}^H(\cdot)$.

It follows from the convexity of f and the uniform convexity of $d_{p+1}(\cdot)$ with degree $p + 1$ and parameter $\sigma_{d_{p+1}} = 2^{1-p}$ that

$$\begin{aligned} f_{y_k,H}^p(z) &\geq f(x) + \langle \nabla f(x), z - x \rangle \\ &\quad + H \left(d_{p+1}(x) + \langle \nabla d_{p+1}(x), z - x \rangle + \frac{1}{p+1} 2^{1-p} \|z - x\|^{p+1} \right) \\ &= f_{y_k,H}^p(x) + \left\langle \nabla f_{y_k,H}^p(x), z - x \right\rangle + \frac{H}{p+1} 2^{1-p} \|z - x\|^{p+1}, \end{aligned}$$

implying the uniform convexity of $f_{y_k,H}^p(\cdot)$ with degree $p + 1$ and parameter $\sigma_{f_{y_k,H}^p} = 2^{1-p}H$. □

Since $\psi(\cdot)$ is convex and $f_{y_k,H}^p(\cdot)$ is uniformly convex with degree $p + 1$ and parameter $\sigma_{f_{y_k,H}^p} = 2^{1-p}H$, it is clear that $\varphi_k(\cdot)$ is uniformly convex with degree $p + 1$ and parameter $\sigma_{\varphi_k} = 2^{1-p}H$. This and Lemma 3.6 yield

$$\varphi_k(y_k) \geq \dots \geq \varphi_k(z_i) \geq \varphi_k(z_{i+1}) \geq \varphi_k(z_k^*) + \frac{\sigma_{\varphi_k}}{p+1} \|z_{i+1} - z_k^*\|^{p+1},$$

leading to

$$\begin{aligned} \|z_{i+1} - z_k^*\| &\leq \left(\frac{(p+1)!2^{p-3}}{pM_{p+1}(f)} (\varphi_k(y_k) - \varphi_k(z_k^*)) \right)^{\frac{1}{p+1}} \\ &\leq \left(\frac{(p+1)!2^{p-3}}{pM_{p+1}(f)} (F(y_k) - F^*) \right)^{\frac{1}{p+1}}. \end{aligned}$$

In the same way, we have

$$\varphi_k(y_k) \geq \varphi_k(z_k^*) + \frac{\sigma_{\varphi_k}}{p+1} \|y_k - z_k^*\|^{p+1},$$

leading to

$$\begin{aligned} \|y_k - z_k^*\| &\leq \left(\frac{(p+1)!2^{p-3}}{pM_{p+1}(f)} (\varphi_k(y_k) - \varphi_k(z_k^*)) \right)^{\frac{1}{p+1}} \\ &\leq \left(\frac{(p+1)!2^{p-3}}{pM_{p+1}(f)} (F(y_k) - F^*) \right)^{\frac{1}{p+1}}. \end{aligned}$$

Hence, we come to the inequality

$$\begin{aligned} \|z_{i+1} - y_k\| &\leq \|z_{i+1} - z_k^*\| + \|z_k^* - y_k\| \leq \bar{\Delta}_k, \\ \bar{\Delta}_k &= \left(\frac{(p+1)!2^{2p-2}}{pM_{p+1}(f)} (F(y_k) - F^*) \right)^{\frac{1}{p+1}} \quad \forall i \in \mathbb{N}. \end{aligned} \tag{3.25}$$

Next, we define the bounded convex set

$$\mathcal{L}_k(y_k, \bar{\Delta}_k) = \{z \in \mathbb{E} : \|y_k - z\| \leq \bar{\Delta}_k, \varphi_k(z) \leq \varphi_k(y_k)\}, \tag{3.26}$$

i.e., $\{z_i\}_{i \geq 0} \subseteq \mathcal{L}_k(y_k, \bar{\Delta}_k)$. We next show that the scaling function $\rho_{y_k, p}^H(\cdot)$ satisfies (H3).

Lemma 3.8 *For any $x \in \mathcal{L}_k(y_k, \bar{\Delta}_k)$ and $\xi > 1$, if $p \geq 2$ and $q = \lfloor p/2 \rfloor$, then*

$$\|\nabla^2 \rho_{y_k, p}^H(\cdot)\| \leq \bar{L}, \quad \bar{L} = M_4(f)\bar{\Delta}_k^2 + 2M_2(f) + \left(\frac{2M_{p+1}}{(p-1)!} + \frac{3pH}{2} \right) \bar{\Delta}_k^{p-1}, \tag{3.27}$$

where $M_2(f) < +\infty$, $M_4(f) < +\infty$, and $M_{p+1}(f) < +\infty$ and on the set $\mathcal{L}_k(y_k, \bar{\Delta}_k)$ with

$$\bar{\Delta}_k = \left(\frac{(p+1)!2^{2p-2}}{pM_{p+1}(f)} (F(y_k) - F^*) \right)^{\frac{1}{p+1}}.$$

Proof It follows from (3.19) that

$$\begin{aligned} \nabla^2 f(y_k + h) &= \nabla^2 f(y_k) + \sum_{i=3}^p \frac{1}{(i-2)!} D^i f(y_k)[h]^{i-2} + r_{p+1}(h), \\ \nabla^2 f(y_k - h) &= \nabla^2 f(y_k) + \sum_{i=3}^p \frac{(-1)^{i-2}}{(i-2)!} D^i f(y_k)[h]^{i-2} + r_{p+1}(-h), \end{aligned}$$

where

$$\begin{aligned} r_{p+1}(h) &= \frac{p+1}{(p-1)!} \int_0^1 (1 - \xi)^p D^{p+1} f(y_k + \xi h)[h]^{p-1} d\xi, \quad \|r_{p+1}(\pm h)\| \\ &\leq \frac{M_{p+1}(f)}{(p-1)!} \|h\|^{p-1}. \end{aligned}$$

Summing up the latter identities, we come to

$$\nabla^2 f(y_k + h) + \nabla^2 f(y_k - h) - (r_{p+1}(h) + r_{p+1}(-h)) = \sum_{j=1}^q \frac{2}{(2j-2)!} D^{2j} f(y_k)[h]^{2j-2}. \tag{3.28}$$

Moreover, it holds that

$$\begin{aligned} \nabla^2 f(y_k + h) &= \nabla^2 f(y_k) + D^3 f(y_k)[h] + r_4(h), \\ \nabla^2 f(y_k - h) &= \nabla^2 f(y_k) - D^3 f(y_k)[h] + r_4(-h), \end{aligned}$$

leading to

$$\|\nabla^2 f(y_k + h) + \nabla^2 f(y_k - h) - 2\nabla^2 f(y_k)\| \leq r_4(h) + r_4(-h) \leq M_4(f)\|h\|^2. \tag{3.29}$$

In light of (1.3), we have

$$\begin{aligned} \|\nabla^2 \rho_{y_k, H}(x)\| &\leq \left\| \sum_{k=1}^q \frac{2}{(2k-2)!} D^{2k} f(y_k)[h]^{2k-2} \right\| + \frac{3pH}{2} \|h\|^{p-1} \\ &\leq \left\| \nabla^2 f(y_k + h) + \nabla^2 f(y_k - h) - (r_{p+1}(h) + r_{p+1}(-h)) \right\| \\ &\quad + \frac{3pH}{2} \|h\|^{p-1} \\ &\leq \|\nabla^2 f(y_k + h) + \nabla^2 f(y_k - h) - 2\nabla^2 f(y_k)\| \\ &\quad + 2\|\nabla^2 f(y_k)\| + \frac{2M_{p+1}}{(p-1)!} \|h\|^{p-1} + \frac{3pH}{2} \|h\|^{p-1} \\ &\leq M_4(f)\|h\|^2 + 2M_2(f) + \frac{2M_{p+1}}{(p-1)!} \|h\|^{p-1} + \frac{3pH}{2} \|h\|^{p-1}. \end{aligned}$$

For $x \in \mathcal{L}_k(y_k, \bar{\Delta}_k)$ and $h = x - y_k$, we come to

$$\|\nabla^2 \rho_{y_k, H}(x)\| \leq M_4(f)\bar{\Delta}_k^2 + 2M_2(f) + \left(\frac{2M_{p+1}}{(p-1)!} + \frac{3pH}{2}\right)\bar{\Delta}_k^{p-1},$$

establishing (3.27). □

From now on and for sake of simplicity, we denote $\rho_{y_k, p}^H(\cdot)$ and $\sigma_{f_{y_k, H}}^p$ by $\rho_k(\cdot)$ and σ_k , respectively. In order to upper bound the Bregman term $\beta_{\rho_k}(\cdot, \cdot)$, we next define the *norm-dominated scaling function* inspiring by [30, Definition 2], which will be needed in the remainder of this section.

Definition 3.9 The scaling function $\rho(\cdot)$ is called *norm-dominated* on the set $S \subseteq \mathbb{E}$ by some function $\theta_S : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ if $\theta_S(\cdot)$ is convex with $\theta_S(0) = 0$ such that

$$\beta_{\rho}(x, y) \leq \theta_S(\|x - y\|), \tag{3.30}$$

for all $x \in S$ and $y \in \mathbb{E}$.

We first verify the norm-dominatedness of the function $d_{p+1}(\cdot)$ over the set

$$B_R = \{x \in \mathbb{E} \mid \|x\| \leq R\},$$

for $R > 0$.

Lemma 3.10 (Norm-dominatedness of the scaling function $d_{p+1}(\cdot)$) *Let $p \geq 2$ and $q = \lfloor p/2 \rfloor$. Then, the scaling function $d_{p+1}(\cdot)$ is norm-dominated on B_R by the function*

$$\tilde{\theta}_{\mathcal{L}_k}(\tau) = p2^{p-2}R^{p-1}\tau^2 + \frac{2^{p-1}}{p+1}\tau^{p+1}. \tag{3.31}$$

Proof From (1.3), it is clear that

$$\|\nabla^2 d_{p+1}(x)\| = \|\|x\|^{p-1}B + (p-1)\|x\|^{p-3}\| \leq p\|x\|^{p-1}B.$$

Together with the definition of Bregman distances, the inequality $\left(a^{\frac{1}{\theta}} + a^{\frac{1}{\theta}}\right)^\theta \leq 2^{\theta-1}(a+b)$ for $a, b \geq 0$ and $\theta \geq 1$, and $\tau = \|y-x\|$, this implies

$$\begin{aligned} \beta_{d_{p+1}}(x, y) &= d_{p+1}(y) - d_{p+1}(x) - \langle \nabla d_{p+1}(x), y-x \rangle \\ &= \int_0^1 (1-t)\nabla^2 d_{p+1}(x+t(y-x))[y-x]^2 dt \\ &\leq p \int_0^1 (1-t)\|y-x\|^2\|x+t(y-x)\|^{p-1} dt \\ &\leq p \int_0^1 (1-t)\|y-x\|^2(\|x\|+t\|y-x\|)^{p-1} dt \\ &\leq p2^{p-1}\tau^2 \int_0^1 (1-t)(R^{p-1}+t^{p-1}\tau^{p-1}) dt \\ &= p2^{p-2}R^{p-1}\tau^2 + \frac{2^{p-1}}{p+1}\tau^{p+1}, \end{aligned}$$

giving (3.31). □

In order to show the norm-dominatedness of the scaling function $\rho_k(\cdot)$, we also need the following technical lemma.

Lemma 3.11 *Let $p \geq 2$ and $q = \lfloor p/2 \rfloor$, and let the function $\widehat{\rho}_k : \mathbb{E} \rightarrow \mathbb{R}$ be defined by*

$$\widehat{\rho}_k(x) = \sum_{j=1}^q \frac{2}{(2j)!} D^{2j} f(y_k)[x-y_k]^{2j},$$

then $\widehat{\rho}_k(x)$ is \widehat{L} -smooth over

$$B_{\overline{\Delta}_k} = \{z \in \mathbb{E} : \|y_k - z\| \leq \overline{\Delta}_k\},$$

with $\widehat{L} = M_4(f)\overline{\Delta}_k^{-2} + 2M_2(f) + \frac{2M_{p+1}(f)}{(p-1)!}\overline{\Delta}_k^{p-1}$ and $\overline{\Delta}_k = \left(\frac{(p+1)!2^{2p-2}}{pM_{p+1}(f)}(F(y_k) - F^*)\right)^{\frac{1}{p+1}}$.

Proof Setting $h = x - y_k$ and using $x \in B_{\overline{\Delta}_k}$, (3.28) and (3.29), we come to

$$\begin{aligned} \|\nabla^2 \rho_{y_k, H}(x)\| &\leq \left\| \sum_{k=1}^q \frac{2}{(2k-2)!} D^{2k} f(y_k)[h]^{2k-2} \right\| \\ &\leq \left\| \nabla^2 f(y_k + h) + \nabla^2 f(y_k - h) - (r_{p+1}(h) + r_{p+1}(-h)) \right\| \\ &\leq \|\nabla^2 f(y_k + h) + \nabla^2 f(y_k - h) - 2\nabla^2 f(y_k)\| + 2\|\nabla^2 f(y_k)\| \\ &\quad + \frac{2M_{p+1}(f)}{(p-1)!} \|h\|^{p-1} \\ &\leq M_4(f)\|h\|^2 + 2M_2(f) + \frac{2M_{p+1}(f)}{(p-1)!} \|h\|^{p-1} \\ &\leq M_4(f)\overline{\Delta}_k^2 + 2M_2(f) + \frac{2M_{p+1}(f)}{(p-1)!} \overline{\Delta}_k^{p-1}, \end{aligned}$$

which guarantees \widehat{L} -smoothness of $\widehat{\rho}_k(x)$ on $B_{\overline{\Delta}_k}$. □

Invoking Lemma 3.11, we next show that the norm-dominatedness of $\rho_k(\cdot)$ on $B_{\overline{\Delta}_k}$.

Lemma 3.12 (Norm-dominatedness of the scaling function $\rho_k(\cdot)$) *Let $p \geq 2$ and $q = \lfloor p/2 \rfloor$. Then, the function $\rho_k(\cdot)$ is norm-dominated over $B_{\overline{\Delta}_k}$ with $\overline{\Delta}_k =$*

$\left(\frac{(p+1)!2^{2p-2}}{pM_{p+1}(f)} (F(y_k) - F^)\right)^{\frac{1}{p+1}}$ by the function*

$$\theta_{\mathcal{L}_k}(\tau) = \frac{\widehat{L} + 3p2^{p-2}\overline{\Delta}_k^{p-1}H}{2} \tau^2 + \frac{3H2^{p-2}}{p+1} \tau^{p+1}, \tag{3.32}$$

where $\tau \geq 0$.

Proof In light of the definition of $\rho_k(\cdot)$ and the \widehat{L} -smoothness of $\widehat{\rho}_k(\cdot)$, it can be concluded that

$$\begin{aligned} \beta_{\rho_k}(x, y) &= \widehat{\rho}_k(y) - \widehat{\rho}_k(x) - \langle \nabla \widehat{\rho}_k(x), y - x \rangle \\ &\quad + \frac{3}{2}H (d_{p+1}(y - y_k) - d_{p+1}(x - y_k) - \langle \nabla d_{p+1}(x - y_k), y - x \rangle) \\ &\leq \frac{\widehat{L}}{2} \|x - y\|^2 + \frac{3}{2}H \beta_{d_{p+1}}(x - y_k, y - y_k). \end{aligned} \tag{3.33}$$

Invoking Lemmas 3.10 and 3.11, we come to

$$\begin{aligned} \beta_{\rho_k}(x, y) &\leq \frac{\widehat{L}}{2} \|x - y\|^2 + \frac{3}{2}H \beta_{d_{p+1}}(x - y_k, y - y_k) \\ &\leq \frac{\widehat{L}}{2} \|x - y\|^2 + \frac{3}{2}H \left[p2^{p-2}\overline{\Delta}_k^{p-1} \|x - y\|^2 + \frac{2^{p-1}}{p+1} \|x - y\|^{p+1} \right], \end{aligned}$$

adjusting (3.32) for $\tau = \|x - y\|$. □

Motivated by the identity (3.22), in the remainder of this section, we have

$$\sigma_k = \sigma_{\rho_k} = \frac{2^{3-p} M_{p+1}(f)}{(p-1)!}. \quad (3.34)$$

Additionally, in view of (2.20), we consider

$$\beta = \frac{1}{p}, \quad A_k = \frac{(p-1)(p-1)!}{p2^{p+2}M_{p+1}(f)} \left(\frac{k}{p+1}\right)^{p+1}, \quad a_{k+1} = A_{k+1} - A_k, \quad \text{for } k \geq 0. \quad (3.35)$$

We now present our accelerated high-order method by combining all above-mentioned facts into Algorithm 2 leading to the following algorithm.

Algorithm 4: (Bi-Level High-Order Algorithm)

<p>Input: $x_0 \in \text{dom}\psi$, $\beta \in [0, 1/p]$, $H = \frac{4}{(p-1)!}M_{p+1}(f)$, $A_0 = 0$, $\Psi_0(x) = d_{p+1}(x - x_0)$, $k = 0$;</p> <p>1 begin</p> <p>2 while $F(x_k) - F^* > \varepsilon$ do</p> <p>3 Compute $v_k = \arg \min_{x \in \mathbb{E}} \Psi_k(x)$ and compute A_{k+1} and a_{k+1} by (2.20);</p> <p>4 Set $y_k = \frac{A_k}{A_{k+1}}x_k + \frac{a_{k+1}}{A_{k+1}}v_k$ and consider the scaling function (3.18);</p> <p>5 Find $(T_k, g_{k+1}) \in \mathcal{A}_p^H(y_k, \beta)$ with $g_{k+1} \in \partial\psi(T_k)$ by Algorithm 3;</p> <p>6 Find x_{k+1} such that $F(x_{k+1}) \leq F(T_k)$;</p> <p>7 Update $\Psi_{k+1}(x)$ by (2.21) and set $k = k + 1$;</p> <p>8 end</p> <p>9 end</p>
--

Since $L = 1$ in our setting, the optimality conditions for the auxiliary problem (3.4) for the p th-order proximal-point operator is given by

$$\nabla f_{y_k, H}^p(z_i) + \partial\psi(z_{i+1}) + 2(\nabla\rho_k(z_{i+1}) - \nabla\rho_k(z_i)) \ni 0,$$

which should be solved exactly. We next translate this inclusion for convex constrained problem (2.2).

Example 3.13 We here revisit the convex constrained problem (2.2) and its unconstrained version (2.3) with $\psi(\cdot) = \delta_Q(\cdot)$. For $z_i \in \mathbb{E}$ and $L = 1$, writing the first-order optimality conditions leads to

$$N_Q(z_{i+1}) \ni 2(\nabla\rho_k(z_i) - \nabla\rho_k(z_{i+1})) - \nabla f_{y_k, H}^p(z_i), \quad (3.36)$$

where $\partial\psi(z_{i+1}) = N_Q(z_{i+1})$ and therefore the *normal cone*

$$N_Q(x) = \begin{cases} \{u \in \mathbb{E} : \langle u, y - x \rangle \leq 0, \forall y \in Q\} & \text{if } x \in Q, \\ \emptyset & \text{if } x \notin Q \end{cases}$$

plays a crucial role for finding a solution of the auxiliary problem (3.4). As an example, let us consider the Euclidean ball $Q = \{x \in \mathbb{R}^n : \|x\| \leq \delta\}$ for which we have

$$N_Q(x) = \begin{cases} \{\alpha x : \alpha > 0\} & \text{if } \|x\| = \delta, \\ \{0\} & \text{if } \|x\| < \delta. \end{cases}$$

For $p = 3$, our scaling function is given by

$$\rho_k(z) = \left\langle \nabla^2 f(y_k)(z - y_k), z - y_k \right\rangle + \frac{3H}{8} \|z - y_k\|^4.$$

We now consider two cases: (i) $\|z_{i+1}\| < \delta$; (ii) $\|z_{i+1}\| = \delta$. In Case (i), we have

$$\begin{aligned} & (4\nabla^2 f(y_k)(z_{i+1} - z_i) + 3H\|z_{i+1} - y_k\|^2 B(z_{i+1} - y_k) - 3H\|z_i - y_k\|^2 B(z_i - y_k)) \\ & - \nabla f_{y_k, H}^3(z_i) = 0, \end{aligned}$$

with $\nabla f_{y_k, H}^3(z_i) = \nabla f(z_i) + H\|z_i - y_k\|^2(z_i - y_k)$, i.e.,

$$\left[4\nabla^2 f(y_k) + 3H\|z_{i+1} - y_k\|^2 B \right] (z_{i+1} - y_k) = b_i,$$

for $b_i = \left[4\nabla^2 f(y_k) + H\|z_i - y_k\|^2 B \right] (z_i - y_k) + \nabla f_{y_k, H}^3(z_i)$. This consequently implies

$$z_{i+1} = y_k + \left[4\nabla^2 f(y_k) + 3Hr^2 B \right]^{-1} b_i,$$

where $r = \|z_{i+1} - y_k\|$ can be computed by solving the one-dimensional equation

$$r = \left\| \left[4\nabla^2 f(y_k) + 3Hr^2 B \right]^{-1} b_i \right\|.$$

In Case (ii) ($\|z_{i+1}\| = \delta$), there exists $\alpha > 0$ such that

$$\left[4\nabla^2 f(y_k) + 3H\|z_{i+1} - y_k\|^2 B \right] (z_{i+1} - y_k) - b_i = \alpha z_{i+1},$$

leading to

$$z_{i+1} = y_k + \left[4\nabla^2 f(y_k) + (3Hr^2 - \alpha)B \right]^{-1} (b_i + \alpha y_k),$$

where $r = \|z_{i+1} - y_k\|$ and α are obtained by solving the system

$$\begin{cases} r = \left\| \left[4\nabla^2 f(y_k) + (3Hr^2 - \alpha)B \right]^{-1} (b_i + \alpha y_k) \right\|, \\ \delta = \left\| y_k + \left[4\nabla^2 f(y_k) + (3Hr^2 - \alpha)B \right]^{-1} (b_i + \alpha y_k) \right\|. \end{cases}$$

Finally, we come to the solution

$$z_{i+1} = \begin{cases} y_k + [4\nabla^2 f(y_k) + 3Hr^2B]^{-1} b_i & \text{if } \|y_k + [4\nabla^2 f(y_k) + 3Hr^2B]^{-1} b_i\| < \delta, \\ y_k + [4\nabla^2 f(y_k) + (3Hr^2 - \alpha)B]^{-1} (b_i + \alpha y_k) & \text{otherwise,} \end{cases}$$

for the r and α computed by solving the above-mentioned nonlinear systems. □

We now have all the ingredients to address the complexities of the upper and lower levels of Algorithm 4, which is the main result of this section. To this end, for the auxiliary minimization problem (3.4), we assume

$$R_0 = \|x_0 - x^*\|, \quad D_0 = \max_{z \in \text{dom}\psi} \{\|z - x^*\| : F(z) \leq F(x_0)\} < +\infty. \quad (3.37)$$

Let us set $S = \{z \in \text{dom}\psi : \|z - x^*\| \leq 2R_0\}$ and assume

$$F_S = \sup_{z \in S} F(z) < +\infty. \quad (3.38)$$

Let us note that the function $f_{y_k,p}^H(\cdot)$ is not strongly convex relative to the scaling function (3.18) for an arbitrary $p \geq 2$ (except for $p = 3$, i.e., $q = 1$), which implies that the results of Theorem 3.4 cannot be applied here. As such, for arbitrary p , we present the following result thanks to the uniform convexity of $f_{y_k,p}^H(\cdot)$ and $\rho_k(\cdot)$, for which we need the next technical lemma.

Lemma 3.14 *Let $q > 1$, and let*

$$\delta_i - \delta_{i+1} \geq \delta_i^q \quad (3.39)$$

for some integer $i \geq m \geq 0$. Then, $\delta_i < 1$ and

$$\delta_i \leq \frac{\delta_m}{[1 + (q - 1)(i - m)\delta_m^{q-1}]^{\frac{1}{q-1}}} \leq \frac{1}{[(q - 1)(i - m)]^{\frac{1}{q-1}}}. \quad (3.40)$$

Proof The proof is a simple generalization of [13, Lemma 1.1]. □

Theorem 3.15 (Complexity of Algorithm 4) *Let us assume that all conditions of Theorem 3.4 hold, let $p \geq 2$ and $q = \lfloor p/2 \rfloor$, let $M_2(f) < +\infty$, $M_4(f) < +\infty$, and $M_{p+1}(f) < +\infty$. Then,*

(i) *Algorithm 4 attains an ε -solution of the problem (2.1) in*

$$(2p + 2) \left(\frac{3pM_{p+1}(f)}{(p-1)(p+1)(p-1)!\varepsilon} \right)^{\frac{1}{p+1}} R_0$$

iterations, for the accuracy parameter $\varepsilon > 0$.

(ii) For $p = 3$, the auxiliary problem (3.4) is approximately solved by Algorithm 3 in at most

$$1 + 24 \log \left(\frac{\frac{D\bar{L}}{\beta} \left(\frac{\theta_{B_{D_1}}(D_1)}{M_4(f)} \right)^{1/(p+1)}}{\varepsilon} \right) \tag{3.41}$$

iterations with $D = D_1 + \max \{D_0, 2R_0\}$, $B_{D_1} = \{z \in \mathbb{E} : \|y_k - z\| \leq D_1\}$ and

$$D_1 = \left(\frac{4}{(1-\vartheta)M_4(f)} (\max \{F(x_0), F_S\} - F^*) \right)^{\frac{1}{4}}, \quad \bar{L} = 2M_4(f)D_1^2 + 2M_2(f).$$

(iii) For arbitrary p , the auxiliary problem (3.4) is approximately solved by Algorithm 3 in at most

$$m + 1 + \frac{\frac{6\bar{L}}{p-1} \beta^{1-p} \sigma_k^{-\frac{2}{p+1}} \bar{D}^{p-1} \bar{C}^{-\frac{p-1}{p+1}} (p+1)^{\frac{p+3}{p+1}}}{\varepsilon^{p-1}} \tag{3.42}$$

iterations, where $\varphi_k(z_i) - \varphi_k(z_k^*) < \sigma_k^{-\frac{2}{p-1}} (p+1)^{\frac{2}{p-1}} (3\bar{L})^{\frac{p+1}{p-1}}$ for $i \geq m$ and

$$\bar{D} = \left(\frac{(p+1)!2^{2p-2}}{pM_{p+1}(f)} (\max \{F(x_0), F_S\}) \right)^{\frac{1}{p+1}} + \max \{D_0, 2R_0\}.$$

Proof Assertion (i) follows directly from the inequality (2.23). To show (3.41) for $p = 3$, we apply the results of Theorem 2.8. From Algorithm 4, we obtain

$$y_k = \frac{A_{k-1}}{A_k} x_k + \frac{a_k}{A_k} v_k, \quad \|v_k - x^*\| \leq 2^{\frac{1}{2}} \|x_0 - x^*\|,$$

leading to

$$\begin{aligned} \|y_k - x^*\| &\leq \frac{A_k}{A_{k+1}} \|x_k - x^*\| + \frac{a_{k+1}}{A_{k+1}} \|v_k - x^*\| \\ &\leq \frac{A_k + a_{k+1}}{A_{k+1}} \max \{ \|x_k - x^*\|, \|v_k - x^*\| \} \\ &\leq \max \left\{ D_0, 2^{\frac{1}{2}} R_0 \right\} \\ &\leq \max \{D_0, 2R_0\}. \end{aligned} \tag{3.43}$$

Invoking Theorem 2.8(iii), it holds that

$$\|v_k - x^*\| \leq 2^{\frac{1}{2}} \|x_0 - x^*\| \leq 2R_0,$$

i.e., $v_k \in S$. Together with the convexity of $\psi(\cdot)$ and the monotonicity of the sequence $\{F(x_k)\}_{k \geq 0}$, this implies

$$F(y_k) \leq \frac{A_k}{A_{k+1}} F(x_k) + \frac{a_{k+1}}{A_{k+1}} F(v_k) \leq \frac{A_k}{A_{k+1}} F(x_0) + \frac{a_{k+1}}{A_{k+1}} F_S \leq \max \{F(x_0), F_S\}. \tag{3.44}$$

It follows from (3.6) that $F(z_i) \leq \varphi_k(z_i) \leq \varphi_k(z_{i-1}) \leq \dots \leq \varphi_k(y_k) = F(y_k)$. Setting $z = y_k$ in (3.7) and using and the latter inequality, we come to

$$\frac{(1-\vartheta)M_4(f)}{4} \|z_i - y_k\|^4 \leq F(y_k) - F(z_i) \leq \max \{F(x_0), F_S\} - F^*,$$

leading to $\|z_i - y_k\| \leq \left(\frac{4}{(1-\vartheta)M_4(f)} (\max \{F(x_0), F_S\} - F^*) \right)^{\frac{1}{4}} = D_1$. Hence, these inequalities lead to

$$\|z_i - x^*\| \leq \|z_i - y_k\| + \|y_k - x^*\| \leq D_1 + \max \{D_0, 2R_0\} = D.$$

For $p = 3$, it is clear that all conditions of Theorem 3.4 are satisfied. On the other hand, from the definition $\theta_{B_{D_1}}(\cdot)$ given in (3.31), we obtain

$$\theta_R(\|z_k^* - y_k\|) \leq \theta_{B_{D_1}}(D_1).$$

Then, from $L = \frac{3}{4}, \mu = \frac{1}{4}, \kappa = \frac{1}{3}$, the uniform convexity of $\rho_k(\cdot)$ with degree $\eta = 4$ and parameter $\sigma_{\rho_k} = \frac{1}{2} M_4(f), C = \frac{3M_4(f)}{2L^4}, (3.31), \bar{L} = 2M_4(f)D_1^2 + 2M_2(f)$ (see Lemma 3.11), and the proof of Theorem 3.4, we come to

$$i_k^* \leq 1 + \frac{4}{-\log\left(1-\frac{1}{6}\right)} \log \left(\frac{\frac{D}{\beta} \left(\frac{\bar{L}^4}{M_4(f) \theta_{B_{D_1}}(D_1)} \right)^{\frac{1}{4}}}{\varepsilon} \right),$$

which leads to (3.41) for $p = 3$.

By Lemma 3.8, $\rho_k(\cdot)$ is \widehat{L} -smooth on $B_{\widehat{\Delta}_k}$ with $\widehat{L} = M_4(f)\overline{\Delta}_k^2 + 2M_2(f) + \left(\frac{2M_{p+1}(f)}{(p-1)!} + \frac{3pH}{2} \right) \overline{\Delta}_k^{p-1}$. Hence, from the convexity of $(\rho_k + f_{y_k, p}^H)(\cdot)$, and the uniform convexity of $\varphi_k(\cdot)$ with degree $p + 1$ and parameter $\sigma_k = 2^{1-p}H$, we obtain

$$\begin{aligned}
 \varphi_k(z_{i+1}) &= \min_{z \in \mathbb{E}} \left\{ f_{y_k, p}^H(z_i) + \left\langle \nabla f_{y_k, p}^H(z_i), z - z_i \right\rangle + \psi(z) + 2\beta\rho_k(z_i, z) \right\} \\
 &\leq \min_{z \in \mathbb{E}} \left\{ f_{y_k, p}^H(z) + \beta\rho_k(z_i, z) + \psi(z) + 2\beta\rho_k(z_i, z) \right\} \\
 &\leq \min_{\alpha \in [0, 1]} \left\{ \varphi_k(z) + 3\beta\rho_k(z_i, z) : z = z_i + \alpha(z_k^* - z_i) \right\} \\
 &\leq \min_{\alpha \in [0, 1]} \left\{ \varphi_k(z) + \frac{3\bar{L}}{2} \|z - z_i\|^2 : z = z_i + \alpha(z_k^* - z_i) \right\} \\
 &\leq \min_{\alpha \in [0, 1]} \left\{ \varphi_k(z_i) - (\varphi_k(z_i) - \varphi_k(z_k^*))\alpha + \frac{3\bar{L}}{2} \|z_k^* - z_i\|^2 \alpha^2 \right\} \\
 &\leq \min_{\alpha \in [0, 1]} \left\{ \varphi_k(z_i) - (\varphi_k(z_i) - \varphi_k(z_k^*))\alpha + \frac{3\bar{L}}{2} \left[\frac{p+1}{\sigma_k} (\varphi_k(z_i) - \varphi_k(z_k^*)) \right]^{\frac{2}{p+1}} \alpha^2 \right\}.
 \end{aligned}$$

Minimizing the right hand-side of the last inequality with respect to α leads to the minimizer

$$\begin{aligned}
 \alpha^* &= \min \left\{ \frac{1}{3\bar{L}} \left(\frac{\sigma_k}{p+1} \right)^{\frac{2}{p+1}} (\varphi_k(z_i) - \varphi_k(z_k^*))^{\frac{p-1}{p+1}}, 1 \right\} \\
 &= \min \left\{ \frac{\sigma_k}{(p+1)(3\bar{L})^{\frac{p+1}{2}}} (\varphi_k(z_i) - \varphi_k(z_k^*))^{\frac{p-1}{2}}, 1 \right\}^{\frac{2}{p+1}} \\
 &= \min \left\{ \alpha_{k,i}, 1 \right\}^{\frac{2}{p+1}},
 \end{aligned}$$

where $\alpha_{k,i} = \sigma_k / [(p+1)(3\bar{L})^{\frac{p+1}{2}}] (\varphi_k(z_i) - \varphi_k(z_k^*))^{\frac{p-1}{2}}$. Let us consider two cases:

(i) $\alpha_{k,i} < 1$; (ii) $\alpha_{k,i} \geq 1$. In Case (i), we have $\alpha^* = \alpha_{k,i}^{\frac{2}{p+1}}$, i.e.,

$$\begin{aligned}
 \varphi_k(z_{i+1}) &\leq \varphi_k(z_i) - \left((\varphi_k(z_i) - \varphi_k(z_k^*)) - \frac{3\bar{L}}{2} \left[\frac{p+1}{\sigma_k} (\varphi_k(z_i) - \varphi_k(z_k^*)) \right]^{\frac{2}{p+1}} \alpha_{k,i}^{\frac{2}{p+1}} \right) \alpha_{k,i}^{\frac{2}{p+1}} \\
 &= \varphi_k(z_i) - \left((\varphi_k(z_i) - \varphi_k(z_k^*)) - \frac{1}{2} (\varphi_k(z_i) - \varphi_k(z_k^*)) \right) \alpha_{k,i}^{\frac{2}{p+1}} \\
 &= \varphi_k(z_i) - \frac{1}{2} \alpha_{k,i}^{\frac{2}{p+1}} (\varphi_k(z_i) - \varphi_k(z_k^*)).
 \end{aligned}$$

In Case (ii), we have $\alpha_{k,i} \geq 1$, i.e., $\alpha^* = 1$ resulting to

$$\begin{aligned}
 \varphi_k(z_{i+1}) &\leq \varphi_k(z_i) - \left((\varphi_k(z_i) - \varphi_k(z_k^*)) - \frac{3\bar{L}}{2} \left[\frac{p+1}{\sigma_k} (\varphi_k(z_i) - \varphi_k(z_k^*)) \right]^{\frac{2}{p+1}} \right) \\
 &= \varphi_k(z_i) - \left(1 - \frac{1}{2} \left[\left(\frac{(p+1)(3\bar{L})^{\frac{p+1}{2}}}{\sigma_k} \right) (\varphi_k(z_i) - \varphi_k(z_k^*))^{\frac{1-p}{2}} \right] \right) \\
 &\quad \times (\varphi_k(z_i) - \varphi_k(z_k^*)) \\
 &= \varphi_k(z_i) - \left(1 - \frac{1}{2} \alpha_{k,i}^{-1} \right) (\varphi_k(z_i) - \varphi_k(z_k^*)) \\
 &= \varphi_k(z_i) - \frac{1}{2} (\varphi_k(z_i) - \varphi_k(z_k^*)).
 \end{aligned}$$

Combining the both cases, we come to

$$\varphi_k(z_{i+1}) \leq \varphi_k(z_i) - \frac{1}{2} \min \{ \alpha_{k,i}, 1 \} \frac{2}{p+1} (\varphi_k(z_i) - \varphi_k(z_k^*)).$$

Since $\{\varphi_k(z_i)\}_{k \geq 0}$ is decreasing, there exists $m \in \mathbb{N}$ such that $\alpha_{k,i} < 1$ for all $i \geq m$, i.e., the latter inequality is translated to

$$\begin{aligned} \varphi_k(z_{i+1}) - \varphi_k(z_k^*) &\leq \varphi_k(z_i) - \varphi_k(z_k^*) \\ &- \left[\left(\frac{1}{6L} \right)^{\frac{p+1}{p-1}} \left(\frac{\sigma_k}{p+1} \right)^{\frac{2}{p-1}} (\varphi_k(z_i) - \varphi_k(z_k^*)) \right]^{\frac{p-1}{p+1}} (\varphi_k(z_i) - \varphi_k(z_k^*)), \end{aligned}$$

for $i \geq m$. Setting $\delta_k = \left(\frac{1}{6L} \right)^{\frac{p+1}{p-1}} \left(\frac{\sigma_k}{p+1} \right)^{\frac{2}{p-1}} (\varphi_k(z_i) - \varphi_k(z_k^*))$, it holds that

$$\delta_i - \delta_{i+1} \geq \delta_i^{\frac{2p}{p+1}},$$

i.e., the inequality (3.39) is satisfied with $\varrho = \frac{2p}{p+1}$. Together with Lemma 3.14, this implies

$$\varphi_k(z_i) - \varphi_k(z_k^*) \leq \sigma_k^{-\frac{2}{p-1}} (p-1)^{-\frac{p+1}{p-1}} (6L)^{\frac{p+1}{p-1}} (p+1)^{\frac{p+3}{p-1}} \left(\frac{1}{i-m} \right)^{\frac{p+1}{p-1}}. \tag{3.45}$$

On the other hand, following the proof of Lemma 3.3 and using 1-smoothness relative to $\rho_k(\cdot)$, (3.13), and the uniform convexity of ρ_k with degree $p + 1$ and parameter σ_{ρ_k} , we come to

$$\|\mathcal{G}_{i+1}\|_* \leq \|(\nabla^2 \rho_k - \nabla^2 f_{y_k,p}^H)(z)\| \|z_{i+1} - z_i\| \leq \bar{L} \left(\frac{p+1}{\sigma_{\rho_k}} \beta_{\rho_k}(z_i, z_{i+1}) \right)^{1/(p+1)},$$

Together with (3.6), this implies

$$\begin{aligned} \varphi_k(z_i) - \varphi_k(z_{i+1}) &\geq \beta_{\rho_k}(z_i, z_{i+1}) \geq \frac{\sigma_{\rho_k}}{(p+1)\bar{L}^{p+1}} \|\mathcal{G}_{i+1}\|_*^{p+1} = \bar{C} \|\mathcal{G}_{i+1}\|_*^{p+1}, \\ \bar{C} &= \frac{\sigma_{\rho_k}}{(p+1)\bar{L}^{p+1}}, \end{aligned}$$

which consequently leads to

$$\bar{C}^{-\frac{1}{p+1}} \left(\varphi_k(z_{i_k^*-1}) - \varphi_k(z_k^*) \right)^{\frac{1}{p+1}} \geq \bar{C}^{-\frac{1}{p+1}} \left(\varphi_k(z_{i_k^*-1}) - \varphi_k(z_{i_k^*}) \right)^{\frac{1}{p+1}} \geq \|\mathcal{G}_{i_k^*}\|_*,$$

with $\mathcal{G}_{i_k^*} = \nabla f_{y_k, p}^H(z_{i_k^*}) - \nabla f_{y_k, p}^H(z_{i_k^*-1}) + \nabla \rho_k(z_{i_k^*-1}) - \nabla \rho_k(z_{i_k^*}) \in \partial \varphi_k(z_{i_k^*})$. Moreover, combining (3.25), (3.43), and (3.44) yields

$$\begin{aligned} \|z_i - x^*\| &\leq \|z_i - y_k\| + \|y_k - x^*\| \\ &\leq \left(\frac{(p+1)!2^{2p-2}}{pM_{p+1}(f)} (\max\{F(x_0), F_S\})\right)^{\frac{1}{p+1}} + \max\{D_0, 2R_0\} = \bar{D}. \end{aligned}$$

It follows from (3.45) that

$$\begin{aligned} \|\mathcal{G}_{i_k^*}\|_* &\leq \bar{C}^{-\frac{1}{p+1}} \left(\varphi_k(z_{i_k^*-1}) - \varphi_k(z_{i_k^*})\right)^{\frac{1}{p+1}} \\ &\leq \sigma_k^{-\frac{2}{p^2-1}} (p-1)^{-\frac{1}{p-1}} \bar{C}^{-\frac{1}{p+1}} (6\bar{L})^{\frac{1}{p-1}} (p+1)^{\frac{p+3}{p^2-1}} \left(\frac{1}{i_k^* - m - 1}\right)^{\frac{1}{p-1}}. \end{aligned}$$

Then, the inequality

$$\sigma_k^{-\frac{2}{p^2-1}} (p-1)^{-\frac{1}{p-1}} \bar{C}^{-\frac{1}{p+1}} (6\bar{L})^{\frac{1}{p-1}} (p+1)^{\frac{p+3}{p^2-1}} \left(\frac{1}{i_k^* - m - 1}\right)^{\frac{1}{p-1}} \geq \frac{\beta\varepsilon}{\bar{D}}$$

gives (3.42). □

Let us fix $q \geq 1$. Then, the function $f_{y_k, p}^H(\cdot)$ is L -smooth relative to the scaling function $\rho_k(\cdot)$ (3.18), which is the same for both $p = 2q$ and $p = 2q + 1$. If p is even (i.e., $p = 2q$), then Algorithm 4 is a $2q$ -order method (requiring the $2q$ -order oracle) and attains the complexity of $\mathcal{O}(\varepsilon^{-1/(2q+1)})$, which worse than the optimal complexity $\mathcal{O}(\varepsilon^{-2/(6q+1)})$. On the other hand, if p is odd ($p = 2q + 1$), then Algorithm 4 is again a $2q$ -order method (requiring the $2q$ -order oracle) obtaining the complexity of $\mathcal{O}(\varepsilon^{-1/(2q+2)})$, which is also worse than the optimal complexity $\mathcal{O}(\varepsilon^{-2/(6q+1)})$ except for $p = 3$ that leads to the complexity $\mathcal{O}(\varepsilon^{-(1/4)})$ overpassing the optimal complexity bound of second-order methods, i.e., $\mathcal{O}(\varepsilon^{-(2/7)})$, as was known from in [29]. However, in the following example, we show that the complexity of our method can overpass the classical bounds for some structured class of problems. It is arguable that one may come up with algorithms with better complexity than (3.42) for finding the acceptable solution satisfying (2.5) for general p (e.g., with a more sophisticated methods or stronger assumptions); however, this out of scope of the current study and will be further investigated in our future work.

Example 3.16 Let us consider the vector $b \in \mathbb{R}^N$, the vectors $a_i \in \mathbb{R}^n$ and the univariate functions $f_i : \mathbb{R} \rightarrow \mathbb{R}$ that are four times continuously differentiable, for $i = 1, \dots, N$. Then, we define the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ as

$$f(x) = \sum_{i=1}^N f_i(\langle a_i, x \rangle - b_i). \tag{3.46}$$

We are interested to apply Algorithm 4 with $p = 4$ and $p = 5$ to the problem (2.1) with this function $f(\cdot)$. In case of $p = 5$, $q = \lfloor 5/2 \rfloor = 2$ and we need to handle the subproblem

$$z_{i+1} = \arg \min_{z \in \mathbb{E}} \left\{ \left\langle \nabla f_{y_k, H}^5(z_i), z - z_i \right\rangle + \psi(z) + 2L\beta_{\rho_k}(z_i, z) \right\},$$

with

$$\rho_k(x) = \left\langle \nabla^2 f(y_k)(x - y_k), x - y_k \right\rangle + \frac{1}{12} D^4 f(y_k)[x - y_k]^4 + \frac{3}{2} H d_6(x - y_k),$$

which readily implies that our method requires fourth-order oracle of $f_i(\cdot)$, for $i = 1, \dots, N$. Let us emphasize that Theorem 3.5 implies that the sacling function $\rho_k(\cdot)$ is convex, which is an interesting result even in one dimension and with $N = 1$, i.e.,

$$2f''(y_k) + f^{iv}(y_k)h^2 + \frac{15}{2}H|x - y_k|^4 \geq 0.$$

In the same way, for $p = 4$, we need fourth-order oracle of $f_i(\cdot)$, for $i = 1, \dots, N$. Moreover, Theorem 3.15 ensures that the sequence generated by Algorithm 4 attains the complexity $\mathcal{O}(\varepsilon^{-1/5})$ for $p = 4$ and $\mathcal{O}(\varepsilon^{-1/6})$ for $p = 5$, which are worse than the optimal complexity $\mathcal{O}(\varepsilon^{-2/13})$, for the accuracy parameter ε . On the other hand, setting $h = x - y_k$, it holds that

$$\begin{aligned} \left\langle \nabla^2 f(y_k)h, h \right\rangle &= \sum_{i=1}^N \nabla^2 f_i(\langle a_i, y_k \rangle - b_i) \langle a_i, h \rangle^2, \\ D^4 f(y_k)[h]^4 &= \sum_{i=1}^N \nabla^4 f_i(\langle a_i, y_k \rangle - b_i) \langle a_i, h \rangle^4. \end{aligned}$$

Let us particularly verify these terms for $f_i(x) = -\log(x)$ ($i = 1, \dots, N$) for $x \in (0, +\infty)$, which consequently leads to

$$\nabla^2 f_i(x) = \frac{1}{x^2}, \quad \nabla^4 f_i(x) = \frac{6}{x^4} = 6 \left(\nabla^2 f_i(x) \right)^2,$$

i.e.,

$$D^4 f(y_k)[h]^4 = 6 \sum_{i=1}^N \left(\nabla^2 f_i(\langle a_i, y_k \rangle - b_i) \right)^2 \langle a_i, h \rangle^4.$$

Thus, in this case, the implementation of Algorithm 4 with $p = 4$ and $p = 5$ only requires the second-order oracle of $f_i(\cdot)$ ($i = 1, \dots, N$) and the first-order oracle of $\psi(\cdot)$. Therefore, we end up with a second-order method with the complexity of $\mathcal{O}(\varepsilon^{-1/5})$ for $p = 4$ and $\mathcal{O}(\varepsilon^{-1/6})$ for $p = 5$, which are much faster than the second-order methods optimal bound $\mathcal{O}(\varepsilon^{-2/7})$; however, choosing the odd order $p = 5$, Algorithm 4 attains the better complexity. \square

4 Conclusion

In this paper, we suggest a bi-level optimization (BiOPT), a novel framework for solving convex composite minimization problems, which is a generalization of the BLUM framework given in [29] and involves two levels of methodologies. In the upper level, we only assume the convexity of the objective function and design some upper-level scheme using a high-order proximal-point iterations with arbitrary order. On the other hand, in the lower level, we need to solve the proximal-point auxiliary problem inexactly by some lower-level scheme. In this step, we require some more properties of the objective function for developing efficient algorithms providing acceptable solutions for this auxiliary problem at a reasonable computational cost. The overall complexity of the method will be the product of the complexities in both levels.

We here develop the basic p th-order inexact proximal-point method and its acceleration using the estimation sequence technique that, respectively, achieve the convergence rate $\mathcal{O}(k^{-p})$ and $\mathcal{O}(k^{-(p+1)})$ for the iteration counter k . Assuming the L -smoothness and μ -strong convexity of the differentiable part of the proximal-point objective relative to some scaling function (for $L, \mu > 0$), we design a non-Euclidean composite gradient method to inexactly solve the proximal-point problem. It turns out that this method attains the complexity $\mathcal{O}(\log \frac{1}{\varepsilon})$, for the accuracy parameter $\varepsilon > 0$.

In the BiOPT framework, we apply the accelerated p th-order proximal-point algorithm in the upper level, introduce a new high-order scaling function and show that the differentiable part of the auxiliary objective is smooth relative to this function, and solve the auxiliary problem by a non-Euclidean composite gradient method in the lower level with the complexity of $\mathcal{O}(\log \frac{1}{\varepsilon})$. We consequently come to a bi-level high-order method with the complexity of $\mathcal{O}(\varepsilon^{-1/(p+1)})$, which overpasses the classical complexity bound of second-order methods for $p = 3$, as was known from [29]. In general, for $p = 2$ and $p \geq 3$, the complexity of our bi-level method is sub-optimal; however, we showed that for some class of structured problems it can overpass the optimal complexity bound $\mathcal{O}(\varepsilon^{-2/(3p+1)})$. Overall, the BiOPT framework paves the way toward methodologies using the p th-order proximal-point operator in the upper level and requiring lower-order oracle than p in the lower level. Therefore, owing to this framework, we can design lower-order methods with convergence rates overpassing the classical complexity bounds for convex composite problems. Hence, this will open up an entirely new ground for developing novel efficient algorithms for convex composite optimization that was not possible in the classical complexity theory.

Several extensions of our framework are possible. As an example, we will present some extension of our framework using a segment search in the upcoming article [3]. Moreover, the proximal-point auxiliary problem can be solved by some more efficient method such as the non-Euclidean line search method [5] with second-order directions [4, 6, 34]. In addition, the introduced high-order scaling function can be employed to extend the second-order methods presented in [27–30] to higher-order methods.

Acknowledgements The authors are also grateful to the anonymous referees for their helpful comments and suggestions that improved the quality of the paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give

appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Agarwal, N., Hazan, E.: Lower bounds for higher-order convex optimization. In: Conference On Learning Theory, pp. 774–792 (2018)
2. Ahoosh, M.: Accelerated first-order methods for large-scale convex optimization: nearly optimal complexity under strong convexity. *Math. Methods Oper. Res.* **89**(3), 319–353 (2019)
3. Ahoosh, M., Nesterov, Y.: High-order methods beyond the classical complexity bounds: inexact high-order proximal-point methods with segment search. Technical report, University of Antwerp (2021)
4. Ahoosh, M., Artacho, F.J.A., Fleming, R.M., Vuong, P.T.: Local convergence of the Levenberg–Marquardt method under Hölder metric subregularity. *Adv. Comput. Math.* 1–36 (2019)
5. Ahoosh, M., Themelis, A., Patrinos, P.: A Bregman forward–backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. *SIAM J. Optim.* **31**(1), 653–685 (2021)
6. Ahoosh, M., Fleming, R.M., Vuong, P.T.: Finding zeros of Hölder metrically subregular mappings via globally convergent Levenberg–Marquardt methods. *Optim. Methods Softw.* **37**, 113–149 (2022)
7. Arjevani, Y., Shamir, O., Shiff, R.: Oracle complexity of second-order methods for smooth convex optimization. *Math. Program.* **178**(1–2), 327–360 (2019)
8. Baes, M.: Estimate sequence methods: extensions and approximations. Institute for Operations Research, ETH, Zürich, Switzerland (2009)
9. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.* **42**(2), 330–348 (2016)
10. Birgin, E.G., Gardenghi, J., Martínez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Math. Program.* **163**(1–2), 359–368 (2017)
11. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.* **28**(3), 2131–2151 (2018)
12. Gasnikov, A., Dvurechensky, P., Gorbunov, E., Vorontsova, E., Selikhanovych, D., Uribe, C.: Optimal tensor methods in smooth convex and uniformly convex optimization. In: Proceedings of the Thirty-Second Conference on Learning Theory, pp. 1374–1391 (2019)
13. Grapiglia, G.N., Nesterov, Y.: Regularized newton methods for minimizing functions with hölder continuous Hessians. *SIAM J. Optim.* **27**(1), 478–506 (2017)
14. Grapiglia, G.N., Nesterov, Y.: On inexact solution of auxiliary problems in tensor methods for convex optimization. *Optim. Methods Softw.* 1–26 (2020)
15. Güler, O.: New proximal point algorithms for convex minimization. *SIAM J. Optim.* **2**(4), 649–664 (1992)
16. Iusem, A.N., Svaiter, B.F., Teboulle, M.: Entropy-like proximal methods in convex programming. *Math. Oper. Res.* **19**(4), 790–814 (1994)
17. Jiang, B., Wang, H., Zhang, S.: An optimal high-order tensor method for convex optimization. In: Conference on Learning Theory, pp. 1799–1801 (2019)
18. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.* **28**(1), 333–354 (2018)
19. Martinet, B.: Brève communication. Régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge* **4**(R3), 154–158 (1970)
20. Martinet, B.: Détermination approchée d'un point fixe d'une application pseudo-contractante. *CR Acad. Sci. Paris* **274**(2), 163–165 (1972)

21. Nemirovsky, A., Yudin, D.: *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, New York (1983)
22. Nesterov, Y.: Smooth minimization of non-smooth functions. *Math. Program.* **103**(1), 127–152 (2005)
23. Nesterov, Y.: Accelerating the cubic regularization of newton's method on convex problems. *Math. Program.* **112**(1), 159–181 (2008)
24. Nesterov, Y.: Gradient methods for minimizing composite functions. *Math. Program.* **140**(1), 125–161 (2013)
25. Nesterov, Y.: Universal gradient methods for convex optimization problems. *Math. Program.* **152**(1–2), 381–404 (2015)
26. Nesterov, Y.: *Lectures on Convex Optimization*, vol. 137. Springer, Berlin (2018)
27. Nesterov, Y.: Implementable tensor methods in unconstrained convex optimization. *Math. Program.* 1–27 (2019)
28. Nesterov, Y.: Inexact basic tensor methods for some classes of convex optimization problems. *Optim. Methods Softw.* 1–29 (2020)
29. Nesterov, Y.: Inexact accelerated high-order proximal-point methods. *Math. Program.* 1–26 (2021)
30. Nesterov, Y.: Superfast second-order methods for unconstrained convex optimization. *J. Optim. Theory Appl.* **191**(1), 1–30 (2021)
31. Nesterov, Y., Nemirovskii, A.: *Interior-Point Polynomial Algorithms in Convex Programming*, vol. 13. SIAM (1994)
32. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14**(5), 877–898 (1976)
33. Teboulle, M.: Entropic proximal mappings with applications to nonlinear programming. *Math. Oper. Res.* **17**(3), 670–690 (1992)
34. Themelis, A., Ahoosh, M., Patrinos, P.: On the acceleration of forward-backward splitting via an inexact Newton method. In: Luke, R., Bauschke, H., Burachik, R. (eds.) *Splitting Algorithms, Modern Operator Theory, and Applications*, pp. 363–412. Springer, Berlin (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.