ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

Nearly optimal first-order methods for convex optimization under gradient norm measure: An adaptive regularization approach

Masaru Ito and Mituhiro Fukuda

February 2020, B–492

Department of Mathematical and Computing Sciences
Tokyo Institute of Technology

SERIES B: Operations Research

Nearly optimal first-order methods for convex optimization under gradient norm measure: An adaptive regularization approach

Masaru Ito* Mituhiro Fukuda†

February 2020

Abstract

In the development of first-order methods for smooth (resp., composite) convex optimization problems minimizing L-smooth functions, the gradient (resp., gradient mapping) norm is a fundamental optimality measure for which the best known iteration complexity to obtain an ε-solution is $O(\sqrt{LD/\varepsilon}\log(1/\varepsilon))$ for the distance D from the initial point to the optimal solution set. In this paper, we report an adaptive regularization approach attaining this iteration complexity without the prior knowledge of D which was required to be known in the existing regularization approach. To obtain further faster convergence adaptively, we secondly apply this approach to construct a first-order method that is adaptive to the Hölderian error bound condition (or equivalently, the Łojasiewicz gradient property) which covers moderately wide class of applications. The proposed method attains nearly optimal iteration complexity with respect to the gradient mapping norm.

Keywords: smooth/composite convex optimization; accelerated proximal gradient methods; Hölderian error bound; adaptive methods

Mathematical Subject Classification (2010): 90C25; 68Q25; 49M37

1 Introduction

The class of proximal gradient methods is a fundamental tool for solving the composite convex optimization problem

$$\varphi^* = \min_x [\varphi(x) \equiv f(x) + \Psi(x)], \tag{1}$$

where f is an L-smooth convex function defined on a Euclidean space, i.e., f is a continuously differentiable convex function with L-Lipschitz continuous gradient, and Ψ is a proper, lower-semicontinuous convex function. Accelerated first-order methods for this class of problems have been well-studied as they provide optimal iteration complexity to obtain an ε-approximate solution under the measure $\varphi(\cdot) - \varphi^*$ for various classes of problems [2, 14, 17, 18, 20, 22].

A major interest focused on this paper is to consider the iteration complexity to obtain an ε-approximate solution with respect to the *gradient mapping norm*. The gradient mapping $g_L(x)$ is defined by

$$g_L(x) = L(x - \mathrm{prox}_{\Psi/L}(x - L^{-1}\nabla f(x))), \quad \text{where } \mathrm{prox}_{\Psi/L}(y) = \operatorname*{argmin}_x \left[\Psi(x) + \frac{L}{2}\|x - y\|^2\right],$$

*Department of Mathematics, College of Science and Technology, Nihon University, 1-8-14 Kanda-Surugadai, Chiyoda, Tokyo 101-8308, Japan (ito.m@math.cst.nihon-u.ac.jp).

†Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1-W8-41 Oh-okayama, Meguro, Tokyo 152-8552, Japan (mituhiro@is.titech.ac.jp).

and $g_L(x) = 0$ holds if and only if x is optimal to (1). The norm $\|g_L(x)\|$ of the gradient mapping is a useful optimality measure as it is computable at each iteration (if prox_Ψ is computable) in contrast to the measure $\varphi(\cdot) - \varphi^*$ that is not verifiable if we do not know φ^*. Note that, for the smooth problems (i.e., the case $\Psi \equiv 0$), the gradient mapping $g_L(x)$ coincides with the gradient $\nabla f(x)$.

To find an approximate solution x such that $\|g_L(x)\| \leq \varepsilon$, the best known iteration complexity of first-order methods for the problem (1) is

$$O\left(\sqrt{\frac{LD}{\varepsilon}} \log \frac{LD}{\varepsilon}\right), \tag{2}$$

where $D := \mathrm{dist}(x_0, X^*)$ for an initial point x_0 and the optimal solution set X^*. This complexity is attained by a regularization technique [16] but it requires D to be known in advance. Without the prior knowledge of D, accelerated gradient methods achieving the iteration complexity $O((LD/\varepsilon)^{2/3})$ seems to be the best known ones [6, 16, 23]. One aim of this paper is to report an adaptive algorithm (Algorithm 3) for the regularization technique that attains the iteration complexity (2) without the requirement of D, as shown in Corollary 4.3.

Another motivation is the development of an adaptive first-order method under the Hölderian Error Bound (HEB) condition, that is, given an initial point x_0, we assume that

$$\varphi(x) - \varphi^* \geq \kappa\, \mathrm{dist}(x, X^*)^\rho, \quad \forall x \text{ such that } \varphi(x) \leq \varphi(x_0), \tag{3}$$

for some $\kappa > 0$ and $\rho \geq 1$. This condition is also related to the concept called the Łojasiewicz gradient inequality [10, 11] which is a useful tool for the development and the analysis of algorithms as well as first-order methods [1, 4]. These concepts are known to be satisfied under moderately mild assumptions such as when φ is coercive and subanalytic (in particular, semi-algebraic) [3]. The coefficient κ and the exponent ρ are critical parameters to determine the convergence rate but they are difficult to estimate in general, so that the development of adaptive algorithms is an important issue.

For the problem (1) under the Hölderian error bound condition, we propose Algorithm 4, a restart scheme of the previous mentioned adaptive regularization algorithm. Our method is inspired by Liu-Yang's method [8] as we employ an (approximated) proximal-point approach, where the main difference is the adaption parameter: Liu-Yang's method adaptively estimate the coefficient κ while our method adaptively determine the regularization parameter σ to define the regularized auxiliary problem

$$\min_x \left[\varphi(x) + \frac{\sigma}{2}\|x - x_0\|^2\right].$$

As a result, without knowing the coefficient κ and the exponent ρ, the proposed method adapts both the parameters κ and ρ. To find an approximate solution x such that $\|g_L(x)\| \leq \varepsilon$, we prove the following iteration complexity result (Corollary 4.5):

- Case $\rho = 1$. The algorithm finds an optimal solution with a finite iteration complexity, i.e., the iteration complexity is independent of ε (if ε is sufficiently small).
- Case $\rho \in (1, 2)$. The iteration complexity is $O(\log\log \varepsilon^{-1})$ (superlinear convergence).
- Case $\rho = 2$. The iteration complexity is $O(\log \varepsilon^{-1})$ (linear convergence).
- Case $\rho > 2$. The iteration complexity is $O(\varepsilon^{-\frac{\rho-2}{2(\rho-1)}} \log \varepsilon^{-1})$ (sublinear convergence).

The finite and the superlinear convergences in $\rho = 1$ and $\rho \in (1, 2)$, respectively, are due to accurate convergence analysis, which were not shown in the existing adaptive methods [8, 20]. Moreover, for

the smooth problems (i.e., the case $\Psi \equiv 0$), we show that the proposed method attains the nearly optimal iteration complexity with respect to the gradient norm. We can also immediately deduce the nearly optimal iteration complexity result with respect to the measure $\varphi(\cdot) - \varphi^*$.

Table 1 shows the relation to existing adaptive first-order methods. All the algorithms in this table attain the nearly optimal iteration complexity with respect to the employed measure. The recent first-order methods [20, 22] are applicable to our problem (for specific Ψ) and they adapt both κ and ρ ensuring the nearly optimal iteration complexity with respect to the measure $\varphi(\cdot) - \varphi^*$. One advantage of our method compared with these methods is the (nearly optimal) *convergence* guarantee; the method of Roulet and d'Aspremont (see [22, Proposition 3.4]) is a fixed step algorithm (remark that, if we know φ^*, [22, Algorithm 3] gives nearly optimal convergence), and Renegar-Grimmer's method [20] fixes the target tolerance ε. Although the other algorithms, i.e., this work and the first four algorithms [5, 7, 8, 17] in Table 1, terminate if an ε-solution is found, they provide the nearly optimal convergence letting $\varepsilon = 0$ as we discuss later.

Table 1: Adaptive first-order methods. The column 'Parameters' indicates the parameters that the algorithm can adapt. The column 'Optimality measures' is the optimality measure for which the nearly optimal iteration complexity was proved. The column 'Convergence' shows whether the algorithm ensures the convergence of the optimality measure to zero (at the nearly optimal rate).

Algorithm	Problem class	Parameters	Optimality measures	Convergence
Nesterov [17] Lin and Xiao [7]	μ-strongly convex φ	μ	gradient mapping norm	yes
Fercoq and Qu [5]	HEB (3) with $\rho = 2$	κ	gradient mapping norm	yes
Liu and Yang [8]	HEB (3) with known ρ	κ	gradient mapping norm	yes
Roulet and d'Aspremont [22, Proposition 3.4]	HEB (3)	κ and ρ	$\varphi(\cdot) - \varphi^*$	no (fixed iteration)
Reneger and Grimmer [20]	HEB (3)	κ and ρ	$\varphi(\cdot) - \varphi^*$	no (fixed tolerance)
This work (Algorithm 4)	HEB (3)	κ and ρ	gradient mapping norm (and $\varphi(\cdot) - \varphi^*$)	yes

This paper is organized as follows. Section 2 collects preliminary facts on the gradient mapping and the Hölderian error bound condition. In particular, in Section 2.2, we deduce a lower iteration complexity bound with respect to the gradient norm for the class of smooth convex functions satisfying the Hölderian error bound condition. We review in Section 3 the regularization technique [16] preparing auxiliary results. In Section 4, we propose adaptive first-order methods and show their iteration complexity results. We show an adaptive regularization algorithm in Section 4.1 and prove the iteration complexity (2). A restart scheme of this algorithm is given in Section 4.2 and we show that it adaptively attains the nearly optimal iteration complexity under the Hölderian error bound condition. Concluding remarks are given in Section 5.

2 Preliminaries

Throughout this paper, let $\mathbb{E}$ be a finite dimensional real Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle$. We denote by $\|x\| = \langle x, x \rangle^{1/2}$ the induced norm on $\mathbb{E}$.

A convex function $f : \mathbb{E} \to \mathbb{R}$ is called *L-smooth* for $L > 0$ if f is continuously differentiable and its gradient is L-Lipschitz continuous on $\mathbb{E}$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{E}.$$

The following inequality is a fundamental property of L-smooth functions (e.g., see [15]):

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{E}. \tag{4}$$

We focus on the the convex composite optimization problem

$$\varphi^* = \min_x [\varphi(x) \equiv f(x) + \Psi(x)] \tag{5}$$

for an L_f-smooth convex function $f : \mathbb{E} \to \mathbb{R}$ and a proper lower-semicontinuous convex function $\Psi : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$. We denote by X^* the set of optimal solutions of $\min_{x \in \mathbb{E}} \varphi(x)$:

$$X^* = \operatorname*{Argmin}_{x \in \mathbb{E}} \varphi(x).$$

The subdifferential of φ at x is denoted by $\partial\varphi(x) = \{g \in \mathbb{E} \mid \varphi(y) \geq \varphi(x) + \langle g, y - x \rangle, \ \forall y \in \mathbb{E}\}$.

For the objective function $\varphi(x) = f(x) + \Psi(x)$, we define

$$m_L(y; x) := f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + \Psi(x),$$
$$T_L(y) := \operatorname*{argmin}_{x \in \mathbb{E}} m_L(y; x) \quad (= \operatorname{prox}_{\Psi/L}(y - L^{-1} \nabla f(y))),$$

where the minimizer $T_L(y)$ is well-defined since $m_L(y; \cdot)$ is strongly convex. It is assumed that $\Psi(\cdot)$ has a "simple" structure, namely, $T_L(\cdot)$ is moderately computable (see [19] for examples). The *gradient mapping* of φ is defined by

$$g_L(y) := L(y - T_L(y)), \quad y \in \mathbb{E}, \ L > 0.$$

For instance, if $\Psi \equiv 0$, we see that $T_L(y) = y - \nabla f(y)/L$ and $g_L(y) = \nabla f(y)$ hold.

Remark that the norm of $g_L(y)$ is given by

$$\|g_L(y)\| = L \|y - T_L(y)\|,$$

from which the quantity $\|g_L(y)\|$ can be used as a computable optimality measure at y (see Lemma 2.1 (ii) below).

The following lemma collects basic properties on $T_L(x)$ and $g_L(x)$ which can be found in [2, 15, 17].

Lemma 2.1. *Let $\varphi = f + \Psi$ be the sum of a continuously differentiable convex function $f : \mathbb{E} \to \mathbb{R}$ and a proper lower-semicontinuous convex function $\Psi : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$. Then, the following assertions hold.*

(i) For all $y \in \mathbb{E}$, the map $L \mapsto \|g_L(y)\|$ is increasing.

(ii) $x^* \in X^*$ holds if and only if $g_L(x^*) = 0$.

(iii) $\nabla f(T_L(y)) - \nabla f(y) + g_L(y) \in \partial\varphi(T_L(y))$ for all $y \in \mathbb{E}$. Moreover, if f is L_f-smooth, then

$$\|\nabla f(T_L(y)) - \nabla f(y) + g_L(y)\| \leq \left(\frac{L_f}{L} + 1\right) \|g_L(y)\|.$$

(iv) If $y \in \mathbb{E}$ and $L > 0$ satisfy $\varphi(T_L(y)) \leq m_L(y; T_L(y))$, then we have

$$\frac{1}{2L} \|g_L(y)\|^2 \leq \varphi(y) - \varphi(T_L(y)) \leq \varphi(y) - \varphi^*,$$

$$\frac{1}{2L} \|g_L(y)\| \leq \operatorname{dist}(y, X^*).$$

(v) If f is L_f-smooth, then $\varphi(T_L(y)) \le m_L(y; T_L(y))$ holds for all $y \in \mathbb{E}$ and $L \ge L_f$.

Proof. (i) is proved in [17, Lemma 2].

(ii) The optimality condition of the problem $\min_{x\in\mathbb{E}} m_L(y;x)$ is given as follows:

$$z = T_L(y) \iff 0 \in \nabla f(y) + L(z-y) + \partial\Psi(z). \tag{6}$$

On the other hand, the optimality of the original problem $\min_x \varphi(x)$ is characterized by

$$z \in \operatorname*{Argmin}_{x\in\mathbb{E}} \varphi(x) \iff 0 \in \nabla f(z) + \partial\Psi(z).$$

(Remark that $\partial\varphi = \nabla f + \partial\Psi$ holds [21, Theorem 23.8]). Plugging $y = z = x^*$, we see that the equivalence $x^* \in \operatorname{Argmin}_{x\in\mathbb{E}} \varphi(x) \iff x^* = T_L(x^*)$ ($\iff g_L(x^*) = 0$) follows.

(iii) By the optimality condition (6), we have

$$0 \in \nabla f(y) + L(T_L(y) - y) + \partial\Psi(T_L(y)) = \nabla f(y) - g_L(y) + \partial\Psi(T_L(y)). \tag{7}$$

Hence,

$$\nabla f(T_L(y)) - \nabla f(y) + g_L(y) \in \nabla f(T_L(y)) + \partial\Psi(T_L(y)) = \partial\varphi(T_L(y)).$$

Now if f is L_f-smooth, we have

$$\|\nabla f(T_L(y)) - \nabla f(y)\| \le L_f \|T_L(y) - y\| = \frac{L_f}{L} \|g_L(y)\|,$$

which yields the latter assertion of (iii).

(iv) It is shown in [2, Lemma 2.3] that, if $\varphi(T_L(y)) \le m_L(y; T_L(y))$ holds, then we have

$$\varphi(x) - \varphi(T_L(y)) \ge \frac{1}{2L} \|g_L(y)\|^2 + \langle g_L(y), x - y\rangle, \quad \forall x \in \mathbb{E}.$$

Letting $x := y$ shows the first assertion. On the other hand, since $\varphi(T_L(y)) \ge \varphi^*$, letting $x := x^* \in X^*$ gives

$$\frac{1}{2L} \|g_L(y)\|^2 \le \langle g_L(y), y - x^*\rangle \le \|g_L(y)\| \|y - x^*\|, \quad \forall x^* \in X^*.$$

Thus, we obtain $\frac{1}{2L}\|g_L(y)\| \le \operatorname{dist}(y, X^*)$.

(v) Since f is L-smooth for any $L \ge L_f$, the inequality (4) implies $\varphi(x) \le m_L(y;x)$ for all $x, y \in \mathbb{E}$ and $L \ge L_f$. □

2.1 Hölderian error bound

Here we introduce the Hölderian error bound condition which is also discussed in the context of Łojasiewicz inequality [3, 9, 11].

Definition 2.2. Fix $x_0 \in \mathbb{E}$. We say that φ satisfies the *Hölderian error bound condition* with coefficient $\kappa > 0$ and exponent $\rho \ge 1$ if

$$\varphi(x) - \varphi^* \ge \kappa \operatorname{dist}(x, X^*)^\rho, \quad \forall x \in \operatorname{lev}_\varphi(\varphi(x_0)), \tag{8}$$

where $\operatorname{lev}_\varphi(\gamma) = \{x \in \mathbb{E} \mid \varphi(x) \le \gamma\}$ and $\operatorname{dist}(x, X^*) = \inf_{x^*\in X^*} \|x - x^*\|$.

According to [3, Theorem 3.3], the Hölderian error bound condition is satisfied for some κ and ρ if $\varphi(x)$ is a proper, lower-semicontinuous, convex, coercive and subanalytic function. As subanalytic functions contain semi-algebraic ones, this condition appears in many applications including popular problems in machine learning; see, e.g., [4, 8] for related studies.

A noteworthy concept related to the Hölderian error bound condition is the *Łojasiewicz gradient inequality* [10, 11], which is of the form

$$\operatorname{dist}(0, \partial\varphi(x)) \geq \lambda(\varphi(x) - \varphi^*)^\theta, \quad \forall x \in \operatorname{lev}_\varphi(\varphi(x_0)) \setminus X^* \tag{9}$$

for $\lambda > 0$ and $\theta \in [0, 1)$. In fact, these concepts (8) and (9) are equivalent for convex functions (see Remark 2.4). The Łojasiewicz gradient inequality is a powerful tool for analyzing the convergence of first-order methods as it covers wide class of applications and algorithms [1, 4].

Lemma 2.3. *Let $\varphi : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous convex function. For $x_0 \in \mathbb{E}$, suppose that φ satisfies the Hölderian error bound condition (8) with coefficient $\kappa > 0$ and exponent $\rho \geq 1$. Then, for every $x \in \operatorname{lev}_\varphi(\varphi(x_0)) \setminus X^*$, we have*

$$\begin{aligned} \kappa \operatorname{dist}(x, X^*)^{\rho-1} &\leq \inf\{\|g\| : g \in \partial\varphi(x)\}, \\ \kappa^{\frac{1}{\rho}} (\varphi(x) - \varphi^*)^{\frac{\rho-1}{\rho}} &\leq \inf\{\|g\| : g \in \partial\varphi(x)\}. \end{aligned} \tag{10}$$

Proof. Let x^* be the projection of x onto X^*, so that $\|x - x^*\| = \operatorname{dist}(x, X^*)$. For every $g \in \partial\varphi(x)$, we have

$$\kappa \operatorname{dist}(x, X^*)^\rho \leq \varphi(x) - \varphi^* \leq -\langle g, x^* - x\rangle \leq \|g\| \operatorname{dist}(x, X^*) \leq \|g\| \frac{1}{\kappa^{1/\rho}} (\varphi(x) - \varphi^*)^{1/\rho}.$$

In particular, we obtain two inequalities

$$\begin{aligned} \kappa \operatorname{dist}(x, X^*)^\rho &\leq \|g\| \operatorname{dist}(x, X^*), \\ \varphi(x) - \varphi^* &\leq \|g\| \frac{1}{\kappa^{1/\rho}} (\varphi(x) - \varphi^*)^{1/\rho}. \end{aligned}$$

Then, since $x \notin X^*$, the assertion follows. □

Remark 2.4. The condition (10) is the Łojasiewicz gradient inequality (9) with the correspondence $\lambda = \kappa^{1/\rho}$ and $\theta = (\rho - 1)/\rho \in [0, 1)$. It is shown in [4] that the Łojasiewicz gradient inequality is essentially equivalent to the Hölderian error bound condition: If (9) holds with $\lambda = \rho\kappa^{1/\rho}$ and $\theta = (\rho - 1)/\rho$, then (8) holds (In [4, Theorem 5 (i)] set $(\kappa^{-1/\rho} s^{1/\rho}, \varphi(x_0))$ in place of $(\varphi(s), r_0)$ and let the radius ρ of $B(\bar{x}, \rho)$ to $+\infty$). □

2.2 Lower complexity bounds

Let us discuss lower bounds on the iteration complexity under the Hölderian error bound condition with respect to the optimality measure $\|g_L(x)\|$. The lower bound is derived in the case $\Psi \equiv 0$ so that $\varphi = f$ is a smooth function and we have $g_L(x) = \nabla\varphi(x)$.

Given a class $\mathcal{F}$ of objective functions and an optimality measure $\delta : \mathcal{F} \times \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, the *iteration complexity* of a first-order method $\mathcal{M}$ applied to $\varphi \in \mathcal{F}$ for an accuracy $\varepsilon > 0$, say $C(\mathcal{M}, \varphi, \delta; \varepsilon)$, is defined as the minimal number of evaluations of a first-order oracle $(\varphi(\cdot), \nabla\varphi(\cdot))$ in the method $\mathcal{M}$ required to find a point $x \in \mathbb{E}$ satisfying $\delta(\varphi, x) \leq \varepsilon$. Then we define the iteration complexity of first-order methods associated with the class $\mathcal{F}$ with respect to the measure δ by

$$C(\mathcal{F}, \delta; \varepsilon) := \inf_{\mathcal{M}} \sup_{\varphi \in \mathcal{F}} C(\mathcal{M}, \varphi, \delta; \varepsilon),$$

where $\mathcal{M}$ runs all first-order methods for the class $\mathcal{F}$ starting from some fixed initial point $x_0 \in \mathbb{E}$.

We are interested in the iteration complexity under the following classes and measures:

- $\mathcal{F}(x_0, R, L)$ denotes the class of L-smooth convex functions φ with $X^* \neq \emptyset$ and $\operatorname{dist}(x_0, X^*) \leq R$, where $X^* = \operatorname{Argmin}_{x \in \mathbb{E}} \varphi(x)$.
- Class $\mathcal{F}(x_0, R, L, \kappa, \rho)$: For $R, L, \kappa > 0$, $\rho \geq 2$, and $x_0 \in \mathbb{E}$, we say that φ belongs to the class $\mathcal{F}(x_0, R, L, \kappa, \rho)$ if $\varphi \in \mathcal{F}(x_0, R, L)$ and it satisfies the Hölderian error bound condition

$$\varphi(x) - \varphi^* \geq \kappa \operatorname{dist}(x, X^*)^\rho, \quad \forall x \in \operatorname{lev}_\varphi(\varphi(x_0)),$$

  where $X^* = \operatorname{Argmin}_{x \in \mathbb{E}} \varphi(x)$.

  Remark that we do not consider the case $\rho \in [1, 2)$ because any L-smooth convex function cannot admit the Hölderian error bound condition with exponent $\rho \in [1, 2)$.[1]

- We consider the optimality measures

$$\delta^*(\varphi, x) := \varphi(x) - \inf \varphi,$$
$$\delta(\varphi, x) := \|\nabla \varphi(x)\|.$$

For the class $\mathcal{F}(x_0, R, L)$, the following lower bound on the iteration complexity holds (by [12, Section 2.3B] applied to the class of L-smooth convex quadratic minimization):

$$C(\mathcal{F}(x_0, R, L), \delta; \varepsilon) = \Omega\left(\min\left\{\dim \mathbb{E}, \sqrt{\frac{LR}{\varepsilon}}\right\}\right). \tag{11}$$

Let us observe a lower bound on the iteration complexity for the class $\mathcal{F}(x_0, R, L, \kappa, \rho)$.

Proposition 2.5. *For the class $\mathcal{F} = \mathcal{F}(x_0, R, L, \kappa, \rho)$, we have*

$$C(\mathcal{F}, \delta; \varepsilon) \geq C(\mathcal{F}, \delta^*; \varepsilon^*), \quad \text{where} \quad \varepsilon^* := \left(\frac{1}{\kappa}\right)^{\frac{1}{\rho-1}} \varepsilon^{\frac{\rho}{\rho-1}}.$$

Proof. If a first-order method $\mathcal{M}$ applied to $\varphi \in \mathcal{F}$ finds an approximate solution $x \in \mathbb{E}$ satisfying $\delta(\varphi, x) \leq \varepsilon$, then Lemma 2.3 implies

$$\delta^*(\varphi, x) = \varphi(x) - \varphi^* \leq \left(\frac{1}{\kappa}\right)^{\frac{1}{\rho-1}} \|\nabla \varphi(x)\|^{\frac{\rho}{\rho-1}} = \left(\frac{1}{\kappa}\right)^{\frac{1}{\rho-1}} \delta(\varphi, x)^{\frac{\rho}{\rho-1}} \leq \left(\frac{1}{\kappa}\right)^{\frac{1}{\rho-1}} \varepsilon^{\frac{\rho}{\rho-1}} = \varepsilon^*.$$

Therefore, it follows that

$$C(\mathcal{M}, \varphi, \delta; \varepsilon) \geq C(\mathcal{M}, \varphi, \delta^*; \varepsilon^*),$$

and we obtain the assertion. □

[1] Suppose that φ is an L-smooth convex function satisfying (8) for some exponent $\rho \in [1, 2)$. For $\rho = 1$, Lemma 2.3 implies $\|\nabla \varphi(x)\| \geq \kappa$ for $x \notin \operatorname{Argmin} \varphi$. If $\rho \in (0, 2)$, on the other hand, Lemmas 2.1 and 2.3 imply $\frac{1}{2L} \|\nabla \varphi(x)\|^2 \leq \varphi(x) - \varphi^* \leq \kappa^{-1/(\rho-1)} \|\nabla \varphi(x)\|^{\frac{\rho}{\rho-1}}$ for all x, which yields $\|\nabla \varphi(x)\| \geq$ const. for all $x \notin \operatorname{Argmin} \varphi$. This contradicts to the continuity of $\nabla \varphi$ at points in $\operatorname{Argmin} \varphi$.

Under the measure δ^*, the following lower bound is known [13]:

$$C(\mathcal{F}(x_0,R,L,\kappa,\rho),\delta^*;\varepsilon)=\begin{cases}\Omega\left(\min\left\{\dim\mathbb{E},\sqrt{\dfrac{L}{\kappa^{\frac{2}{\rho}}\varepsilon^{\frac{\rho-2}{\rho}}}}\right\}\right) & :\rho>2,\\ \Omega\left(\min\left\{\dim\mathbb{E},\sqrt{\dfrac{L}{\kappa}}\log\dfrac{\kappa R^2}{\varepsilon}\right\}\right) & :\rho=2.\end{cases} \tag{12}$$

Consequently, by Proposition 2.5, we obtain lower bounds under the gradient norm measure δ:

$$C(\mathcal{F}(x_0,R,L,\kappa,\rho),\delta;\varepsilon)=\begin{cases}\Omega\left(\min\left\{\dim\mathbb{E},\sqrt{\dfrac{L}{\kappa^{\frac{1}{\rho-1}}\varepsilon^{\frac{\rho-2}{\rho-1}}}}\right\}\right) & :\rho>2,\\ \Omega\left(\min\left\{\dim\mathbb{E},\sqrt{\dfrac{L}{\kappa}}\log\dfrac{\kappa R}{\varepsilon}\right\}\right) & :\rho=2.\end{cases} \tag{13}$$

3 Accelerated proximal gradient method applied to regularized problems

This section is devoted to review the regularization strategy [8, 16], from which we construct adaptive methods in the next section. We consider to apply an accelerated proximal gradient method to the regularized problem

$$\min_{x\in E}\left[\varphi_\sigma(x):=\varphi(x)+\frac{\sigma}{2}\|x-x_0\|^2\right],$$

where $x_0\in\mathbb{E}$ is a fixed initial point and $\sigma>0$ is a regularization parameter. Since φ_σ is strongly convex, it has a unique minimizer. Let

$$x_\sigma^*:=\operatorname*{argmin}_{x\in\mathbb{E}}\varphi_\sigma(x),\quad \varphi_\sigma^*:=\min_{x\in\mathbb{E}}\varphi_\sigma(x).$$

We define the gradient mapping $g_L^\sigma(x)$ for the regularized function φ_σ in the following manner:

$$\begin{aligned}f_\sigma(x)&:=f(x)+\frac{\sigma}{2}\|x-x_0\|^2,\\ m_L^\sigma(y;x)&:=f_\sigma(y)+\langle\nabla f_\sigma(y),x-y\rangle+\frac{L}{2}\|x-y\|^2+\Psi(x),\\ T_L^\sigma(y)&:=\operatorname*{argmin}_{x\in\mathbb{E}}m_L^\sigma(y;x),\\ g_L^\sigma(y)&:=L(y-T_L^\sigma(y)).\end{aligned}$$

The following relations between $\varphi_\sigma(x)$ and $\varphi(x)$ are useful.

Lemma 3.1. *(i) $\varphi_\sigma(x)\le\varphi_\sigma(x_0)$ implies $\varphi(x)\le\varphi(x_0)$.*

(ii) $\|x_0-x_\sigma^*\|\le\operatorname{dist}(x_0,X^*)$.

(iii) We have $\|g_L(y)-g_L^\sigma(y)\|\le\sigma\|y-x_0\|$ for any $y\in\mathbb{E}$.

Proof. (i) is immediate by $\varphi(x)\le\varphi_\sigma(x)$ and $\varphi_\sigma(x_0)=\varphi(x_0)$.
(ii) For any $x^*\in X^*$, the strong convexity of φ_σ implies $\varphi_\sigma(x^*)\ge\varphi_\sigma^*+\frac{\sigma}{2}\|x^*-x_\sigma^*\|^2$, which can be rewritten as

$$\varphi(x^*)-\varphi(x_\sigma^*)\ge\frac{\sigma}{2}(-\|x^*-x_0\|^2+\|x_\sigma^*-x_0\|^2+\|x^*-x_\sigma^*\|^2).$$

As $\varphi(x^*) - \varphi(x^*_\sigma) \leq 0$, we conclude $\|x^* - x_0\|^2 \geq \|x^*_\sigma - x_0\|^2 + \|x^* - x^*_\sigma\|^2 \geq \|x^*_\sigma - x_0\|^2$, which proves the assertion.
(iii) In general, if h is a lower-semicontinuous and μ-strongly convex function, we have for any $a, b \in \mathbb{E}$ that[2]

$$\|x^*_a - x^*_b\| \leq \frac{\|a - b\|}{\mu}, \quad \text{where} \quad x^*_a = \operatorname*{argmin}_{x \in \mathbb{E}} \{\langle a, x\rangle + h(x)\}, \quad x^*_b = \operatorname*{argmin}_{x \in \mathbb{E}} \{\langle b, x\rangle + h(x)\}.$$

This fact implies

$$\|T_L(y) - T^\sigma_L(y)\| \leq \frac{\|\nabla f(y) - \nabla f_\sigma(y)\|}{L} = \frac{\sigma \|y - x_0\|}{L}.$$

Therefore, we conclude (iii) because of $\|g_L(y) - g^\sigma_L(y)\| = L \|T_L(y) - T^\sigma_L(y)\|$. □

3.1 Accelerated proximal gradient method

We employ Nesterov's accelerated proximal gradient method [17] for solving $\min_{x \in E} \varphi_\sigma(x)$, by regarding the objective function as

$$\varphi_\sigma(x) = f(x) + \Psi_\sigma(x), \quad \Psi_\sigma(x) := \Psi(x) + \frac{\sigma}{2} \|x - x_0\|^2 . \tag{14}$$

The analogy of the definition of $T_L(\cdot)$ for this regularization is

$$\tilde{T}_L(y) = \operatorname*{argmin}_{x \in \mathbb{E}} \left[f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2} \|x - y\|^2 + \Psi_\sigma(x) \right].$$

Observe that $\tilde{T}_L(y) = T^\sigma_{L+\sigma}(y)$ holds (because of the identity $m^\sigma_{L+\sigma}(y; x) = f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2} \|x - y\|^2 + \Psi_\sigma(x)$). Therefore, the accelerated method $\mathcal{A}(x_0, L_0, \sigma)$ given in [17, Algorithm (4.9)] applied to the regularized function (14) can be described as follows: *Let* $x_0 \in \mathbb{E}$, $\psi_0(x) = \frac{1}{2} \|x - x_0\|^2$, $A_0 = 0$, $L_0 \geq L_{\min}$. *Generate the sequence* $\{x_k, \psi_k, M_k, L_k, A_k\}$ *by the iteration*

$$\{x_{k+1}, \psi_{k+1}, M_k, L_{k+1}, A_{k+1}\} \leftarrow \texttt{APGIter}_\sigma(x_k, \psi_k, L_k, A_k)$$

for each $k \geq 0$. The scheme `APGIter` at each iteration is shown in Algorithm 1. Nesterov's method involves the backtracking line-search procedure to adapt the Lipschitz constant L_f with the estimate M_k (and L_{k+1}) which is controlled by the parameters $\gamma_{\text{inc}} > 1$ and $\gamma_{\text{dec}} \geq 1$.

Remark 3.2. In one execution of the loop (Lines 3–10) in `APGIter`, we have three evaluations of $\varphi(x)$ (at $x \in \{z, T_L(z), T^\sigma_{L+\sigma}(z)\}$), two evaluations of $\nabla f(x)$ (at $x \in \{y, z\}$), and three proximal operations $T^\sigma_{L+\sigma}(y), T^\sigma_{L+\sigma}(z), T_L(z)$. There is one proximal operation to compute v_k outside the loop at Line 1. Remark that $v_0 = x_0$ holds and, by the recurrence formula for ψ_k at Line 13, v_k for $k \geq 1$ can be computed as

$$v_k = \operatorname{prox}_{\gamma_k \Psi}(w_k), \text{ where } \gamma_k = \frac{A_k}{1 + \sigma A_k}, \quad w_k = x_0 - \frac{1}{1 + \sigma A_k} \sum_{i=1}^{k} a_i \nabla f(x_i). \tag{16}$$

□

[2] By the strong convexity of $\langle a, x\rangle + h(x)$ and $\langle b, x\rangle + h(x)$, we have

$$\frac{\mu}{2} \|x^*_a - x^*_b\|^2 \leq [\langle a, x^*_b\rangle + h(x^*_b)] - [\langle a, x^*_a\rangle + h(x^*_a)],$$
$$\frac{\mu}{2} \|x^*_a - x^*_b\|^2 \leq [\langle b, x^*_a\rangle + h(x^*_a)] - [\langle b, x^*_b\rangle + h(x^*_b)],$$

respectively. Adding them implies $\mu \|x^*_a - x^*_b\|^2 \leq \langle a - b, x^*_b - x^*_a\rangle \leq \|a - b\| \|x^*_b - x^*_a\|$.

Algorithm 1: Accelerated Proximal Gradient Iteration

$\{x_{k+1}, \psi_{k+1}, M_k, L_{k+1}, A_{k+1}\} \leftarrow \texttt{APGIter}_\sigma(x_k, \psi_k, L_k, A_k)$

Parameters: $\gamma_{\text{inc}} > 1$, $\gamma_{\text{dec}} \geq 1$, $L_{\min} \in (0, L_f]$.

Input: $\sigma > 0$, $x_k \in \mathbb{E}$, $\psi_k : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, $L_k > 0$, $A_k > 0$.

1: Compute $v_k := \operatorname{argmin}_{x \in \mathbb{E}} \psi_k(x)$ (cf. (16)).

2: Set $L := L_k / \gamma_{\text{inc}}$.

3: **repeat**

4: Set $L := L\gamma_{\text{inc}}$.

5: Find the largest root $a > 0$ of the equation $\frac{a^2}{A_k + a} = 2\frac{1+\sigma A_k}{L}$.

6: Set $y = \frac{A_k x_k + a v_k}{A_k + a}$.

7: Compute $z = T^\sigma_{L+\sigma}(y) = \operatorname{argmin}_x [f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + \Psi_\sigma(x)]$.

8: Compute $T_L(z)$ and $T^\sigma_{L+\sigma}(z)$.

9: Test the conditions

$$\left\langle \nabla f(y) - \nabla f(T^\sigma_{L+\sigma}(y)), y - T^\sigma_{L+\sigma}(y) \right\rangle \geq \frac{1}{L} \left\| \nabla f(y) - \nabla f(T^\sigma_{L+\sigma}(y)) \right\|^2, \tag{15a}$$

$$\varphi_\sigma(T^\sigma_{L+\sigma}(z)) \leq m^\sigma_{L+\sigma}(z; T^\sigma_{L+\sigma}(z)), \tag{15b}$$

$$\text{[Optional]} \quad \varphi(T_L(z)) \leq \varphi(z). \tag{15c}$$

10: **until** the conditions in (15) hold.

11: Define $y_k := y$, $M_k := L$, $x_{k+1} := z = T^\sigma_{M_k+\sigma}(y_k)$,

12: $a_{k+1} := a$, $A_{k+1} := A_k + a_{k+1}$, $L_{k+1} := \max\{L_{\min}, M_k/\gamma_{\text{dec}}\}$,

13: $\psi_{k+1}(x) := \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle + \Psi_\sigma(x)]$.

14: Output $\{x_{k+1}, \psi_{k+1}, M_k, L_{k+1}, A_{k+1}\}$.

Remark 3.3. Compared with the original Nesterov's method, there are small modifications for our development. The update $L_{k+1} := \max\{L_{\min}, M_k/\gamma_{\text{dec}}\}$ at Line 12 is slightly different from the original one $L_{k+1} := M_k/\gamma_{\text{dec}}$ in [17], which affects Lemma 3.5. Moreover, we have additional computations $T_L(z)$ and $T^\sigma_{L+\sigma}(z)$ at Line 8 in order to test the conditions (15b) and (15c). Remark that, when $L = L_f$ is known, the computation of $T^\sigma_{L+\sigma}(z)$ at Line 8 can be omitted as it is only used to check the second condition (15b). We let the third criterion (15c) optional; it is independent of the analysis of Algorithm 3 while we need it in Algorihtm 4. The first condition (15a) is equivalent to the one in Nesterov's method (see the proof of Lemma 5 in [17]). It will be verified in Lemmas 3.4 and 3.5 that our modification does not affect the original complexity analysis. □

Lemma 3.4. *The condition* (15) *of* `APGIter` *holds whenever* $L \geq L_f$.

Proof. The first condition (15a) is satisfied since f is L-smooth (e.g., see [15, Theorem 2.1.5]). Since f_σ is $(L_f + \sigma)$-smooth, the second one (15b) holds if $L \geq L_f$ by Lemma 2.1 (v). The third one (15c) can be verified again by Lemma 2.1 (v): If $L \geq L_f$, we have

$$\varphi(T_L(z)) \leq m_L(z; T_L(z)) \leq \varphi(z),$$

where the second inequality follows from

$$m_L(z; T_L(z)) = \min_{x \in \mathbb{E}} \left\{ f(z) + \langle \nabla f(z), x - z \rangle + \Psi(x) + \frac{L}{2} \|x - z\|^2 \right\} \overset{x=z}{\leq} f(z) + \Psi(z) = \varphi(z). \tag{17}$$

□

The following lemma is given in [17, Lemma 6] while we rewrite its proof due to the difference in the update of L_{k+1}.

Lemma 3.5. *Suppose that $L_k \le \gamma_{\mathrm{inc}} L_f$ holds. Then, `APGIter` ensures $M_k \le \gamma_{\mathrm{inc}} L_f$ and $L_{k+1} \le \max\{L_{\min}, \frac{\gamma_{\mathrm{inc}}}{\gamma_{\mathrm{dec}}} L_f\} \le \gamma_{\mathrm{inc}} L_f$. The number of the executions of the loop (Lines 3–10) is bounded by*

$$1 + \frac{\log \gamma_{\mathrm{dec}}}{\log \gamma_{\mathrm{inc}}} + \frac{1}{\log \gamma_{\mathrm{inc}}} \log \frac{L_{k+1}}{L_k}.$$

Proof. Let $n_k \ge 1$ be the number of inner loops so that $M_k = L_k \gamma_{\mathrm{inc}}^{n_k - 1}$. If $n_k = 1$, then $M_k = L_k \le \gamma_{\mathrm{inc}} L_f$. If $n_k > 1$, then $M_k \le \gamma_{\mathrm{inc}} L_f$ must hold because otherwise $L_k \gamma_{\mathrm{inc}}^{n_k - 2} = M_k / \gamma_{\mathrm{inc}} > L_f$ and the n_k-th loop cannot occur (by Lemma 3.4). Hence, we conclude $M_k \le \gamma_{\mathrm{inc}} L_f$ and also $L_{k+1} = \max\{L_{\min}, M_k/\gamma_{\mathrm{dec}}\} \le \max\{L_{\min}, \frac{\gamma_{\mathrm{inc}}}{\gamma_{\mathrm{dec}}} L_f\}$.

Now, the relation $M_k = L_k \gamma_{\mathrm{inc}}^{n_k - 1}$ implies

$$L_{k+1} = \max\{L_{\min}, M_k/\gamma_{\mathrm{dec}}\} \ge \frac{1}{\gamma_{\mathrm{dec}}} L_k \gamma_{\mathrm{inc}}^{n_k - 1}.$$

Then, we have $(n_k - 1) \log \gamma_{\mathrm{inc}} \le \log \frac{\gamma_{\mathrm{dec}} L_{k+1}}{L_k} = \log \gamma_{\mathrm{dec}} + \log \frac{L_{k+1}}{L_k}$ and therefore

$$n_k \le 1 + \frac{\log \gamma_{\mathrm{dec}}}{\log \gamma_{\mathrm{inc}}} + \frac{1}{\log \gamma_{\mathrm{inc}}} \log \frac{L_{k+1}}{L_k}.$$

□

The complexity estimate of Nesterov's method is given as follows.

Proposition 3.6. *Let $\{x_k, \psi_k, M_k, L_k, A_k\}$ be generated by the accelerated proximal gradient method applied to the regularized objective function (14), that is,*

$$\{x_{k+1}, \psi_{k+1}, M_k, L_{k+1}, A_{k+1}\} \leftarrow \texttt{APGIter}_\sigma(x_k, \psi_k, L_k, A_k), \quad k = 0, 1, 2, \ldots,$$

with the initialization $x_0 \in \mathbb{E}$, $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2$, $A_0 = 0$, $L_0 > 0$.

(i) If $L_0 \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$, then we have $L_{\min} \le L_k \le M_k \le \gamma_{\mathrm{inc}} L_f$ for all $k \ge 0$. The total number of the executions of the loop (Lines 3–10) until the k-th iteration is bounded by

$$\left[1 + \frac{\log \gamma_{\mathrm{dec}}}{\log \gamma_{\mathrm{inc}}}\right](k+1) + \frac{1}{\log \gamma_{\mathrm{inc}}} \log \frac{L_{k+1}}{L_0}.$$

(ii) For each $k \ge 1$, we have

$$A_k \ge \frac{2}{\gamma_{\mathrm{inc}} L_f} \left[1 + \sqrt{\frac{\sigma}{2\gamma_{\mathrm{inc}} L_f}}\right]^{2(k-1)}.$$

(iii) $\varphi_\sigma(x_k) - \varphi_\sigma^* \le \frac{\|x_0 - x_\sigma^*\|^2}{2A_k}$ holds for all $k \ge 1$.

(iv) $\varphi(x_k) \le \varphi(x_0)$ holds for all $k \ge 1$.

(v) For every $k \ge 1$, we have

$$\|x_k - x_0\| \le \left(1 + \frac{1}{\sqrt{\sigma A_k}}\right) \mathrm{dist}(x_0, X^*),$$

$$\left\|g_{M_{k-1}}(x_k)\right\| \le \left\|g^\sigma_{M_{k-1}+\sigma}(x_k)\right\| + \sigma \|x_k - x_0\| \le \left(2\sqrt{\frac{M_{k-1} + \sigma}{A_k}} + \sigma\right) \mathrm{dist}(x_0, X^*).$$

Proof. (i) $L_{\min} \le L_k \le M_k$ is clear by the construction. $M_k \le \gamma_{\mathrm{inc}} L_f$ is obtained by applying Lemma 3.5 inductively.

Since Ψ_σ is σ-strongly convex, (ii) follows by [17, Lemma 8] applied to the objective function $\varphi_\sigma = f + \Psi_\sigma$.

According to [17, Lemma 7], the following relations hold for all $k \ge 0$:

$$\begin{cases} A_k \varphi_\sigma(x_k) \le \min_{x \in \mathbb{E}} \psi_k(x), \\ \psi_k(x) \le A_k \varphi_\sigma(x) + \frac{1}{2} \|x - x_0\|^2, \quad \forall x \in \mathbb{E}. \end{cases}$$

Combining them, we obtain

$$A_k \varphi_\sigma(x_k) \le \min_{x \in \mathbb{E}} \left\{ A_k \varphi_\sigma(x) + \frac{1}{2} \|x - x_0\| \right\}, \quad \forall k \ge 0.$$

Taking $x = x_\sigma^*$ on the right hand side, we obtain (iii). On the other hand, taking $x = x_0$ yields $\varphi_\sigma(x_k) \le \varphi_\sigma(x_0)$. Then Lemma 3.1 (i) gives the assertion $\varphi(x_k) \le \varphi(x_0)$.

(v) By the σ-strong convexity of φ_σ and using (iii), we have

$$\frac{\sigma}{2} \|x_k - x_\sigma^*\|^2 \le \varphi_\sigma(x_k) - \varphi_\sigma^* \le \frac{\|x_0 - x_\sigma^*\|^2}{2A_k} \quad \Longrightarrow \quad \|x_k - x_\sigma^*\| \le \frac{1}{\sqrt{\sigma A_k}} \|x_0 - x_\sigma^*\|.$$

This shows the former inequality of (v):

$$\|x_k - x_0\| \le \|x_k - x_\sigma^*\| + \|x_0 - x_\sigma^*\| \le \left(1 + \frac{1}{\sqrt{\sigma A_k}}\right) \|x_0 - x_\sigma^*\| \le \left(1 + \frac{1}{\sqrt{\sigma A_k}}\right) \mathrm{dist}(x_0, X^*).$$

where the last inequality follows by Lemma 3.1 (ii).

Since $\varphi_\sigma(T^\sigma_{M_{k-1}+\sigma}(x_k)) \le m^\sigma_{M_{k-1}+\sigma}(x_k; T^\sigma_{M_{k-1}+\sigma}(x_k))$ holds by (15b), Lemma 2.1 (iv) yields

$$\frac{1}{2(M_{k-1}+\sigma)} \left\| g^\sigma_{M_{k-1}+\sigma}(x_k) \right\|^2 \le \varphi_\sigma(x_k) - \varphi_\sigma^* \le \frac{\|x_0 - x_\sigma^*\|^2}{2A_k} \le \frac{\mathrm{dist}(x_0, X^*)}{2A_k}.$$

Thus, $\left\| g^\sigma_{M_{k-1}+\sigma}(x_k) \right\| \le \sqrt{\frac{M_{k-1}+\sigma}{A_k}}\, \mathrm{dist}(x_0, X^*)$ holds. Consequently, using Lemma 2.1 (i) and Lemma 3.1 (iii), we obtain

$$\begin{aligned} \left\| g_{M_{k-1}}(x_k) \right\| &\le \left\| g_{M_{k-1}+\sigma}(x_k) \right\| \le \left\| g^\sigma_{M_{k-1}+\sigma}(x_k) \right\| + \sigma \|x_k - x_0\| \\ &\le \sqrt{\frac{M_{k-1}+\sigma}{A_k}}\, \mathrm{dist}(x_0, X^*) + \sigma \left(1 + \frac{1}{\sqrt{\sigma A_k}}\right) \mathrm{dist}(x_0, X^*) \\ &\le \left(2\sqrt{\frac{M_{k-1}+\sigma}{A_k}} + \sigma\right) \mathrm{dist}(x_0, X^*). \end{aligned}$$

□

3.2 Proximal gradient method

We end this section by presenting Algorithm 2, a basic proximal gradient descent with a backtracking strategy to estimate L_f [17, Algorithm (3.3)], which will be used in the initialization of our method. The method consists of evaluations of $\varphi(x)$ at $x \in \{x_k, T_L(x_k)\}$ (which can be omitted if L_f is known), one gradient evaluation $\nabla f(x_k)$, and proximal operations $T_L(x_k)$, for each guess L of the Lipschitz constant.

Algorithm 2: Proximal Gradient Iteration
$\{T_{M_k}(x_k), M_k, L_{k+1}\} \leftarrow \texttt{PGIter}(x_k, L_k)$

Parameters: $\gamma_{\mathrm{inc}} > 1$, $\gamma_{\mathrm{dec}} \geq 1$, $L_{\min} \in (0, L_f]$.
Input: $x_k \in \mathbb{E}$, $L_k > 0$.

1: Set $L := L_k/\gamma_{\mathrm{inc}}$.
2: **repeat**
3: Set $L := \gamma_{\mathrm{inc}} L$.
4: Compute $T_L(x_k)$.
5: **until** the following condition holds:

$$\varphi(T_L(x_k)) \leq m_L(x_k; T_L(x_k)). \tag{18}$$

6: Define $M_k := L$, $L_{k+1} := \max\{L_{\min}, M_k/\gamma_{\mathrm{dec}}\}$.
7: Output $\{T_{M_k}(x_k), M_k, L_{k+1}\}$.

Lemma 3.7. *Let $\{T_{M_k}(x_k), M_k, L_{k+1}\}$ be given by $\texttt{PGIter}(x_k, L_k)$. Then the following assertions hold.*

(i) $\varphi(T_{M_k}(x_k)) \leq \varphi(x_k)$ *holds.*

(ii) If $L_k \leq \gamma_{\mathrm{inc}} L_f$, *then we have* $M_k \leq \gamma_{\mathrm{inc}} L_f$ *and* $L_{k+1} \leq \max\{L_{\min}, \frac{\gamma_{\mathrm{inc}}}{\gamma_{\mathrm{dec}}} L_f\} \leq \gamma_{\mathrm{inc}} L_f$. *Moreover, the number of the executions of the loop (Lines 2–5) is bounded by*

$$1 + \frac{\log \gamma_{\mathrm{dec}}}{\log \gamma_{\mathrm{inc}}} + \frac{1}{\log \gamma_{\mathrm{inc}}} \log \frac{L_{k+1}}{L_k}.$$

Proof. (i) The condition (18) ensures $\varphi(T_{M_k}(x_k)) \leq m_{M_k}(x_k; T_{M_k}(x_k)) \leq \varphi(x_k)$ (recall (17) for the second inequality).

(ii) can be verified in the same way as Lemma 3.5. □

4 Adaptive proximal gradient methods

In this section, we propose an adaptive proximal gradient method (Algorithm 3) and its restart scheme (Algorithm 4). We show that these two are nearly optimal for the classes $\mathcal{F}(x_0, R, L)$ and $\mathcal{F}(x_0, R, L, \kappa, \rho)$, respectively.

4.1 Adaptive determination of the regularization parameter

For solving (5) under the measure $\|g_L(x)\|$, we propose the adaptive accelerated proximal gradient method `AdaAPG` shown in Algorithm 3, which can be seen as a simple extension of the regularization technique [16] introducing a guess-and-check procedure to adapt the regularization parameter σ. The j-th outer loop of `AdaAPG` corresponds to applying Nesterov's accelerated proximal gradient method to the regularized problem $\min_{x \in \mathbb{E}} \varphi_{\sigma_j}(x)$ where $\sigma_j = \sigma_0/\gamma_{\mathrm{reg}}^j$ ($\gamma_{\mathrm{reg}} > 1$). We stop Nesterov's method if it successfully finds an ε-solution or it iterates excessively as detected by the growth condition of A_{k+1} at Line 10. The growth of A_{k+1} is used as the criterion that the current guess of σ_j is not desirable and then we restart Nesterov's method reallocating the regularization parameter as $\sigma_{j+1} := \sigma_j/\gamma_{\mathrm{reg}}$. The proposed method involves the parameter $\beta \in (0, 1]$ which controls the accuracy of solving $\min_x \varphi_{\sigma_j}(x)$ by Nesterov's method (recall Proposition 3.6 (iii)).

The next lemma shows what happens if an outer loop fails to terminate the algorithm.

Algorithm 3: Adaptive Accelerated Proximal Gradient Method
`AdaAPG`$(x_0, L_{-1}, \sigma_0, \varepsilon)$

Parameters: $\gamma_{\mathrm{inc}} > 1$, $\gamma_{\mathrm{dec}} \geq 1$, $L_{\min} \in (0, L_f]$, $\gamma_{\mathrm{reg}} > 1$, $\beta \in (0, 1]$.
Input: $x_0 \in \mathbb{E}$, $L_{-1} \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$, $\sigma_0 > 0$, $\varepsilon > 0$.

$L_0 := L_{-1}$.
for $j = 0, 1, 2, \ldots$ **do**
  $\sigma_j := \sigma_0 / \gamma_{\mathrm{reg}}^j$.
  $\psi_0(x) := \frac{1}{2}\|x - x_0\|^2$, $A_0 := 0$.
  repeat for $k = 0, 1, 2, \ldots$
    $\{x_{k+1}, \psi_{k+1}, M_k, L_{k+1}, A_{k+1}\} \leftarrow \mathtt{APGIter}_{\sigma_j}(x_k, \psi_k, L_k, A_k)$.
    if $\|g_{M_k}(x_{k+1})\| \leq \varepsilon$ **then**
      output $\{\sigma_j, x_{k+1}, T_{M_k}(x_{k+1}), M_k, L_{k+1}\}$ and terminate the algorithm.
    end if
  until $A_{k+1} \geq \frac{2(M_{k+1} + \sigma_j)}{\beta^2 \sigma_j^2}$.
  $L_0 := L_{k+1}$.
end for

Lemma 4.1. *In* `AdaAPG`*, suppose that a j-th outer loop finished with the criterion $A_{k+1} \geq \frac{2(M_k + \sigma_j)}{\beta^2 \sigma_j^2}$ for some $k \geq 0$. Then we have*

$$\|x_{k+1} - x_0\| \leq \left(1 + \frac{\beta}{\sqrt{2}}\right) \operatorname{dist}(x_0, X^*),$$

$$\|g_{M_k}(x_{k+1})\| \leq (1 + \sqrt{2}\beta)\sigma_j \operatorname{dist}(x_0, X^*).$$

Moreover, the number of inner iterations is bounded as follows.

$$k + 1 \leq 2 + \left(\sqrt{\frac{2\gamma_{\mathrm{inc}} L_f}{\sigma_j}} + 1\right) \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_j}{\beta \sigma_j}.$$

Proof. The bounds on $\|x_{k+1} - x_0\|$ and $\|g_{M_k}(x_{k+1})\|$ can be obtained by Proposition 3.6 (v) applying the assumption $A_{k+1} \geq \frac{2(M_k + \sigma_j)}{\beta^2 \sigma_j^2}$. To show the bound on $k + 1$, suppose $k \geq 1$ (the result is clear when $k = 0$). By the assumption on k and Proposition 3.6 (i), we have

$$A_k < \frac{2(M_{k-1} + \sigma_j)}{\beta^2 \sigma_j^2} \leq \frac{2(\gamma_{\mathrm{inc}} L_f + \sigma_j)}{\beta^2 \sigma_j^2}.$$

On the other hand, Proposition 3.6 (ii) implies

$$A_k \geq \frac{2}{\gamma_{\mathrm{inc}} L_f} \left[1 + \sqrt{\frac{\sigma_j}{2\gamma_{\mathrm{inc}} L_f}}\right]^{2(k-1)} \geq \frac{2}{\gamma_{\mathrm{inc}} L_f} \exp\left(2(k-1) \frac{1}{\sqrt{\frac{2\gamma_{\mathrm{inc}} L_f}{\sigma_j}} + 1}\right),$$

where the second inequality is due to the fact[3] $1+x \geq \exp \frac{x}{1+x}$ $(\forall x > -1)$. Therefore,

$$\frac{2}{\gamma_{\mathrm{inc}} L_f} \exp\left(2(k-1)\frac{1}{\sqrt{\frac{2\gamma_{\mathrm{inc}} L_f}{\sigma_j}}+1}\right) \leq \frac{2(\gamma_{\mathrm{inc}} L_f + \sigma_j)}{\beta^2 \sigma_j^2}$$
$$\Longrightarrow k+1 \leq 2 + \frac{1}{2}\left(\sqrt{\frac{2\gamma_{\mathrm{inc}} L_f}{\sigma_j}}+1\right)\log\frac{\gamma_{\mathrm{inc}} L_f(\gamma_{\mathrm{inc}} L_f + \sigma_j)}{\beta^2\sigma_j^2}.$$

The assertion follows by relaxing $\gamma_{\mathrm{inc}} L_f \leq \gamma_{\mathrm{inc}} L_f + \sigma_j$. □

In view of Proposition 3.6 (i) and Remark 3.2, the total number of the executions of `APGIter`, say N, determines the iteration complexity of `AdaAPG`. For instance, the total number of the evaluations of $\nabla f(\cdot)$ in `AdaAPG` is bounded by

$$2\left[1 + \frac{\log \gamma_{\mathrm{dec}}}{\log \gamma_{\mathrm{inc}}}\right] N + \frac{2}{\log \gamma_{\mathrm{inc}}} \log \frac{\gamma_{\mathrm{inc}} L_f}{L_{-1}}.$$

Theorem 4.2. *In* `AdaAPG`*, let N be the total number of the executions of* `APGIter`*. Then the following assertions hold.*

(i) The algorithm terminates at the j-th outer loop whenever $\sigma_j \leq \sigma(x_0, \varepsilon)$, where

$$\sigma(x_0, \varepsilon) := \frac{\varepsilon}{(1+\sqrt{2}\beta)\operatorname{dist}(x_0, X^*)} \tag{19}$$

(We let $\sigma(x_0, \varepsilon) := +\infty$ when $x_0 \in X^*$). Moreover, we have

$$\sigma_j \geq \sigma(x_0, \varepsilon)/\gamma_{\mathrm{reg}}, \quad \forall j > 0. \tag{20}$$

(ii) Suppose that the algorithm terminates at $j = \ell$ for some $\ell \geq 0$. Then N is at most

$$\frac{\sqrt{2\gamma_{\mathrm{inc}} L_f}}{\sqrt{\gamma_{\mathrm{reg}}}-1}\left[\sqrt{\gamma_{\mathrm{reg}}}\sqrt{\frac{1}{\sigma_\ell}} - \sqrt{\frac{1}{\sigma_0}}\right]\log\frac{\gamma_{\mathrm{inc}} L_f + \sigma_\ell}{\beta\sigma_\ell}$$
$$+ \left(1 + \log_{\gamma_{\mathrm{reg}}}\frac{\sigma_0}{\sigma_\ell}\right)\left(2 + \log\frac{\gamma_{\mathrm{inc}} L_f + \sigma_\ell}{\beta\sigma_\ell}\right).$$

(iii) If $\sigma_0 \leq \sigma(x_0, \varepsilon)$, then

$$N \leq 2 + \left(\sqrt{\frac{2\gamma_{\mathrm{inc}} L_f}{\sigma_0}}+1\right)\log\frac{\gamma_{\mathrm{inc}} L_f + \sigma_0}{\beta\sigma_0}.$$

(iv) If $\sigma_0 \geq \sigma(x_0, \varepsilon)$, then N is at most

$$\frac{\sqrt{2\gamma_{\mathrm{inc}} L_f}}{\sqrt{\gamma_{\mathrm{reg}}}-1}\left[\gamma_{\mathrm{reg}}\sqrt{\frac{1}{\sigma(x_0,\varepsilon)}} - \sqrt{\frac{1}{\sigma_0}}\right]\log\left(\frac{\gamma_{\mathrm{reg}}\gamma_{\mathrm{inc}} L_f}{\beta\sigma(x_0,\varepsilon)} + \frac{1}{\beta}\right)$$
$$+ \left(2 + \log_{\gamma_{\mathrm{reg}}}\frac{\sigma_0}{\sigma(x_0,\varepsilon)}\right)\left[2 + \log\left(\frac{\gamma_{\mathrm{reg}}\gamma_{\mathrm{inc}} L_f}{\beta\sigma(x_0,\varepsilon)} + \frac{1}{\beta}\right)\right]$$
$$= O\left(\sqrt{\frac{L_f \operatorname{dist}(x_0, X^*)}{\varepsilon}}\log\frac{L_f\operatorname{dist}(x_0,X^*)}{\varepsilon} + \log\frac{\sigma_0\operatorname{dist}(x_0,X^*)}{\varepsilon}\log\frac{L_f\operatorname{dist}(x_0,X^*)}{\varepsilon}\right). \tag{21}$$

[3] In fact, since the derivative $\log(1+x)$ of the function $h(x) := (1+x)\log(1+x) - x$ is increasing and vanishes at $x = 0$, we have $\min_{x>-1} h(x) = h(0) = 0$.

Proof. (i) By Lemma 4.1, the algorithm must terminate at the j-th loop whenever

$$(1+\sqrt{2}\beta)\sigma_j \operatorname{dist}(x_0, X^*) \le \varepsilon, \quad \text{i.e.,} \quad \sigma_j \le \frac{\varepsilon}{(1+\sqrt{2}\beta)\operatorname{dist}(x_0, X^*)} = \sigma(x_0, \varepsilon).$$

To see the latter assertion, assume that $\sigma_j < \sigma(x_0,\varepsilon)/\gamma_{\mathrm{reg}}$ holds for some $j > 0$. Then the previous $(j-1)$-th loop satisfies $\sigma_{j-1} < \sigma(x_0,\varepsilon)$ so that the j-th loop is not executed by the former assertion. This verifies the assertion (20).

(ii) Since $\sigma_\ell = \sigma_0/\gamma_{\mathrm{reg}}^\ell$, we have $\ell = \log_{\gamma_{\mathrm{reg}}} \sigma_0/\sigma_\ell$. Using Lemma 4.1, N is bounded as follows.

$$\begin{aligned} N &\le \sum_{j=0}^{\ell} \left[2 + \left(\sqrt{\frac{2\gamma_{\mathrm{inc}} L_f}{\sigma_j}} + 1 \right) \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_j}{\beta \sigma_j} \right] \\ &\le 2(\ell+1) + \sqrt{2\gamma_{\mathrm{inc}} L_f} \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_\ell}{\beta\sigma_\ell} \sum_{j=0}^{\ell} \sqrt{\frac{1}{\sigma_j}} + (\ell+1) \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_\ell}{\beta\sigma_\ell}, \end{aligned}$$

where the second inequality is due to $\sigma_j \ge \sigma_\ell$ $(\forall j \le \ell)$. Note that

$$\begin{aligned} \sum_{j=0}^{\ell} \sqrt{\frac{1}{\sigma_j}} &= \sum_{j=0}^{\ell} \sqrt{\frac{1}{\sigma_0/\gamma_{\mathrm{reg}}^j}} = \sqrt{\frac{1}{\sigma_0}} \sum_{j=0}^{\ell} \sqrt{\gamma_{\mathrm{reg}}}^j = \sqrt{\frac{1}{\sigma_0}} \frac{\sqrt{\gamma_{\mathrm{reg}}}^{\ell+1} - 1}{\sqrt{\gamma_{\mathrm{reg}}} - 1} \\ &= \sqrt{\frac{1}{\sigma_0}} \frac{\sqrt{\gamma_{\mathrm{reg}}}\sqrt{\sigma_0/\sigma_\ell} - 1}{\sqrt{\gamma_{\mathrm{reg}}} - 1} = \frac{1}{\sqrt{\gamma_{\mathrm{reg}}} - 1} \left[\sqrt{\gamma_{\mathrm{reg}}} \sqrt{\frac{1}{\sigma_\ell}} - \sqrt{\frac{1}{\sigma_0}} \right] \end{aligned}$$

Therefore, we see that

$$N \le (\ell+1) \left(2 + \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_\ell}{\beta\sigma_\ell} \right) + \frac{\sqrt{2\gamma_{\mathrm{inc}} L_f}}{\sqrt{\gamma_{\mathrm{reg}}} - 1} \left[\sqrt{\gamma_{\mathrm{reg}}} \sqrt{\frac{1}{\sigma_\ell}} - \sqrt{\frac{1}{\sigma_0}} \right] \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_\ell}{\beta\sigma_\ell}.$$

The assertion follows by substituting $\ell = \log_{\gamma_{\mathrm{reg}}} \sigma_0/\sigma_\ell$.

In view of (i), the assertions (iii) and (iv) follow by applying (ii) with $\sigma_\ell = \sigma_0$ and $\sigma_\ell \ge \sigma(x_0,\varepsilon)/\gamma_{\mathrm{reg}}$, respectively. The big-$O$ expression (21) is obtained by substituting the definition of $\sigma(x_0,\varepsilon)$. □

If σ_0 is chosen appropriately, then the complexity estimates in Theorem 4.2 (iii) and (vi) match the lower complexity bound (11) for the class $\mathcal{F}(x_0, R, L)$, up to a logarithmic factor. Nesterov's regularization technique [16] essentially corresponds to the ideal choice $\sigma_0 = \sigma(x_0,\varepsilon)$, which requires $\operatorname{dist}(x_0, X^*)$ to be known. Here we show a simple example to choose σ_0 so that `AdaAPG` attains the near optimality without knowing $\operatorname{dist}(x_0, X^*)$.

Corollary 4.3. *Given $x_0 \in \mathbb{E}$ and $L_{-1} \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$, apply $\{T_M(x_0), M, L\} \leftarrow$ `PGIter`(x_0, L_{-1}). For any $\varepsilon > 0$ such that $\|g_M(x_0)\| \ge \varepsilon$, choose σ_0 from the interval*

$$\sigma_0 \in \left[\frac{2\varepsilon M}{(1+\sqrt{2}\beta)\,\|g_M(x_0)\|}, \frac{2M}{1+\sqrt{2}\beta} \right]. \tag{22}$$

Then we have

$$\sigma(x_0,\varepsilon) \le \sigma_0 \le \frac{2\gamma_{\mathrm{inc}} L_f}{1+\sqrt{2}\beta}. \tag{23}$$

Consequently, `AdaAPG`$(x_0, L_{-1}, \sigma_0, \varepsilon)$ *finds $x_k \in \mathbb{E}$ and $M_k > 0$ satisfying $\|g_{M_k}(x_k)\| \le \varepsilon$ with the iteration complexity at most*

$$O\left(\sqrt{\frac{L_f \operatorname{dist}(x_0, X^*)}{\varepsilon}} \log \frac{L_f \operatorname{dist}(x_0, X^*)}{\varepsilon} \right). \tag{24}$$

Proof. Since $\varphi(T_M(x_0)) \le m_M(x_0; T_M(x_0))$ holds by the condition (18) in `PGIter`, Lemma 2.1 (iv) implies

$$\operatorname{dist}(x_0, X^*) \ge \frac{1}{2M} \|g_M(x_0)\| \ge \frac{1}{2M}\varepsilon \ge \frac{1}{2\gamma_{\mathrm{inc}} L_f}\varepsilon,$$

where the last inequality follows from Lemma 3.7 (ii). This can be rewritten as

$$\sigma(x_0, \varepsilon) \equiv \frac{\varepsilon}{(1+\sqrt{2}\beta)\operatorname{dist}(x_0, X^*)} \le \frac{2\varepsilon M}{(1+\sqrt{2}\beta)\|g_M(x_0)\|} \le \frac{2M}{1+\sqrt{2}\beta} \le \frac{2\gamma_{\mathrm{inc}} L_f}{1+\sqrt{2}\beta}.$$

Therefore, (22) implies (23). In particular, we can apply Theorem 4.2 (iv) to obtain the complexity bound (21) whose second term is dominated by the first one due to the upper bound (23) of σ_0. □

Finally, we make an observation on the convergence of the proposed method. Let us consider the execution of `AdaAPG` with $\varepsilon = 0$ which is possibly an infinite step algorithm. Then the choice

$$\sigma_0 := 2M/(1+2\sqrt{\beta})$$

given in Corollary 4.3 ensures the nearly optimal complexity bound (24) for *every* $\varepsilon \in (0, \|g_M(x_0)\|_*)$. This means that, with this choice of σ_0, the algorithm `AdaAPG` with $\varepsilon = 0$ yields a *nearly optimal convergence* with respect to $\|g_L(x)\|$.

4.2 Adaptive restart algorithm under the Hölderian error bound condition

In this section, given an initial point $x^{(0)} \in \mathbb{E}$, we assume that the objective function φ admits the Hölderian error bound condition

$$\varphi(x) - \varphi^* \ge \kappa \operatorname{dist}(x, X^*)^\rho, \quad \forall x \in \operatorname{lev}_\varphi(\varphi(x^{(0)})), \tag{25}$$

for some $\kappa > 0$ and $\rho \ge 1$. For this case, we propose the restart scheme `rAdaAPG` described in Algorithm 4. Namely, given a current solution $x^{(t)}$, we apply a proximal gradient iteration $x_+^{(t)} := T_{M^{(t)}}(x^{(t)})$ from which we start `AdaAPG` to find the next $x^{(t+1)}$ reducing the gradient mapping norm at the ratio $\theta \in (0,1)$:

$$\left\| g_{M^{(t+1)}}(x^{(t+1)}) \right\| \le \theta \left\| g_{M^{(t)}}(x^{(t)}) \right\|.$$

Remarkably, the regularization parameter $\sigma^{(t)}$ is input to `AdaAPG` which adaptively finds the next $\sigma^{(t+1)}$. Therefore, the next $x^{(t+1)}$ can be seen as an approximate solution to the regularized problem

$$\min_{x \in \mathbb{E}} \left[\varphi(x) + \frac{\sigma^{(t+1)}}{2} \left\| x - x_+^{(t)} \right\|^2 \right]$$

computed by an accelerated proximal gradient method. Finally, we note that the initial regularization parameter $\sigma^{(0)}$ may be determined depending on the result of Line 1, as observed in Corollary 4.5.

4.2.1 Iteration complexity results and near optimality

To show the iteration complexity result, observe that the total number N of the executions of `APGIter` determines the complexity of `rAdaAPG`. For instance, the total number of the evaluations of $\nabla f(\cdot)$ in `rAdaAPG` can be bounded by (recall Proposition 3.6 (i), Lemma 3.7 (ii), and Remark 3.2)

$$\left[1 + \frac{\log \gamma_{\mathrm{dec}}}{\log \gamma_{\mathrm{inc}}}\right](2N+1) + \frac{2}{\log \gamma_{\mathrm{inc}}} \log \frac{\gamma_{\mathrm{inc}} L_f}{L^{(-1)}}.$$

Algorithm 4: Restart Scheme for Adaptive Accelerated Proximal Gradient Method `rAdaAPG`$(x^{(0)}, L^{(-1)}, \sigma^{(0)}, \varepsilon)$

Parameters:
$\gamma_{\mathrm{inc}} > 1$, $\gamma_{\mathrm{dec}} \geq 1$, $L_{\min} \in (0, L_f]$ (for backtracking line-search);
$\gamma_{\mathrm{reg}} > 1$ (the ratio to reallocate the regularization parameter);
$\beta \in (0, 1]$ (controls the accuracy applying accelerated proximal gradient method);
$\theta \in (0, 1)$ (the ratio reducing the residue per iteration).
Input: $x^{(0)} \in \mathbb{E}$, $L^{(-1)} \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$, $\sigma^{(0)} > 0$, $\varepsilon > 0$.

1: Compute $\{x_+^{(0)}, M^{(0)}, L^{(0)}\} \leftarrow$ `PGIter`$(x^{(0)}, L^{(-1)})$.
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: **if** $\left\|g_{M^{(t)}}(x^{(t)})\right\| \leq \varepsilon$ **then** output $\{x^{(t)}, M^{(t)}, x_+^{(t)}\}$ and terminate the algorithm.
4: $\varepsilon^{(t)} := \theta \left\|g_{M^{(t)}}(x^{(t)})\right\|$.
5: $\{\sigma^{(t+1)}, x^{(t+1)}, x_+^{(t+1)}, M^{(t+1)}, L^{(t+1)}\} \leftarrow$ `AdaAPG`$(x_+^{(t)}, L^{(t)}, \sigma^{(t)}, \varepsilon^{(t)})$
6: (with testing the optional condition (15c) of `APGIter`).
7: **end for**

We focus on the bound on N in the remaining of this section.

To describe the iteration complexity bounds, we define $N(\varepsilon, \sigma_*, C)$ and $\bar{\sigma}$ as follows. Given $\varepsilon, \sigma_*, C > 0$, let $N(\varepsilon, \sigma_*, C) := 0$ if $\left\|g_{M^{(0)}}(x^{(0)})\right\| \leq \varepsilon$ and otherwise

$$\begin{aligned} N(\varepsilon, \sigma_*, C) := & \left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon} + \log_{\gamma_{\mathrm{reg}}} \frac{\sigma^{(0)}}{\sigma_*}\right) \left(2 + \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*}\right) \\ & + \frac{\sqrt{2\gamma_{\mathrm{inc}} L_f}}{\sqrt{\gamma_{\mathrm{reg}}} - 1} \left[\sqrt{\frac{1}{\sigma_*}} - \sqrt{\frac{1}{\sigma^{(0)}}}\right] \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*} \\ & + C\sqrt{2\gamma_{\mathrm{inc}} L_f} \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*}. \end{aligned} \tag{26}$$

Moreover, we define

$$\bar{\sigma} = \begin{cases} \dfrac{\theta}{1 + \sqrt{2}\beta} \kappa^{\frac{1}{\rho - 1}} \left(\dfrac{L_f}{L_{\min}} + 1\right)^{-\frac{1}{\rho - 1}} \varepsilon^{\frac{\rho - 2}{\rho - 1}} & (\text{if } \rho \geq 2), \\ \dfrac{\theta}{1 + \sqrt{2}\beta} \kappa^{\frac{2}{\rho}} \left(\dfrac{L_f}{L_{\min}} + 1\right)^{-1} (\varphi(x^{(0)}) - \varphi^*)^{-\frac{2 - \rho}{\rho}} & (\text{if } 1 \leq \rho < 2). \end{cases} \tag{27}$$

The next theorem provides the descriptions of iteration complexity bounds which have complicated expressions to explicit their dependence on parameters. A simplified form is presented in Corollary 4.5. Their proofs are given in Section 4.2.2.

Theorem 4.4. *Assume that the Hölderian error bound condition (25) holds. In* `rAdaAPG`*, let N be the total number of the executions of* `APGIter`*. Denote*

$$\sigma_* := \begin{cases} \sigma^{(0)} & (\textit{if } \sigma^{(0)} \leq \bar{\sigma}), \\ \bar{\sigma}/\gamma_{\mathrm{reg}} & (\textit{otherwise}), \end{cases} \tag{28}$$

for $\bar{\sigma}$ defined in (27). Also, let $N(\cdot, \cdot, \cdot)$ be defined by (26). Then the following assertions hold.

(i) If $\rho = 2$, then $N \leq N(\varepsilon, \sigma_*, C)$ holds with

$$C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon}\right).$$

(ii) Suppose $\rho > 2$. If $\sigma^{(0)} \geq \bar{\sigma}$, then $N \leq N(\varepsilon, \sigma_*, C)$ holds with

$$C = \sqrt{\frac{1}{\sigma^{(0)}}} \left(1 + \frac{\rho - 1}{\rho - 2} \log_{1/\theta} \frac{\sigma_*^{(0)}}{\min(\sigma_*^{(0)}, \sigma^{(0)})}\right) + \frac{\sqrt{\gamma_{\mathrm{reg}}}}{1 - \sqrt{\theta}^{\frac{\rho-2}{\rho-1}}} \left(\sqrt{\frac{1}{\bar{\sigma}}} - \sqrt{\theta}^{\frac{\rho-2}{\rho-1}} \sqrt{\frac{1}{\sigma^{(0)}}}\right),$$

where $\sigma_*^{(0)} := \frac{\theta}{1+\sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1\right)^{-\frac{1}{\rho-1}} \|g_{M^{(0)}}(x^{(0)})\|_*^{\frac{\rho-2}{\rho-1}}$. Otherwise, it follows $\sigma^{(t)} = \sigma^{(0)}$ $(\forall t \geq 0)$ and $N \leq N(\varepsilon, \sigma^{(0)}, C)$ holds with

$$C = \sqrt{\frac{1}{\sigma^{(0)}}} \left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon}\right).$$

(iii) Suppose $\rho \in (1, 2)$. Then we have

$$N \leq 1 + N(\max(\varepsilon, \varepsilon_*), \sigma_*, C) + \left(\log \frac{1}{\rho - 1}\right)^{-1} \left(\log\log \frac{\varepsilon_*}{\theta^{\frac{\rho-1}{2-\rho}} \min(\varepsilon, \varepsilon_*)} - \log\log \frac{1}{\theta^{\frac{\rho-1}{2-\rho}}}\right), \quad (29)$$

where

$$\varepsilon_* = \left[(1 + \sqrt{2})\theta^{-1}(\gamma_{\mathrm{inc}} L_f + \sigma^{(0)})\right]^{-\frac{\rho-1}{2-\rho}} \kappa^{\frac{1}{2-\rho}} \left(\frac{L_f}{L_{\min}} + 1\right)^{-\frac{1}{2-\rho}}, \quad (30)$$

$$C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\max(\varepsilon, \varepsilon_*)}\right).$$

(iv) If $\rho = 1$, then we have

$$N \leq 1 + N(\max(\varepsilon, \varepsilon_*), \sigma_*, C),$$

where

$$\varepsilon_* = \kappa \left(\frac{L_f}{L_{\min}} + 1\right)^{-1}, \quad C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\max(\varepsilon, \varepsilon_*)}\right).$$

Moreover, if $\varepsilon < \varepsilon_*$, then $x_+^{(t)}$ in the output of the algorithm must be an optimal solution to the original problem (5).

Corollary 4.5. *Let $x^{(0)} \in \mathbb{E}$ and $L^{(-1)} \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$. Assume that the Hölderian error bound condition (25) holds. After the initialization step (Line 1) in* `rAdaAPG`*, determine $\sigma^{(0)}$ by*

$$\sigma^{(0)} := \frac{2\varepsilon^{(0)} M}{(1 + \sqrt{2}\beta) \left\|g_M(x_+^{(0)})\right\|}, \quad \text{where} \quad \{T_M(x_+^{(0)}), M, L\} \leftarrow \texttt{PGIter}(x_+^{(0)}, L^{(0)}). \quad (31)$$

Then, `rAdaAPG`$(x^{(0)}, L^{(-1)}, \sigma^{(0)}, \varepsilon)$ *finds $x^{(t)} \in \mathbb{E}$ and $M^{(t)} > 0$ such that $\left\|g_{M^{(t)}}(x^{(t)})\right\| \leq \varepsilon$ with iteration complexity at most the following quantities, where we denote*

$$g_0 := g_{M^{(0)}}(x^{(0)}), \quad \Delta_0 := \varphi(x^{(0)}) - \varphi^*,$$

ε_* *is defined by (30), and we regard* $\theta, \gamma_{\mathrm{inc}}, \gamma_{\mathrm{reg}}, \beta, \frac{L_f}{L_{\min}}$ *as constants.*

$$\begin{cases} O\left(\left[\log \frac{\|g_0\|}{\varepsilon} + \sqrt{\frac{L_f}{\kappa^{\frac{1}{\rho-1}} \varepsilon^{\frac{\rho-2}{\rho-1}}}}\right] \log \frac{L_f}{\varepsilon^{\frac{\rho-2}{\rho-1}}}\right) & (\rho > 2), \\ O\left(\sqrt{\frac{L_f}{\kappa}} \log \frac{L_f}{\kappa} \log \frac{\|g_0\|}{\varepsilon}\right) & (\rho = 2), \\ O\left(\sqrt{\frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}}} \log \frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}} \log \frac{\|g_0\| L_f^{\frac{\rho-1}{2-\rho}}}{\kappa^{\frac{1}{2-\rho}}} + (\log \frac{1}{\rho-1})^{-1} \log\log \frac{\kappa^{\frac{1}{2-\rho}}}{L_f^{\frac{\rho-1}{2-\rho}} \varepsilon}\right) & (\rho \in (1,2), \varepsilon \le \varepsilon_*), \\ O\left(\sqrt{\frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}}} \log \frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}} \log \frac{\|g_0\|}{\varepsilon}\right) & (\rho \in (1,2), \varepsilon \ge \varepsilon_*), \\ O\left(\sqrt{\frac{L_f \Delta_0}{\kappa^2}} \log \frac{L_f \Delta_0}{\kappa^2} \log \frac{\|g_0\|}{\max(\kappa, \varepsilon)}\right) & (\rho = 1). \end{cases} \tag{32}$$

Let us observe the consequences of Corollary 4.5. Notice that the iteration complexity bounds (32) in the cases $\rho = 2$ and $\rho > 2$ match the lower bounds (13) up to a logarithmic factor. Therefore, `rAdaAPG` achieves the near optimality for the class $\mathcal{F}(x_0, R, L, \kappa, \rho)$.

Note that $\sigma^{(0)}$ defined in (31) is independent of ε. Therefore, if we consider `rAdaAPG` with $\varepsilon = 0$, the algorithm ensures the above iteration complexity for every ε so that we obtain a convergence result.

The algorithm `rAdaAPG` with $\varepsilon = 0$ can also provide an iteration complexity result with respect to the measure $\varphi(\cdot) - \varphi^*$. As we prove the inequality (35), the following relation holds if $\rho \neq 1$.

$$\varphi(x_+^{(t)}) - \varphi^* \le \frac{1}{\kappa^{\frac{1}{\rho-1}}} \left(\frac{L_f}{L_{\min}} + 1\right)^{\frac{\rho}{\rho-1}} \left\| g_{M^{(t)}}(x^{(t)}) \right\|^{\frac{\rho}{\rho-1}}. \tag{33}$$

This means that, given $\delta > 0$, we have the following implication:

$$\left\| g_{M^{(t)}}(x^{(t)}) \right\| \le \varepsilon := \kappa^{\frac{1}{\rho}} \left(\frac{L_f}{L_{\min}} + 1\right)^{-1} \delta^{\frac{\rho-1}{\rho}} \quad \Longrightarrow \quad \varphi(x_+^{(t)}) - \varphi^* \le \delta.$$

Substituting this ε to our complexity bound (32), we also obtain an iteration complexity bound under the measure $\varphi(\cdot) - \varphi^*$ which is nearly optimal in view of the lower complexity bound (12). Although it enjoys an adaptive and nearly optimal convergence, the proposed method does not provide a stopping criterion for the measure $\varphi(\cdot) - \varphi^*$ since the right hand side of (33) is not verifiable unless we know κ and ρ.

4.2.2 Proof of the main results

Here we complete the proofs of Theorem 4.4 and Corollary 4.5. We prepare some lemmas below.

Lemma 4.6. *Assume that the Hölderian error bound condition (25) holds. In the execution of* `rAdaAPG`*, the following assertions hold.*

(i) $L^{(t)}, M^{(t)} \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$ *for all* $t \ge 0$.

(ii) $\varphi(x_+^{(t+1)}) \le \varphi(x_+^{(t)}) \le \cdots \le \varphi(x_+^{(0)}) \le \varphi(x^{(0)})$ *for all* $t \ge 0$.

(iii) $t \le \log_{1/\theta} \dfrac{\left\| g_{M^{(0)}}(x^{(0)}) \right\|}{\left\| g_{M^{(t)}}(x^{(t)}) \right\|}$ *for each* $t \ge 0$.

(iv) Whenever $x_+^{(t)} \notin X^*$, we have

$$\operatorname{dist}(x_+^{(t)}, X^*)^{\rho-1} \le \frac{1}{\kappa}\left(\frac{L_f}{L_{\min}} + 1\right)\left\|g_{M^{(t)}}(x^{(t)})\right\|, \tag{34}$$

$$(\varphi(x_+^{(t)}) - \varphi^*)^{\rho-1} \le \frac{1}{\kappa}\left(\frac{L_f}{L_{\min}} + 1\right)^{\rho}\left\|g_{M^{(t)}}(x^{(t)})\right\|^{\rho}. \tag{35}$$

Proof. (i) Since $L^{(-1)} \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$, it follows $L^{(0)}, M^{(0)} \in [L_{\min}, \gamma_{\mathrm{inc}} L_f]$ by Lemma 3.7 (ii). Then, using Proposition 3.6 (i) inductively, we obtain (i).

(ii) For $t = 0$, we have $\varphi(x_+^{(0)}) = \varphi(T_{M^{(0)}}(x^{(0)})) \le \varphi(x^{(0)})$ by Lemma 3.7 (i). Moreover, according to the criterion (15c) and Proposition 3.6 (iv) applied to the subroutine `AdaAPG` in `rAdaAPG` satisfies

$$\varphi(x_+^{(t+1)}) \le \varphi(x^{(t+1)}) \quad \text{and} \quad \varphi(x^{(t+1)}) \le \varphi(x_+^{(t)}),$$

respectively. This shows $\varphi(x_+^{(t+1)}) \le \varphi(x_+^{(t)})$.

(iii) Since the recurrence $\varepsilon^{(t)}/\theta = \left\|g_{M^{(t)}}(x^{(t)})\right\| \le \varepsilon^{(t-1)}$ holds for $t \ge 1$, we have $\varepsilon^{(0)} \ge \varepsilon^{(t)}/\theta^t$ for each $t \ge 0$. Therefore,

$$t \le \log_{1/\theta} \frac{\varepsilon^{(0)}}{\varepsilon^{(t)}} = \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\left\|g_{M^{(t)}}(x^{(t)})\right\|}.$$

(iv) Since $x_+^{(t)} = T_{M^{(t)}}(x^{(t)})$, Lemma 2.1 (iii) implies that $g_t := \nabla f(x_+^{(t)}) - \nabla f(x^{(t)}) + g_{M^{(t)}}(x^{(t)})$ belongs to $\partial\varphi(x_+^{(t)})$. Then, Lemma 2.3 shows (note that $x_+^{(t)}$ belongs to $\operatorname{lev}_{\varphi}(\varphi(x^{(0)})) \setminus X^*$ by (ii))

$$\kappa \operatorname{dist}(x_+^{(t)}, X^*)^{\rho-1} \le \|g_t\| \le \left(\frac{L_f}{M^{(t)}} + 1\right)\left\|g_{M^{(t)}}(x^{(t)})\right\| \le \left(\frac{L_f}{L_{\min}} + 1\right)\left\|g_{M^{(t)}}(x^{(t)})\right\|,$$

where the second inequality is due to Lemma 2.1 (iii) and the last follows by (i). This shows the assertion (34). Similarly, (35) can be obtained using (10). □

The following lemma plays an essential role to derive our iteration complexity results.

Lemma 4.7. *Let $N^{(t)}$ be the number of the executions of* `APGIter` *at the t-th outer loop of* `rAdaAPG`*. Assume that, for some $T_* \ge 0$, $\sigma_* > 0$, and $\varepsilon_* > 0$, we have*

$$\sigma^{(t+1)} \ge \sigma_* \quad (t = 0, \dots, T_*) \quad \text{and} \quad \left\|g_{M^{(T_*)}}(x^{(T_*)})\right\| \ge \varepsilon_*.$$

Then, under the definition (26) of $N(\cdot,\cdot,\cdot)$, the following inequality holds.

$$\sum_{t=0}^{T_*} N^{(t)} \le N(\varepsilon_*, \sigma_*, C) \le N(\varepsilon_*, \sigma_*, C_*),$$

where

$$C := \sum_{t=0}^{T_*} \sqrt{\frac{1}{\sigma^{(t+1)}}} \le C_* := \sqrt{\frac{1}{\sigma_*}}\left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon_*}\right).$$

Proof. Since $\sigma^{(t+1)} \geq \sigma_*$ for each $t = 0, \ldots, T_*$, Theorem 4.2 (ii) gives the following bound for $t = 0, \ldots, T_*$:

$$\begin{aligned} N^{(t)} &\leq \left(1 + \log_{\gamma_{\mathrm{reg}}} \frac{\sigma^{(t)}}{\sigma^{(t+1)}}\right)\left(2 + \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma^{(t+1)}}{\beta \sigma^{(t+1)}}\right) \\ &\quad + \frac{\sqrt{2\gamma_{\mathrm{inc}} L_f}}{\sqrt{\gamma_{\mathrm{reg}}} - 1}\left[\sqrt{\gamma_{\mathrm{reg}}}\sqrt{\frac{1}{\sigma^{(t+1)}}} - \sqrt{\frac{1}{\sigma^{(t)}}}\right] \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma^{(t+1)}}{\beta \sigma^{(t+1)}} \\ &\leq \left(1 + \log_{\gamma_{\mathrm{reg}}} \frac{\sigma^{(t)}}{\sigma^{(t+1)}}\right)\left(2 + \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*}\right) \\ &\quad + \frac{\sqrt{2\gamma_{\mathrm{inc}} L_f}}{\sqrt{\gamma_{\mathrm{reg}}} - 1}\left[\sqrt{\gamma_{\mathrm{reg}}}\sqrt{\frac{1}{\sigma^{(t+1)}}} - \sqrt{\frac{1}{\sigma^{(t)}}}\right] \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*}. \end{aligned}$$

By Lemma 4.6 (iii), T_* is bounded by

$$T_* \leq \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\left\|g_{M^{(T_*)}}(x^{(T_*)})\right\|} \leq \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon_*}.$$

To estimate the sum $\sum_{t=0}^{T_*} N^{(t)}$, note that

$$\sum_{t=0}^{T_*}\left(1 + \log_{\gamma_{\mathrm{reg}}} \frac{\sigma^{(t)}}{\sigma^{(t+1)}}\right) = 1 + T_* + \log_{\gamma_{\mathrm{reg}}} \frac{\sigma^{(0)}}{\sigma^{(T_*+1)}} \leq 1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon_*} + \log_{\gamma_{\mathrm{reg}}} \frac{\sigma^{(0)}}{\sigma_*},$$

and

$$\begin{aligned} \sum_{t=0}^{T_*}\left[\sqrt{\gamma_{\mathrm{reg}}}\sqrt{\frac{1}{\sigma^{(t+1)}}} - \sqrt{\frac{1}{\sigma^{(t)}}}\right] &= (\sqrt{\gamma_{\mathrm{reg}}} - 1)\sum_{t=0}^{T_*}\sqrt{\frac{1}{\sigma^{(t+1)}}} + \sum_{t=0}^{T_*}\left[\sqrt{\frac{1}{\sigma^{(t+1)}}} - \sqrt{\frac{1}{\sigma^{(t)}}}\right] \\ &= (\sqrt{\gamma_{\mathrm{reg}}} - 1)\sum_{t=0}^{T_*}\sqrt{\frac{1}{\sigma^{(t+1)}}} + \sqrt{\frac{1}{\sigma^{(T_*+1)}}} - \sqrt{\frac{1}{\sigma^{(0)}}}. \end{aligned}$$

Therefore, $\sum_{t=0}^{T_*} N^{(t)}$ is bounded by

$$\begin{aligned} \sum_{t=0}^{T_*} N^{(t)} &\leq \left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon_*} + \log_{\gamma_{\mathrm{reg}}} \frac{\sigma^{(0)}}{\sigma_*}\right)\left(2 + \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*}\right) \\ &\quad + \frac{\sqrt{2\gamma_{\mathrm{inc}} L_f}}{\sqrt{\gamma_{\mathrm{reg}}} - 1}\left[\sqrt{\frac{1}{\sigma_*}} - \sqrt{\frac{1}{\sigma^{(0)}}}\right] \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*} \\ &\quad + \sqrt{2\gamma_{\mathrm{inc}} L_f}\sum_{t=0}^{T_*}\sqrt{\frac{1}{\sigma^{(t+1)}}} \log \frac{\gamma_{\mathrm{inc}} L_f + \sigma_*}{\beta \sigma_*} \\ &= N(\varepsilon_*, \sigma_*, C). \end{aligned}$$

Finally, C has the following bound:

$$C = \sum_{t=0}^{T_*}\sqrt{\frac{1}{\sigma^{(t+1)}}} \leq \sqrt{\frac{1}{\sigma_*}}(1 + T_*) \leq \sqrt{\frac{1}{\sigma_*}}\left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\varepsilon_*}\right) = C_*,$$

which also concludes $N(\varepsilon_*, \sigma_*, C) \leq N(\varepsilon_*, \sigma_*, C_*)$. □

In view of this lemma, the complexity analysis boils down to analyze lower bounds of $\sigma^{(t)}$ as we discuss next.

Lemma 4.8. *Assume that the Hölderian error bound condition (25) holds. Suppose that* `rAdaAPG` *terminated with the stopping criterion at the $(T+1)$-th outer loop for some $T \geq 0$. Let $\bar{\sigma}$ be defined by (27).*

(i) For each $t = 0, \ldots, T$, we have

$$\sigma^{(t+1)} \begin{cases} = \sigma^{(t)} & (\sigma^{(t)} \leq \sigma(x_+^{(t)}, \varepsilon^{(t)})), \\ \geq \sigma(x_+^{(t)}, \varepsilon^{(t)})/\gamma_{\mathrm{reg}} & (\textit{otherwise}), \end{cases}$$

where $\sigma(\cdot, \cdot)$ is defined by (19).

(ii) It follows that

$$\sigma(x_+^{(t)}, \varepsilon^{(t)}) \geq \bar{\sigma}, \quad t = 0, \ldots, T. \tag{36}$$

When $\rho \geq 2$, we further obtain

$$\sigma(x_+^{(t)}, \varepsilon^{(t)}) \geq \frac{\theta^{\frac{1}{\rho-1}}}{1+\sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-\frac{1}{\rho-1}} (\varepsilon^{(t)})^{\frac{\rho-2}{\rho-1}} \geq \bar{\sigma} \tag{37}$$

for each $t = 0, \ldots, T$.

(iii) For each $t = 0, \ldots, T+1$, we obtain

$$\sigma^{(t)} \begin{cases} = \sigma^{(0)} & (\sigma^{(0)} \leq \bar{\sigma}), \\ \geq \bar{\sigma}/\gamma_{\mathrm{reg}} & (\textit{otherwise}). \end{cases}$$

Proof. (i) In the case $\sigma^{(t)} \leq \sigma(x_+^{(t)}, \varepsilon^{(t)})$, Theorem 4.2 (i) implies that the subroutine `AdaAPG` at the t-th outer loop must terminate at the first loop $j = 0$ so that $\sigma^{(t+1)}$ is defined by $\sigma^{(t+1)} = \sigma^{(t)}$. On the other hand, consider the case $\sigma^{(t)} > \sigma(x_+^{(t)}, \varepsilon^{(t)})$. When the subroutine `AdaAPG` at the t-th outer loop terminates at the first loop $j = 0$, then it is clear that

$$\sigma^{(t+1)} = \sigma^{(t)} > \sigma(x_+^{(t)}, \varepsilon^{(t)}) \geq \sigma(x_+^{(t)}, \varepsilon^{(t)})/\gamma_{\mathrm{reg}}.$$

In the another case, (20) implies $\sigma^{(t+1)} \geq \sigma(x_+^{(t)}, \varepsilon^{(t)})/\gamma_{\mathrm{reg}}$.

(ii) We may assume $x_+^{(t)} \notin X^*$ which allows us to apply Lemma 4.6 (iv) (Note that the assertion is clear if $x_+^{(t)} \in X^*$ since then $\sigma(x_+^{(t)}, \varepsilon^{(t)}) = +\infty$). In the case $\rho \geq 2$, using (34) implies

$$\sigma(x_+^{(t)}, \varepsilon^{(t)}) = \frac{\theta \left\| g_{M^{(t)}}(x^{(t)}) \right\|}{(1+\sqrt{2}\beta)\operatorname{dist}(x_+^{(t)}, X^*)} \geq \frac{\theta}{1+\sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-\frac{1}{\rho-1}} \left\| g_{M^{(t)}}(x^{(t)}) \right\|^{\frac{\rho-2}{\rho-1}}.$$

Since $\varepsilon^{(t)} \equiv \theta \|g_{M^{(t)}}(x^{(t)})\|_*$ and $\|g_{M^{(t)}}(x^{(t)})\|_* \geq \varepsilon$ (by the definition of T), we obtain (37).

If $\rho \in [1, 2)$, on the other hand, since $\varphi(x_+^{(t)}) \leq \varphi(x^{(0)})$ by Lemma 4.6 (ii), the inequalities (25) and (35) imply (remark that $\frac{2-\rho}{\rho} > 0$)

$$\begin{aligned} \operatorname{dist}(x_+^{(t)}, X^*) &\leq \frac{1}{\kappa^{\frac{1}{\rho}}} (\varphi(x_+^{(t)}) - \varphi^*)^{\frac{1}{\rho}} = \frac{1}{\kappa^{\frac{1}{\rho}}} (\varphi(x_+^{(t)}) - \varphi^*)^{\frac{\rho-1}{\rho}} (\varphi(x_+^{(t)}) - \varphi^*)^{\frac{2-\rho}{\rho}} \\ &\leq \frac{1}{\kappa^{\frac{2}{\rho}}} \left(\frac{L_f}{L_{\min}} + 1 \right) \left\| g_{M^{(t)}}(x^{(t)}) \right\| (\varphi(x_+^{(t)}) - \varphi^*)^{\frac{2-\rho}{\rho}} \\ &\leq \frac{1}{\kappa^{\frac{2}{\rho}}} \left(\frac{L_f}{L_{\min}} + 1 \right) \left\| g_{M^{(t)}}(x^{(t)}) \right\| (\varphi(x^{(0)}) - \varphi^*)^{\frac{2-\rho}{\rho}}. \end{aligned}$$

Therefore, we conclude that

$$\sigma(x_+^{(t)}, \varepsilon^{(t)}) = \frac{\theta \left\| g_{M^{(t)}}(x^{(t)}) \right\|}{(1+\sqrt{2}\beta)\operatorname{dist}(x_+^{(t)}, X^*)} \geq \frac{\theta}{1+\sqrt{2}\beta} \kappa^{\frac{2}{\rho}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-1} (\varphi(x^{(0)}) - \varphi^*)^{-\frac{2-\rho}{\rho}} = \bar{\sigma}.$$

This proves (36).

(iii) In the case $\sigma^{(0)} \leq \bar{\sigma}$, (ii) implies $\sigma^{(0)} \leq \varepsilon(x_+^{(0)}, \varepsilon^{(0)})$. Then, using (i), we have $\sigma^{(1)} = \sigma^{(0)}$ and also $\sigma^{(1)} \leq \bar{\sigma}$. Continuing this argument inductively, we conclude that $\sigma^{(t)} = \sigma^{(t-1)} = \cdots = \sigma^{(0)}$ for all t.

In the case $\sigma^{(0)} \geq \bar{\sigma}$, on the other hand, let us show $\sigma^{(t)} \geq \bar{\sigma}/\gamma_{\mathrm{reg}}$ $(t = 0, \ldots, T+1)$ by induction. The assertion for $t = 0$ is clear since $\sigma^{(0)} \geq \bar{\sigma} \geq \bar{\sigma}/\gamma_{\mathrm{reg}}$. Assume that $\sigma^{(t)} \geq \bar{\sigma}/\gamma_{\mathrm{reg}}$ holds for some $t \geq 0$. By (i) and (ii), we have

$$\sigma^{(t+1)} \geq \min\{\sigma^{(t)}, \varepsilon(x_+^{(t)}, \varepsilon^{(t)})/\gamma_{\mathrm{reg}}\} \geq \min\{\bar{\sigma}/\gamma_{\mathrm{reg}}, \bar{\sigma}/\gamma_{\mathrm{reg}}\} = \bar{\sigma}/\gamma_{\mathrm{reg}}.$$

This completes the proof of (iii). □

In order to provide more accurate complexity analysis, we prove bounds of $\sigma^{(t)}$ specialized to the case $\rho > 2$.

Lemma 4.9. *Assume that the Hölderian error bound condition (25) holds with $\rho > 2$. Suppose that* `rAdaAPG` *terminated with the stopping criterion at the $(T+1)$-th outer loop for some $T \geq 0$. If $\sigma^{(0)} \geq \bar{\sigma}$ holds for $\bar{\sigma}$ defined by (27), then there exists $t_0 \in \{0, \ldots, T+1\}$ such that the following conditions hold, where $\sigma_*^{(0)} := \frac{\theta}{1+\sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-\frac{1}{\rho-1}} \| g_{M^{(0)}}(x^{(0)}) \|_*^{\frac{\rho-2}{\rho-1}}$.*

$$\textit{(i)} \qquad \sigma^{(0)} = \cdots = \sigma^{(t_0)}. \tag{38}$$

$$\textit{(ii)} \qquad \sigma^{(t+1)} \geq \theta^{-\frac{\rho-2}{\rho-1}(T-t)} \bar{\sigma}/\gamma_{\mathrm{reg}}, \quad t = t_0, \ldots, T. \tag{39}$$

$$\textit{(iii)} \qquad t_0 \leq 1 + \frac{\rho-1}{\rho-2} \log_{1/\theta} \frac{\sigma_*^{(0)}}{\min(\sigma^{(0)}, \sigma_*^{(0)})}. \tag{40}$$

$$\textit{(iv)} \qquad \theta^{\frac{\rho-2}{\rho-1}(T-t_0)} \geq \bar{\sigma}/\sigma^{(0)}. \tag{41}$$

Proof. Define

$$\sigma_*^{(t)} := \frac{\theta^{\frac{1}{\rho-1}}}{1+\sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-\frac{1}{\rho-1}} (\varepsilon^{(t)})^{\frac{\rho-2}{\rho-1}}, \quad t \geq 0.$$

Then Lemma 4.8 (ii) can be written as

$$\sigma(x_+^{(t)}, \varepsilon^{(t)}) \geq \sigma_*^{(t)} \geq \bar{\sigma}, \quad t = 0, \ldots, T. \tag{42}$$

For simplicity, denote

$$\omega := \frac{\rho-2}{\rho-1} \in (0,1), \quad c := \frac{\theta^{\frac{1}{\rho-1}}}{1+\sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-\frac{1}{\rho-1}}$$

so that we have

$$\bar{\sigma} = c(\theta\varepsilon)^{\omega}, \quad \sigma_*^{(t)} = c(\varepsilon^{(t)})^{\omega}, \quad t \geq 0.$$

Note that $\{\sigma_*^{(t)}\}$ is non-increasing; in fact, the relation $\varepsilon^{(t+1)} \le \theta\varepsilon^{(t)}$ implies

$$\sigma_*^{(t+1)} = c(\varepsilon^{(t+1)})^\omega \le c\theta^\omega(\varepsilon^{(t)})^\omega = \theta^\omega\sigma_*^{(t)} \le \sigma_*^{(t)}. \tag{43}$$

Let t_0 be the smallest integer in $\{0,\dots,T+1\}$ such that $\sigma^{(0)} \ge \sigma_*^{(t_0)}$. Remark that, by the definition of T, we have $\varepsilon^{(T+1)} = \theta\|g_{M^{(T+1)}}(x^{(T+1)})\|_* \le \theta\varepsilon$. This implies that

$$\sigma^{(0)} \ge \bar\sigma = c(\theta\varepsilon)^\omega \ge c(\varepsilon^{(T+1)})^\omega = \sigma_*^{(T+1)}.$$

Therefore, t_0 is well-defined.

(i) By the definition of t_0 and (42), we have

$$\sigma^{(0)} < \sigma_*^{(t)} \le \sigma(x_+^{(t)}, \varepsilon^{(t)}), \quad 0 \le \forall t < t_0.$$

Therefore, by induction, we obtain $\sigma^{(0)} = \sigma^{(1)} = \cdots = \sigma^{(t_0)}$ due to Lemma 4.8 (i).

(ii) Let us show $\sigma^{(t+1)} \ge \sigma_*^{(t)}/\gamma_{\mathrm{reg}}$ $(t_0 \le t \le T)$. To prove this, we verity $\sigma^{(t)} \ge \sigma_*^{(t)}/\gamma_{\mathrm{reg}}$ for $t = t_0,\dots,T+1$ by induction. Note that $\sigma^{(t_0)} = \sigma^{(0)} \ge \sigma_*^{(t_0)} > \sigma_*^{(t_0)}/\gamma_{\mathrm{reg}}$ holds by (i) and the definition of t_0. Now under the hypothesis $\sigma^{(t)} \ge \sigma_*^{(t)}/\gamma_{\mathrm{reg}}$ for t with $t_0 \le t \le T$, Lemma 4.8 (i) and (42) imply

$$\sigma^{(t+1)} \ge \min\{\sigma^{(t)}, \sigma(x_+^{(t)},\varepsilon^{(t)})/\gamma_{\mathrm{reg}}\} \ge \min\{\sigma_*^{(t)}/\gamma_{\mathrm{reg}}, \sigma_*^{(t)}/\gamma_{\mathrm{reg}}\} = \sigma_*^{(t)}/\gamma_{\mathrm{reg}} \ge \sigma_*^{(t+1)}/\gamma_{\mathrm{reg}}.$$

Therefore, this completes the induction; in addition, the above inequality proves the desired inequality $\sigma^{(t+1)} \ge \sigma_*^{(t)}/\gamma_{\mathrm{reg}}$ $(t_0 \le t \le T)$. This yields (ii) combined with (42) and (43):

$$\sigma^{(t+1)} \ge \sigma_*^{(t)}/\gamma_{\mathrm{reg}} \ge \theta^{-\omega(T-t)}\sigma_*^{(T)}/\gamma_{\mathrm{reg}} \ge \theta^{-\omega(T-t)}\bar\sigma/\gamma_{\mathrm{reg}}, \quad t = t_0,\dots,T.$$

(iii) If $t_0 = 0$, then (iii) is trivial since $\sigma^{(0)} \ge \sigma_*^{(t_0)} = \sigma_*^{(0)}$. If $t_0 > 0$, then the definition of t_0 and using (43) imply

$$\sigma^{(0)} < \sigma_*^{(t_0-1)} \le \theta^{\omega(t_0-1)}\sigma_*^{(0)},$$

which yields $t_0 \le 1 + \frac{1}{\omega}\log_{1/\theta}\frac{\sigma_*^{(0)}}{\sigma^{(0)}}$.

(iv) By (42), (43), and the definition of t_0, remark that

$$\bar\sigma \le \sigma_*^{(T)} \le \theta^{\omega(T-t_0)}\sigma_*^{(t_0)} \le \theta^{\omega(T-t_0)}\sigma^{(0)}.$$

Hence, $\theta^{\omega(T-t_0)} \ge \bar\sigma/\sigma^{(0)}$ holds. □

Finally, we present the proofs of Theorem 4.4 and Corollary 4.5.

Proof of Theorem 4.4. We may assume that $\left\|g_{M^{(0)}}(x^{(0)})\right\| > \varepsilon$ since $N = 0$ on the other case. Suppose that `rAdaAPG` terminated with the stopping criterion at the $(T+1)$-th outer loop for some $T \ge 0$. Denote by $N^{(t)}$ the number of the executions of `APGIter` at the t-th outer loop so that $N = \sum_{t=0}^{T} N^{(t)}$. Let $\bar\sigma$ be defined by (27).

By the definition of T, we have

$$\left\|g_{M^{(t)}}(x^{(t)})\right\| > \varepsilon, \quad t = 0,\dots,T. \tag{44}$$

Moreover, by the definition of σ_* in (28), using Lemma 4.8 (iii) implies

$$\sigma^{(t)} \ge \sigma_*, \quad t = 0,\dots,T+1.$$

Therefore, applying Lemma 4.7 with $T_* = T$, we obtain the assertion

$$N \le N(\varepsilon, \sigma_*, C) \quad \text{with} \quad C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\left\| g_{M^{(0)}}(x^{(0)}) \right\|}{\varepsilon}\right). \tag{45}$$

For the case $\rho = 2$, (45) proves the assertion (i). We discuss the other cases to improve (45).

(ii) Case $\rho > 2$. If $\sigma^{(0)} < \bar{\sigma}$, then σ_* is defined as $\sigma_* = \sigma^{(0)}$. Therefore, the latter bound of (ii) is obtained by (45).

Now consider the case $\sigma^{(0)} \ge \bar{\sigma}$. Then, there exists $t_0 \in \{0, \dots, T+1\}$ satisfying the conditions in Lemma 4.9. By (41), remark that

$$\sum_{t=t_0}^{T} \sqrt{\theta^{\frac{\rho-2}{\rho-1}(T-t)}} = \sum_{i=0}^{T-t_0} \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}^{i} = \frac{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}^{T-t_0+1}}{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}} \le \frac{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}} \cdot \frac{\bar{\sigma}}{\sigma^{(0)}}}}{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}}. \tag{46}$$

Therefore, we conclude that

$$\begin{aligned}
\sum_{t=0}^{T} \sqrt{\frac{1}{\sigma^{(t+1)}}} &= \sum_{t=0}^{t_0-1} \sqrt{\frac{1}{\sigma^{(t+1)}}} + \sum_{t=t_0}^{T} \sqrt{\frac{1}{\sigma^{(t+1)}}} \\
&\le t_0 \sqrt{\frac{1}{\sigma^{(0)}}} + \sqrt{\frac{\gamma_{\mathrm{reg}}}{\bar{\sigma}}} \sum_{t=t_0}^{T} \sqrt{\theta^{\frac{\rho-2}{\rho-1}(T-t)}} && \text{(by (38) and (39))} \\
&\le \sqrt{\frac{1}{\sigma^{(0)}}} \left(1 + \frac{\rho-1}{\rho-2} \log_{1/\theta} \frac{\sigma_*^{(0)}}{\min(\sigma_*^{(0)}, \sigma^{(0)})}\right) \\
&\quad + \frac{\sqrt{\gamma_{\mathrm{reg}}}}{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}} \left(\sqrt{\frac{1}{\bar{\sigma}}} - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}} \sqrt{\frac{1}{\sigma^{(0)}}}\right) && \text{(by (40) and (46))} \\
&=: C.
\end{aligned}$$

With this definition of C, Lemma 4.7 gives $N \le N(\varepsilon, \sigma_*, C)$.

(iii) Case $\rho \in (1, 2)$. If $\varepsilon \ge \varepsilon_*$, then (45) gives our assertion because the second term of (29) vanishes. Suppose, on the other hand, that $\varepsilon < \varepsilon_*$. Denote

$$\xi_t := \left\| g_{M^{(t)}}(x^{(t)}) \right\|,$$

and let $T_* \ge 0$ be the smallest integer such that $\xi_{T_*} \le \varepsilon_*$. Then, since $\xi_{T_*-1} > \varepsilon_*$ holds, Lemma 4.7 shows that

$$\sum_{t=0}^{T_*-1} N^{(t)} \le N(\varepsilon_*, \sigma_*, C) \quad \text{with} \quad C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\left\| g_{M^{(0)}}(x^{(0)}) \right\|}{\varepsilon_*}\right).$$

Note that, under the convention $\sum_{t=0}^{-1}(\cdot) = 0$, this inequality also holds if $T_* = 0$ since then $\xi_0 \le \varepsilon_*$ and $N(\varepsilon_*, \sigma_*, C) = 0$.

It remains to observe $\sum_{t=T_*}^{T} N^{(t)}$. Take $t \in \{T_*, \dots, T\}$. We shall prove $N^{(t)} = 1$. In the t-th outer loop, consider the first iteration of the subroutine $\texttt{AdaAPG}(x_+^{(t)}, L^{(t)}, \sigma^{(t)}, \varepsilon^{(t)})$, which executes

$$\{x_1, \psi_1, M_0, L_1, A_1\} \leftarrow \texttt{APGIter}_{\sigma^{(t)}}(x_+^{(t)}, \psi_0, L^{(t)}, A_0),$$

where $\psi_0(x) = \frac{1}{2}\|x - x_+^{(t)}\|^2$ and $A_0 = 0$. Then Proposition 3.6 (v) implies

$$\|g_{M_0}(x_1)\| \le \left(2\sqrt{\frac{M_0 + \sigma^{(t)}}{A_1}} + \sigma^{(t)}\right) \operatorname{dist}(x_+^{(t)}, X^*).$$

Moreover, according to the equation at Line 5 in Algorithm 1, A_1 can be calculated as $A_1 = 2/M_0$. Now remark that, using $M_0 \le \gamma_{\mathrm{inc}} L_f$ (Proposition 3.6 (i)), we have

$$\begin{aligned} 2\sqrt{\frac{M_0 + \sigma^{(t)}}{A_1}} + \sigma^{(t)} &= 2\sqrt{\frac{M_0(M_0 + \sigma^{(t)})}{2}} + \sigma^{(t)} \le 2\sqrt{\frac{(M_0 + \sigma^{(t)})^2}{2}} + (M_0 + \sigma^{(t)}) \\ &= (1 + \sqrt{2})(M_0 + \sigma^{(t)}) \le (1 + \sqrt{2})(\gamma_{\mathrm{inc}} L_f + \sigma^{(t)}). \end{aligned}$$

Combining them and using (34)[4], we conclude that

$$\begin{aligned} \|g_{M_0}(x_1)\| &\le (1 + \sqrt{2})(\gamma_{\mathrm{inc}} L_f + \sigma^{(0)}) \operatorname{dist}(x_+^{(t)}, X^*) \\ &\le (1 + \sqrt{2})(\gamma_{\mathrm{inc}} L_f + \sigma^{(0)}) \left(\frac{1}{\kappa}\right)^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1\right)^{\frac{1}{\rho-1}} \xi_t^{\frac{1}{\rho-1}} \\ &= \theta \varepsilon_*^{\frac{\rho-2}{\rho-1}} \xi_t^{\frac{1}{\rho-1}} \\ &\le \theta \xi_t^{\frac{\rho-2}{\rho-1}} \xi_t^{\frac{1}{\rho-1}} = \theta \xi_t = \varepsilon^{(t)}, \end{aligned} \tag{47}$$

where the last inequality is due to $\xi_t \le \varepsilon^*$ for $t \ge T_*$ (and remark $\frac{\rho-2}{\rho-1} < 0$). This shows $N^{(t)} = 1$ and (47) yields the recurrence

$$\xi_{t+1} \le \theta \varepsilon_*^{\frac{\rho-2}{\rho-1}} \xi_t^{\frac{1}{\rho-1}}, \quad t = T_*, \dots, T.$$

Since $\frac{1}{\rho-1} > 1$, it reduces ξ_t superlinearly. In particular, solving this recurrence implies

$$\log \frac{\varepsilon_*}{\theta^{\frac{\rho-1}{2-\rho}} \xi_T} \ge \left(\frac{1}{\rho - 1}\right)^{T - T_*} \log \frac{\varepsilon_*}{\theta^{\frac{\rho-1}{2-\rho}} \xi_{T_*}}.$$

Since $\xi_T > \varepsilon$ and $\xi_{T_*} \le \varepsilon_*$, we obtain

$$\sum_{t=T_*}^{T} N^{(t)} = T - T_* + 1 \le 1 + \left(\log \frac{1}{\rho - 1}\right)^{-1} \left(\log\log \frac{\varepsilon_*}{\theta^{\frac{\rho-1}{2-\rho}} \varepsilon} - \log\log \frac{1}{\theta^{\frac{\rho-1}{2-\rho}}}\right).$$

Consequently, N is bounded as follows.

$$N = \sum_{t=0}^{T_*-1} N^{(t)} + \sum_{t=T_*}^{T} N^{(t)} \le N(\varepsilon_*, \sigma_*, C) + 1 + \left(\log \frac{1}{\rho - 1}\right)^{-1} \left(\log\log \frac{\varepsilon_*}{\theta^{\frac{\rho-1}{2-\rho}} \varepsilon} - \log\log \frac{1}{\theta^{\frac{\rho-1}{2-\rho}}}\right).$$

(iv) Case $\rho = 1$. We have $\sigma^{(t)} \ge \sigma_*$ for each $t = 0, \dots, T+1$, by Lemma 4.8 (iii). Moreover, Lemma 4.6 (vi) shows that, if $x_+^{(t)} \notin X^*$, then we have

$$1 \le \frac{1}{\kappa}\left(\frac{L_f}{L_{\min}} + 1\right) \left\|g_{M^{(t)}}(x^{(t)})\right\|, \quad \text{i.e.,} \quad \left\|g_{M^{(t)}}(x^{(t)})\right\| \ge \varepsilon_* := \kappa \left(\frac{L_f}{L_{\min}} + 1\right)^{-1}.$$

[4] Although (34) is asserted in the case $x_+^{(t)} \notin X^*$, it trivially holds if $x_+^{(t)} \in X^*$ *unless* $\rho = 1$.

In other words, the condition $\left\|g_{M^{(t)}}(x^{(t)})\right\| < \varepsilon_*$ must imply $x_+^{(t)} \in X^*$ which also yields $N^{(t)} = 1$ and $x^{(t+1)} = x_+^{(t)} \in X^*$ by Proposition 3.6 (v) (then the algorithm terminates at the $(t+1)$-th outer loop). Therefore, we have

$$\left\|g_{M^{(t)}}(x^{(t)})\right\| \geq \varepsilon_*, \quad 0 \leq t \leq T-1. \tag{48}$$

Now we consider two cases. If $\left\|g_{M^{(T)}}(x^{(T)})\right\| \geq \varepsilon_*$ holds, then combining with (44) yields $\left\|g_{M^{(T)}}(x^{(T)})\right\| \geq \max(\varepsilon, \varepsilon_*)$, from which Lemma 4.7 with $T_* = T$ concludes

$$N \leq N(\max(\varepsilon, \varepsilon_*), \sigma_*, C) \quad \text{with} \quad C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\left\|g_{M^{(0)}}(x^{(0)})\right\|}{\max(\varepsilon, \varepsilon_*)}\right).$$

On the other case $\left\|g_{M^{(T)}}(x^{(T)})\right\| < \varepsilon_*$, we have $N^{(T)} = 1$. Since (44) and (48) implies $\left\|g_{M^{(T-1)}}(x^{(T-1)})\right\| \geq \max(\varepsilon, \varepsilon_*)$, Lemma 4.7 with $T_* = T-1$ shows $\sum_{t=0}^{T-1} N^{(t)} \leq N(\max(\varepsilon, \varepsilon_*), \sigma_*, C)$. Hence,

$$N = N^{(T)} + \sum_{t=0}^{T-1} N^{(t)} \leq 1 + N(\max(\varepsilon, \varepsilon_*), \sigma_*, C).$$

This proves the desired bound on N. To show the latter assertion of (iv), suppose $\varepsilon < \varepsilon_*$. Then the output $\{x^{(T+1)}, M^{(T+1)}, x_+^{(T+1)}\}$ satisfies $\left\|g_{M^{(T+1)}}(x^{(T+1)})\right\| \leq \varepsilon < \varepsilon_*$. Therefore, $x_+^{(T+1)}$ must be optimal.

The proof of Theorem 4.4 is completed. □

Proof of Corollary 4.5. The function $N(\cdot, \cdot, \cdot)$ defined in (26) has the following expression.

$$N(\varepsilon, \sigma_*, C) = O\left(\log \frac{L_f + \sigma_*}{\sigma_*} \left[\log \frac{\|g_0\|}{\varepsilon} + \log \frac{\sigma^{(0)}}{\sigma_*} + \sqrt{\frac{L_f}{\sigma_*}} + C\sqrt{L_f}\right]\right).$$

By the choice (31) of $\sigma^{(0)}$, we can apply Corollary 4.3 and then the bound (23) becomes

$$\varepsilon(x_+^{(0)}, \varepsilon^{(0)}) \leq \sigma^{(0)} \leq \frac{2\gamma_{\text{inc}} L_f}{1 + \sqrt{2}\beta}. \tag{49}$$

Then, we have $\sigma^{(0)} \geq \bar{\sigma}$ since $\varepsilon(x_+^{(0)}, \varepsilon^{(0)}) \geq \bar{\sigma}$ holds by Lemma 4.8 (ii). Therefore, σ_* in (28) becomes $\sigma_* = \Theta(\bar{\sigma})$, which also implies $\sigma_* = O(\sigma^{(0)}) = O(L_f)$ combined with (49). Applying the bounds $\sigma^{(0)} = O(L_f)$, $\sigma_* = O(L_f)$, and $\sigma_* = \Omega(\bar{\sigma})$, we obtain

$$N(\varepsilon, \sigma_*, C) = O\left(\log \frac{L_f}{\bar{\sigma}} \left[\log \frac{\|g_0\|}{\varepsilon} + \sqrt{\frac{L_f}{\bar{\sigma}}} + C\sqrt{L_f}\right]\right). \tag{50}$$

If $\rho = 2$, Theorem 4.4 (i) implies $N \leq N(\varepsilon, \sigma_*, C)$ with $C = O\left(\sqrt{\frac{1}{\bar{\sigma}}} \log \frac{\|g_0\|}{\varepsilon}\right)$. In particular, (50) yields

$$N(\varepsilon, \sigma_*, C) = O\left(\sqrt{\frac{L_f}{\bar{\sigma}}} \log \frac{L_f}{\bar{\sigma}} \log \frac{\|g_0\|}{\varepsilon}\right). \tag{51}$$

Since $\bar{\sigma} = \Theta(\kappa)$ by (27), we conclude the bound (32) in the case $\rho = 2$.

In the case $\rho = 1$, using Theorem 4.4 (iv), the same argument as the case $\rho = 2$ can be applied to obtain (51) replacing ε by $\max(\varepsilon, \varepsilon_*)$, where $\varepsilon_* = \kappa(L_f/L_{\min} + 1)^{-1} = O(\kappa)$. Since $\bar{\sigma} = \Omega(\kappa^2/\Delta_0)$ by (27), we obtain the bound (32) in the case $\rho = 1$.

In the case $\rho \in (1,2)$, we apply Theorem 4.4 (iii) and the argument is similar to the previous cases. Therefore, the bound (32) in this case can be obtained based on the estimate (51) replacing ε by $\max(\varepsilon, \varepsilon_*)$ and applying $\bar{\sigma} = \Omega(\kappa^{\frac{2}{\rho}} \Delta_0^{-\frac{2-\rho}{\rho}})$, $\varepsilon_* = O(\kappa^{\frac{1}{2-\rho}} L_f^{-\frac{\rho-1}{2-\rho}})$.

Finally, consider the case $\rho > 2$. By Lemma 4.8 (ii) and (49), remark that $\sigma^{(0)} \geq \sigma_*^{(0)} \geq \bar{\sigma}$ holds. Then, Theorem 4.4 (ii) implies $N \leq N(\varepsilon, \sigma_*, C)$ with $C = O\left(\sqrt{\frac{1}{\sigma^{(0)}}} + \sqrt{\frac{1}{\bar{\sigma}}}\right) = O\left(\sqrt{\frac{1}{\bar{\sigma}}}\right)$. Therefore, (50) applying $\bar{\sigma} = \Omega(\kappa^{\frac{1}{\rho-1}} \varepsilon^{\frac{\rho-2}{\rho-1}})$ concludes the bound (32) in the case $\rho > 2$. □

5 Concluding remarks

In this paper, we proposed two adaptive proximal gradient methods, Algorithms 3 and 4. The former algorithm is nearly optimal for the class of problems where f is L-smooth. If we additionally assume the Hölderian error bound condition, the latter algorithm ensures near optimality. It is unclear whether the latter algorithm also provides the near optimality without the Hölderian error bound.

A remarkable fact of the proposed method (Algorithm 4) is the near optimal complexity with respect to the gradient norm under the Hölderian error bound condition, thanks to the lower complexity bound (13). Remark that the optimal complexity of the first-order methods for L-smooth convex functions under the gradient norm is unknown [16], namely, it is open whether we can reduce the logarithmic factor of the complexity bound (24). Similarly, it is an important question whether we can improve the complexity (32) to attain the lower bounds (13).

The key idea of this work is the adaptive determination of the regularization parameter σ used to define the regularization $\varphi_\sigma(x)$. As proved in Theorem 4.2, our method (Algorithm 3) adapts the unknown desired regularization parameter $\sigma(x_0, \varepsilon) = \frac{\varepsilon}{(1+\sqrt{2}\beta)\operatorname{dist}(x_0, X^*)}$. This feature is also critical for the development of the restart scheme (Algorithm 4) to adapt the Hölderian error bound condition. Basically, this adaption is obtained thanks to the relation between $\sigma(x_0, \varepsilon)$ (in other words, $\operatorname{dist}(x_0, X^*)$) and the "problem structure" denoted as $\bar{\sigma}$ (cf. Lemma 4.8 (ii)). This might suggest the possibility of dealing with this adaptive regularization approach under other kinds of problem structures.

Acknowledgements

This work was partially supported by the Grant-in-Aid for Young Scientists (B) (17K12645) and the Grant-in-Aid for Scientific Research (C) (18K11178) from Japan Society for the Promotion of Science.

References

[1] H. Attouch, J. Bolte, and B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, *Mathematical Programming*, **137**(1–2), pp. 91–129, 2013.

[2] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences*, **2**(1), pp. 183–202, 2009.

[3] J. Bolte, A. Daniilidis, and A. Lewis, The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM Journal on Optimization*, **17**(4), pp. 1205-1223, 2007.

[4] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, From error bounds to the complexity of first-order descent methods for convex functions, *Mathematical Programming*, **165**(2), pp. 471–507, 2017.

[5] O. Fercoq and Z. Qu, Adaptive restart of accelerated gradient methods under local quadratic growth condition, *ArXiv preprint*, arXiv:1709.02300v1, 2017.

[6] D. Kim and J. Fessler, Generalizing the optimized gradient method for smooth convex minimization, *SIAM Journal on Optimization*, **28**(2), pp. 1920–1950, 2018.

[7] Q. Lin and L. Xiao, An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization, *Computational Optimization and Applications*, **60**(3), pp. 633–674, 2015.

[8] M. Liu and T. Yang, Adaptive accelerated gradient converging methods under Hölderian error bound condition, *Advances in Neural Information Processing Systems 30*, 2017.

[9] S. Łojasiewicz, Sur le problème de la division, *Studia Mathematica*, **18**, pp. 87–136, 1959.

[10] S. Łojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, in *Les Équations aux Dérivées Partielles* (Éditions du Centre National de la Recherche Scientifique, Paris, 1963), pp. 87–89.

[11] S. Łojasiewicz, *Ensembles semi-analytiques*, preprint, Institut des Hautes Études Scientifiques, 1965.

[12] A. Nemirovsky, Information-based complexity of linear operator equations, *Journal of Complexity*, **8**, pp. 153–175, 1992.

[13] A. Nemirovsky and Y. Nesterov, Optimal methods for smooth convex minimization, *USSR Computational Mathematics and Mathematical Physics*, **25**(2), pp. 21–30, 1985.

[14] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, *Soviet Mathematics Doklady*, **27**(2), pp. 372–376, 1983.

[15] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, Kluwer Academic Publishers, Norwell, 2004.

[16] Y. Nesterov, How to make the gradients small, *Optima*, **88**, pp. 10–11, 2012.

[17] Y. Nesterov, Gradient methods for minimizing composite functions, *Mathematical Programming*, **140**(1), pp. 125–161, 2013.

[18] Y. Nesterov, Universal gradient methods for convex optimization problems, *Mathematical Programming*, **152**(1–2), pp. 381–404, 2015.

[19] N. Parikh and S. Boyd, Proximal algorithms, *Foundations and Trends in Optimization*, **1**(3), pp. 123–231, 2014.

[20] J. Renegar and B. Grimmer, A simple nearly-optimal restart scheme for speeding-up first order methods, *ArXiv preprint*, arXiv:1803.00151v1, 2018.

[21] R. T. Rockafellar, *Convex analysis*, Princeton University Press, New Jersey, 1970.

[22] V. Roulet and A. d'Aspremont, Sharpness, restart and acceleration, *Advances in Neural Information Processing Systems 30*, 2017.

[23] A. B. Taylor, J. M. Hendrickx, and F. Glineur, Smooth strongly convex interpolation and exact worst-case performance of first-order methods, *Mathematical Programming*, **161**(1–2), pp. 307–345, 2017.