

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

Masaru Ito* Mituhiro Fukuda†

October 2020

Abstract

In the development of first-order methods for smooth (resp., composite) convex optimization problems, where smooth functions with Lipschitz continuous gradients are minimized, the gradient (resp., gradient mapping) norm becomes a fundamental optimality measure. Under this measure, a fixed iteration algorithm with the optimal iteration complexity is known, while determining this number of iteration to obtain a desired accuracy requires the prior knowledge of the distance from the initial point to the optimal solution set. In this paper, we report an adaptive regularization approach, which attains the nearly optimal iteration complexity without knowing the distance to the optimal solution set. 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 classes 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 the composite convex optimization problems, which consists in minimizing the sum of a differentiable convex function with Lipschitz continuous gradient and a (possibly non-smooth) convex function with a “simple” structure (called the regularizer). Accelerated first-order methods for this class of problems have been well-studied as they provide optimal iteration complexity to obtain an approximate solution under the measure of the *functional residual*, i.e., the gap between the objective function value at an iterate and the optimal value, for various classes of problems [1–6].

A major interest we focus on in this paper is to consider the iteration complexity under the measure of the norm of the *gradient mapping*. The gradient mapping generalizes the gradient in the smooth case (i.e., the case when there is no regularizer) to the composite case, and the gradient mapping vanishes at a point if and only if that point is an optimal solution. The norm of the gradient mapping

*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).

is a useful optimality measure as it is computable at each iteration (if the proximal mapping of the regularizer is computable), while the functional residual itself is not verifiable if we do not know the optimal value. Remark that, in some cases, we are able to provide computable upper bounds on the functional residual, such as the duality gap [7–9], and the convergence is guaranteed for that bounds.

To find an approximate solution under the gradient norm measure, a first-order method with optimal iteration complexity for the smooth case was recently announced by Nesterov et al. (see [10, Remark 2.1] and [11, Remark 3.1]). This is a fixed iteration algorithm and, in order to determine the number of iterations to attain a desired accuracy, it requires the prior knowledge of the distance, say D , from the initial point to the optimal solution set. Without the two requirements of fixing the number of iterations and knowing the distance D , variants of accelerated first-order methods (see, e.g., [12, 13]) seem to be the best known ones, however their iteration complexity results become worse. One aim of this paper is to provide an algorithm (Algorithm 3), that attains the nearly optimal iteration complexity without these two requirements, as shown in Corollary 4.3. This result is attained by an arrangement of the regularization technique [8], which applies the accelerated first-order method to the regularized auxiliary problem adding a quadratic regularizer to the original objective function. Our algorithm ensures the same iteration complexity as the original regularization technique [8], with an additional advantage that we do not need the distance D .

Another motivation in this paper is the development of an adaptive first-order method under the additional assumption, the *Hölderian Error Bound* (HEB) condition, in order to provide better convergence guarantees as the linear convergence for strongly convex functions. This condition is also related to the concept called the Lojasiewicz gradient inequality [14, 15], which is a useful tool for the development and the analysis of algorithms as well as first-order methods [16, 17]. These concepts are known to be satisfied under moderately mild assumptions such as when the objective function is coercive and subanalytic (in particular, semi-algebraic) [18]. The HEB condition involves two parameters, the coefficient κ and the exponent ρ , which are critical to determine the convergence rate, although they are difficult to estimate in general. Therefore, the development of adaptive algorithms is an important issue.

Under the HEB condition, we propose Algorithm 4, a restart scheme of the previous mentioned adaptive regularization algorithm. Our method is inspired by the Liu-Yang’s method [19] as we employ an (approximate) proximal-point approach, where the main difference is the adaption parameter: Liu-Yang’s method adaptively estimates the coefficient κ , but requires the prior knowledge of the exponent ρ , while our method adaptively determines the regularization parameter to define the regularized auxiliary problem. As a result, without knowing the coefficient κ and the exponent ρ , the proposed method adapts both the parameters κ and ρ . Under the measure of the gradient mapping norm, the proposed method ensures the same iteration complexity as the one by Liu and Yang [19], dismissing the prior knowledge of ρ (Corollary 4.5). In addition, we show the finite or the superlinear convergence when the exponent ρ is less than two, which were not shown in the existing adaptive methods [5, 19]. Furthermore, for the smooth case, we prove that the proposed method attains the nearly optimal iteration complexity with respect to the gradient norm. The same assertion can also be derived with respect to the measure of the functional residual.

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 [5, 6] are applicable to our problem (for a specific regularizer) and they adapt both κ and ρ ensuring the nearly optimal iteration complexity under the measure of the functional residual. One advantage of our method compared with these methods is the (nearly optimal) *convergence* guarantee; the method of Roulet and d’Aspremont (see [6, Proposition 3.4]) is a fixed step algorithm (remark that, if we know the optimal value, [6, Algorithm 3] gives nearly optimal convergence), and Renegar-

Grimmer’s method [5] fixes the target tolerance. Although the other algorithms, i.e., this work and the first four algorithms [3, 19–21] in Table 1, terminate if a desired approximate solution is found, they provide the nearly optimal convergence by omitting the stopping criterion as we discuss later.

This paper is organized as follows. Section 2 collects preliminary facts on the gradient mapping and the HEB 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 HEB condition. We review in Section 3 the regularization technique [12] preparing auxiliary results. Two adaptive first-order methods and the main results for them are established in Section 4. In Section 4.1, we show an adaptive regularization algorithm and prove the nearly optimal iteration complexity for the class of smooth convex functions. 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 HEB condition. Concluding remarks are given in Section 5.

Table 1: Adaptive first-order methods. The column ‘Problem class’ indicates the class of problems to which the algorithm is applicable. The column ‘Optimality measures’ is the optimality measure for which the nearly optimal iteration complexity was proved. The column ‘Nearly optimal convergence’ shows whether the algorithm ensures the convergence of the optimality measure to zero at the nearly optimal rate (while all the algorithms attain the nearly optimal iteration complexity for a fixed tolerance).

Algorithm	Problem class	Optimality measures	Nearly optimal convergence
Nesterov [3] Lin and Xiao [21]	μ -strongly convex φ with unknown μ	gradient mapping norm	yes
Fercoq and Qu [20]	HEB with $\rho = 2$ and unknown κ	gradient mapping norm	yes
Liu and Yang [19]	HEB with known ρ and unknown κ	gradient mapping norm	yes
Roulet and d’Aspremont [6, Proposition 3.4]	HEB with unknown κ and ρ	functional residual	only for predefined number of iterations
Reneger and Grimmer [5]	HEB with unknown κ and ρ	functional residual	only for predefined tolerance
This work (Algorithm 4)	HEB with unknown κ and ρ	gradient mapping norm (and functional residual)	yes

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} . The distance between a point $x \in \mathbb{E}$ and a set $S \subset \mathbb{E}$ is defined by $\text{dist}(x, S) = \inf_{z \in S} \|z - x\|$.

We focus on the the convex composite optimization problem

$$\varphi^* = \min_{x \in \mathbb{E}} [\varphi(x) \equiv f(x) + \Psi(x)], \quad (1)$$

where $f : \mathbb{E} \rightarrow \mathbb{R}$ is an L_f -smooth convex function, that is, it is continuously differentiable and its gradient is L_f -Lipschitz continuous:

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

and $\Psi : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ is a proper, lower-semicontinuous, and convex function. We denote by $X^* := \text{Argmin}_{x \in \mathbb{E}} \varphi(x)$ the set of optimal solutions of $\min_{x \in \mathbb{E}} \varphi(x)$. Throughout the paper, we assume that X^* is nonempty. 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

$$m_L(y; x) := f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + \Psi(x), \quad T_L(y) := \underset{x \in \mathbb{E}}{\text{argmin}} m_L(y; x),$$

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, the proximal mapping $\text{prox}_\Psi(y) := \text{argmin}_{x \in \mathbb{E}} \{\Psi(x) + \frac{1}{2} \|x - y\|^2\}$ as well as the map $T_L(y) \equiv \text{prox}_{\Psi/L}(y - \nabla f(x)/L)$ is moderately computable (see [23] for examples). The *gradient mapping* of φ is defined by

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

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 [1, 3, 22].

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

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

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

Proof. (i) is proved in [3, Lemma 2].

(ii) According to the optimality condition of the problem $\min_{x \in \mathbb{E}} m_L(y; x)$, we have the following equivalence for $y, z \in \mathbb{E}$:

$$z = T_L(y) \iff 0 \in \nabla f(y) + L(z - y) + \partial\Psi(z). \quad (2)$$

On the other hand, the optimality of the original problem $\min_x \varphi(x)$ is characterized as follows. Given $w \in \mathbb{E}$, we have $w \in \text{Argmin}_{x \in \mathbb{E}} \varphi(x)$ if and only if $0 \in \nabla f(w) + \partial\Psi(w)$ holds (Remark that we have $\partial\varphi = \nabla f + \partial\Psi$ [24, Theorem 23.8]). Let us show the equivalence

$$x^* \in \underset{x \in \mathbb{E}}{\text{Argmin}} \varphi(x) \iff x^* = T_L(x^*) \quad (\iff g_L(x^*) = 0). \quad (3)$$

If $x^* = T_L(x^*)$, then (2) implies $0 \in \nabla f(x^*) + \partial\Psi(x^*)$ and so $x^* \in \text{Argmin}_{x \in \mathbb{E}} \varphi(x)$ holds. Conversely, if $x^* \in \text{Argmin}_{x \in \mathbb{E}} \varphi(x)$, then we have the condition $0 \in \nabla f(x^*) + \partial\Psi(x^*)$, which coincides with the right hand one of (2) in the case $y = z = x^*$. Thus, we conclude its equivalent counterpart $x^* = T_L(x^*)$. This proves the equivalence (3) as well as the assertion (ii).

(iii) By the condition (2), we have $0 \in \nabla f(y) + L(T_L(y) - y) + \partial\Psi(T_L(y)) = \nabla f(y) - g_L(y) + \partial\Psi(T_L(y))$. Hence, we obtain $\nabla f(T_L(y)) - \nabla f(y) + g_L(y) \in \nabla f(T_L(y)) + \partial\Psi(T_L(y)) = \partial\varphi(T_L(y))$, giving the former assertion of (iii). Now if f is L_f -smooth, we have $\|\nabla f(T_L(y)) - \nabla f(y)\| \leq L_f \|T_L(y) - y\| = \frac{L_f}{L} \|g_L(y)\|$, which yields the latter assertion of (iii).

(iv) It is shown in [1, Lemma 2.3] that, if $\varphi(T_L(y)) \leq m_L(y; T_L(y))$ holds, then we have

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

Letting $x := y$ shows the first assertion. On the other hand, since $\varphi(T_L(y)) \geq \varphi^*$, letting $x := x^* \in X^*$ gives $\frac{1}{2L} \|g_L(y)\|^2 \leq \langle g_L(y), y - x^* \rangle \leq \|g_L(y)\| \|y - x^*\|$. Thus, we obtain $\frac{1}{2L} \|g_L(y)\| \leq \text{dist}(y, X^*)$.

(v) Note that f is L -smooth for any $L \geq L_f$. Since L -smooth functions satisfy (e.g., see [22])

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

we obtain $\varphi(x) \leq m_L(y; x)$ for all $x, y \in \mathbb{E}$ and $L \geq L_f$. \square

2.1 Hölderian Error Bound

Here, we introduce the Hölderian error bound condition, which is also discussed in the context of Lojasiewicz inequality [15, 18, 25].

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 $\rho \geq 1$ if

$$\varphi(x) - \varphi^* \geq \kappa \text{dist}(x, X^*)^\rho, \quad \forall x \in \text{lev}_\varphi(\varphi(x_0)), \quad (4)$$

where $\text{lev}_\varphi(\gamma) = \{x \in \mathbb{E} \mid \varphi(x) \leq \gamma\}$.

According to [18, 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., [17, 19] for related studies.

A noteworthy concept related to the Hölderian error bound condition is the *Lojasiewicz gradient inequality* [14, 15], which is of the form

$$\text{dist}(0, \partial\varphi(x)) \geq \lambda(\varphi(x) - \varphi^*)^\theta, \quad \forall x \in \text{lev}_\varphi(\varphi(x_0)) \setminus X^*, \quad (5)$$

for $\lambda > 0$ and $\theta \in [0, 1[$. In fact, these concepts (4) and (5) are equivalent for convex functions (see Remark 2.4). The Lojasiewicz gradient inequality is a powerful tool for analyzing the convergence of first-order methods as it covers wide class of applications and algorithms [16, 17].

Lemma 2.3. Let $\varphi : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, lower-semicontinuous, and convex function. For $x_0 \in \mathbb{E}$, suppose that φ satisfies the Hölderian error bound condition (4) with coefficient $\kappa > 0$ and exponent $\rho \geq 1$. Then, for every $x \in \text{lev}_\varphi(\varphi(x_0)) \setminus X^*$, we have the inequalities

$$\kappa \text{dist}(x, X^*)^{\rho-1} \leq \inf\{\|g\| : g \in \partial\varphi(x)\}, \quad \kappa^{\frac{1}{\rho}}(\varphi(x) - \varphi^*)^{\frac{\rho-1}{\rho}} \leq \inf\{\|g\| : g \in \partial\varphi(x)\}. \quad (6)$$

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

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

It involves two inequalities $\kappa \text{dist}(x, X^*)^\rho \leq \|g\| \text{dist}(x, X^*)$ and $\varphi(x) - \varphi^* \leq \|g\| \frac{1}{\kappa^{1/\rho}} (\varphi(x) - \varphi^*)^{1/\rho}$. Since $x \notin X^*$, the assertion follows dividing them by $\text{dist}(x, X^*)$ and $\kappa^{-1/\rho} (\varphi(x) - \varphi^*)^{1/\rho}$, respectively. \square

Remark 2.4. The latter condition in (6) is the Łojasiewicz gradient inequality (5) with the correspondence $\lambda = \kappa^{\frac{1}{\rho}}$ and $\theta = \frac{\rho-1}{\rho} \in [0, 1[$. It is shown in [17] that the Łojasiewicz gradient inequality is essentially equivalent to the Hölderian error bound condition: If (5) holds with $\lambda = \rho\kappa^{\frac{1}{\rho}}$ and $\theta = \frac{\rho-1}{\rho}$, then (4) holds (In [17, Theorem 5 (i)], set $(\kappa^{-\frac{1}{\rho}}s^{\frac{1}{\rho}}, \varphi(x_0))$ in place of $(\varphi(s), r_0)$ and let the radius ρ of $B(\bar{x}, \rho)$ to $+\infty$). \square

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 $g_L(x) = \nabla\varphi(x)$; we assume this fact only at this subsection.

Given a class \mathcal{F} of objective functions and an optimality measure $\delta : \mathcal{F} \times \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$, 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 $\text{dist}(x_0, X^*) \leq R$.
- 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 (4).

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 two optimality measures, the functional residual $\delta_{\text{F}}(\varphi, x) := \varphi(x) - \inf \varphi$ and the gradient norm $\delta_{\text{G}}(\varphi, x) := \|\nabla\varphi(x)\|$.

For the class $\mathcal{F}(x_0, R, L)$, the following lower bound on the iteration complexity holds (by [26, Section 2.3B] applied to the class of L -smooth convex quadratic minimization):

$$C(\mathcal{F}(x_0, R, L), \delta_{\text{G}}; \varepsilon) = \Omega \left(\min \left\{ \dim \mathbb{E}, \sqrt{\frac{LR}{\varepsilon}} \right\} \right). \quad (7)$$

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*

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

¹ Suppose that φ is an L -smooth convex function satisfying (4) for some exponent $\rho \in [1, 2[$. For $\rho = 1$, Lemma 2.3 implies $\|\nabla\varphi(x)\| \geq \kappa$ for $x \notin \text{Argmin} \varphi$. If $\rho \in]1, 2[$, on the other hand, Lemmas 2.1 and 2.3 imply $\frac{1}{2L} \|\nabla\varphi(x)\|^2 \leq \varphi(x) - \varphi^* \leq \kappa^{-\frac{1}{\rho-1}} \|\nabla\varphi(x)\|^{\frac{\rho}{\rho-1}}$ for all x , which yields $\|\nabla\varphi(x)\| \geq \text{const.}$ for all $x \notin \text{Argmin} \varphi$. This contradicts to the continuity of $\nabla\varphi$ at points in $\text{Argmin} \varphi$.

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

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

Therefore, it follows $C(\mathcal{M}, \varphi, \delta_G; \varepsilon) \geq C(\mathcal{M}, \varphi, \delta_F; \varepsilon^*)$ and we obtain the assertion. \square

Under the measure δ_F , the following lower bound is known [27]:

$$C(\mathcal{F}(x_0, R, L, \kappa, \rho), \delta_F; \varepsilon) = \begin{cases} \Omega \left(\min \left\{ \dim \mathbb{E}, \sqrt{\frac{L}{\kappa^{\frac{2}{\rho}} \varepsilon^{\frac{\rho-2}{\rho}}}} \right\} \right) & : \rho > 2, \\ \Omega \left(\min \left\{ \dim \mathbb{E}, \sqrt{\frac{L}{\kappa} \log \frac{\kappa R^2}{\varepsilon}} \right\} \right) & : \rho = 2. \end{cases} \quad (8)$$

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

$$C(\mathcal{F}(x_0, R, L, \kappa, \rho), \delta_G; \varepsilon) = \begin{cases} \Omega \left(\min \left\{ \dim \mathbb{E}, \sqrt{\frac{L}{\kappa^{\frac{1}{\rho-1}} \varepsilon^{\frac{\rho-2}{\rho-1}}}} \right\} \right) & : \rho > 2, \\ \Omega \left(\min \left\{ \dim \mathbb{E}, \sqrt{\frac{L}{\kappa} \log \frac{\kappa R}{\varepsilon}} \right\} \right) & : \rho = 2. \end{cases} \quad (9)$$

3 Accelerated Proximal Gradient Method Applied to Regularized Problems

This section is devoted to review the regularization strategy [12, 19], from which we construct adaptive methods in the next section. We consider to apply an accelerated proximal gradient method to the regularized problem

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

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:

$$f_\sigma(x) := f(x) + \frac{\sigma}{2} \|x - x_0\|^2, \quad m_L^\sigma(y; x) := f_\sigma(y) + \langle \nabla f_\sigma(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + \Psi(x),$$

$$T_L^\sigma(y) := \operatorname{argmin}_{x \in \mathbb{E}} m_L^\sigma(y; x), \quad g_L^\sigma(y) := L(y - T_L^\sigma(y)).$$

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

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

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

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

Proof. (i) is immediate by $\varphi(x) \leq \varphi_\sigma(x)$ and $\varphi_\sigma(x_0) = \varphi(x_0)$.

(ii) For any $x^* \in X^*$, the strong convexity of φ_σ implies $\varphi_\sigma(x^*) \geq \varphi_\sigma^* + \frac{\sigma}{2} \|x^* - x_\sigma^*\|^2$, which can be rewritten as $\varphi(x^*) - \varphi(x_\sigma^*) \geq \frac{\sigma}{2} (-\|x^* - x_0\|^2 + \|x_\sigma^* - x_0\|^2 + \|x^* - x_\sigma^*\|^2)$. 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 $a, b \in \mathbb{E}$ that²

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

This fact implies $\|T_L(y) - T_L^\sigma(y)\| \leq \frac{\|\nabla f(y) - \nabla f_\sigma(y)\|}{L} = \frac{\sigma \|y - x_0\|}{L}$. Therefore, we conclude (iii) because of $\|g_L(y) - g_L^\sigma(y)\| = L \|T_L(y) - T_L^\sigma(y)\|$. \square

3.1 Accelerated Proximal Gradient Method

We employ Nesterov's accelerated proximal gradient method [3] for solving $\min_{x \in E} \varphi_\sigma(x)$, by regarding the objective function as

$$\varphi_\sigma(x) = f(x) + \Psi_\sigma(x), \quad \Psi_\sigma(x) := \Psi(x) + \frac{\sigma}{2} \|x - x_0\|^2. \quad (10)$$

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

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

By the identity $m_{L+\sigma}^\sigma(y; x) = f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + \Psi_\sigma(x)$, observe that $\tilde{T}_L(y) = T_{L+\sigma}^\sigma(y)$ holds. Therefore, the accelerated method $\mathcal{A}(x_0, L_0, \sigma)$ given in [3, Algorithm (4.9)] applied to the regularized function (10) 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*

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

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 $x \in \{z, T_L(z), T_{L+\sigma}^\sigma(z)\}$), two evaluations of $\nabla f(x)$ (at $x \in \{y, z\}$), and three proximal operations $T_{L+\sigma}^\sigma(y), T_{L+\sigma}^\sigma(z), T_L(z)$. 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

$$v_k = \operatorname{prox}_{\gamma_k \Psi}(w_k), \quad \text{where } \gamma_k = \frac{A_k}{1 + \sigma A_k}, \quad w_k = x_0 - \frac{1}{1 + \sigma A_k} \sum_{i=1}^k a_i \nabla f(x_i). \quad (12)$$

\square

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

$$\frac{\mu}{2} \|x_a^* - x_b^*\|^2 \leq [\langle a, x_b^* \rangle + h(x_b^*)] - [\langle a, x_a^* \rangle + h(x_a^*)] \quad \text{and} \quad \frac{\mu}{2} \|x_a^* - x_b^*\|^2 \leq [\langle b, x_a^* \rangle + h(x_a^*)] - [\langle b, x_b^* \rangle + h(x_b^*)],$$

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

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

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

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

- 1: Compute $v_k := \operatorname{argmin}_{x \in \mathbb{E}} \psi_k(x)$ (cf. (12)).
- 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 $z = T_{L+\sigma}^\sigma(y) = \operatorname{argmin}_x [f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 + \Psi_\sigma(x)]$.
- 8: Compute $T_L(z)$ and $T_{L+\sigma}^\sigma(z)$.
- 9: Test the conditions

$$\langle \nabla f(y) - \nabla f(T_{L+\sigma}^\sigma(y)), y - T_{L+\sigma}^\sigma(y) \rangle \geq \frac{1}{L} \|\nabla f(y) - \nabla f(T_{L+\sigma}^\sigma(y))\|^2, \quad (11a)$$

$$\varphi_\sigma(T_{L+\sigma}^\sigma(z)) \leq m_{L+\sigma}^\sigma(z; T_{L+\sigma}^\sigma(z)), \quad (11b)$$

$$\text{[Optional]} \quad \varphi(T_L(z)) \leq \varphi(z). \quad (11c)$$

10: **until** the conditions in (11) hold.

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

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 [3], which affects Lemma 3.5. Moreover, we have additional computations of $T_L(z)$ and $T_{L+\sigma}^\sigma(z)$ at Line 8 in order to test the conditions (11b) and (11c). Remark that, when $L = L_f$ is known, the computation of $T_{L+\sigma}^\sigma(z)$ at Line 8 can be omitted as it is only used to check the second condition (11b). We let the third criterion (11c) optional; it is independent of the analysis of Algorithm 3, while we need it in Algorithm 4. The first condition (11a) is equivalent to the one in Nesterov's method (see the proof of Lemma 5 in [3]). It will be verified in Lemmas 3.4 and 3.5 that our modification does not affect the original complexity analysis. \square

Lemma 3.4. *The condition (11) of APGIter holds whenever $L \geq L_f$.*

Proof. The first condition (11a) is satisfied since f is L -smooth (e.g., see [22, Theorem 2.1.5]). Since f_σ is $(L_f + \sigma)$ -smooth, the second one (11b) holds if $L \geq L_f$ by Lemma 2.1 (v). The third one (11c) 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

$$m_L(z; T_L(z)) = \min_{x \in \mathbb{E}} \left\{ f(z) + \langle \nabla f(z), x - z \rangle + \Psi(x) + \frac{L}{2} \|x - z\|^2 \right\} \stackrel{x=z}{\leq} f(z) + \Psi(z) = \varphi(z). \quad (13)$$

\square

The following lemma is given in [3, Lemma 6], while we rewrite its proof due to the difference in the update of L_{k+1} .

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

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

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

Now, the relation $M_k = L_k \gamma_{\text{inc}}^{n_k-1}$ implies $L_{k+1} = \max\{L_{\min}, M_k / \gamma_{\text{dec}}\} \geq \frac{1}{\gamma_{\text{dec}}} L_k \gamma_{\text{inc}}^{n_k-1}$. Then, we have $(n_k - 1) \log \gamma_{\text{inc}} \leq \log \frac{\gamma_{\text{dec}} L_{k+1}}{L_k} = \log \gamma_{\text{dec}} + \log \frac{L_{k+1}}{L_k}$ and therefore $n_k \leq 1 + \frac{\log \gamma_{\text{dec}}}{\log \gamma_{\text{inc}}} + \frac{1}{\log \gamma_{\text{inc}}} \log \frac{L_{k+1}}{L_k}$. \square

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 (10), that is,*

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

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

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

(ii) *For each $k \geq 1$, we have $A_k \geq \frac{2}{\gamma_{\text{inc}} L_f} \left[1 + \sqrt{\frac{\sigma}{2\gamma_{\text{inc}} L_f}} \right]^{2(k-1)}$.*

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

(iv) *$\varphi(x_k) \leq \varphi(x_0)$ holds for all $k \geq 1$.*

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

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

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

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

Since Ψ_σ is σ -strongly convex, (ii) follows by [3, Lemma 8] applied to the objective function $\varphi_\sigma = f + \Psi_\sigma$.

According to [3, Lemma 7], the following relations hold for all $k \geq 0$:

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

Combining them, we obtain

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

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

(v) By the σ -strong convexity of φ_σ and using (iii), we have $\frac{\sigma}{2} \|x_k - x_\sigma^*\|^2 \leq \varphi_\sigma(x_k) - \varphi_\sigma^* \leq \frac{\|x_0 - x_\sigma^*\|^2}{2A_k}$, and thus $\|x_k - x_\sigma^*\| \leq \frac{1}{\sqrt{\sigma A_k}} \|x_0 - x_\sigma^*\|$. This shows the former inequality of (v):

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

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

Since $\varphi_\sigma(T_{M_{k-1}+\sigma}^\sigma(x_k)) \leq m_{M_{k-1}+\sigma}^\sigma(x_k; T_{M_{k-1}+\sigma}^\sigma(x_k))$ holds by (11b), Lemma 2.1 (iv) yields

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

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

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

□

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 [3, 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 a proximal operation $T_L(x_k)$, for each guess L of the Lipschitz constant.

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

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

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

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

Proof. Since the condition (14) ensures $\varphi(T_{M_k}(x_k)) \leq m_{M_k}(x_k; T_{M_k}(x_k)) \leq \varphi(x_k)$ (recall (13) for the second inequality), we conclude (i). The assertion (ii) can be verified in the same way as Lemma 3.5. □

Algorithm 2: Proximal Gradient Iteration

$$\{T_{M_k}(x_k), M_k, L_{k+1}\} \leftarrow \text{PGIter}(x_k, L_k)$$

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

Input: $x_k \in \mathbb{E}$, $L_k > 0$.

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

$$\varphi(T_L(x_k)) \leq m_L(x_k; T_L(x_k)). \tag{14}$$

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

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

Algorithm 3: Adaptive Accelerated Proximal Gradient Method

$$\text{AdaAPG}(x_0, L_{-1}, \sigma_0, \varepsilon)$$

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

Input: $x_0 \in \mathbb{E}$, $L_{-1} \in [L_{\min}, \gamma_{\text{inc}}L_f]$, $\sigma_0 > 0$, $\varepsilon > 0$.

- 1: $L_0 := L_{-1}$.
 - 2: **for** $j = 0, 1, 2, \dots$ **do**
 - 3: $\sigma_j := \sigma_0/\gamma_{\text{reg}}^j$.
 - 4: $\psi_0(x) := \frac{1}{2} \|x - x_0\|^2$, $A_0 := 0$.
 - 5: **repeat for** $k = 0, 1, 2, \dots$
 - 6: $\{x_{k+1}, \psi_{k+1}, M_k, L_{k+1}, A_{k+1}\} \leftarrow \text{APGIter}_{\sigma_j}(x_k, \psi_k, L_k, A_k)$.
 - 7: **if** $\|g_{M_k}(x_{k+1})\| \leq \varepsilon$ **then**
 - 8: output $\{\sigma_j, x_{k+1}, T_{M_k}(x_{k+1}), M_k, L_{k+1}\}$ and terminate the algorithm.
 - 9: **end if**
 - 10: **until** $A_{k+1} \geq \frac{2(M_k + \sigma_j)}{\beta^2 \sigma_j^2}$.
 - 11: $L_0 := L_{k+1}$.
 - 12: **end for**
-

For solving (1) 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 [12] 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 $\sigma_j = \sigma_0/\gamma_{\text{reg}}^j$ ($\gamma_{\text{reg}} > 1$). We stop Nesterov's method when it successfully finds an ε -solution or when 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 $\sigma_{j+1} := \sigma_j/\gamma_{\text{reg}}$. The proposed method involves the parameter $\beta \in]0, 1]$, which controls the timing when σ_j should be decreased; smaller β imposes larger A_{k+1} in the condition at Line 10, and then we examine the current guess σ_j for a longer time.

The next lemma shows what happens if an outer loop is completed without the termination of the algorithm.

Lemma 4.1. *In AdaAPG, suppose that a j -th outer loop is completed satisfying 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 the inequalities*

$$\|x_{k+1} - x_0\| \leq \left(1 + \frac{\beta}{\sqrt{2}}\right) \text{dist}(x_0, X^*), \quad \|g_{M_k}(x_{k+1})\| \leq (1 + \sqrt{2}\beta)\sigma_j \text{dist}(x_0, X^*).$$

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

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

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

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

On the other hand, Proposition 3.6 (ii) implies

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

where the second inequality is due to the fact³ $1 + x \geq \exp \frac{x}{1+x}$ ($\forall x > -1$). Therefore,

$$\begin{aligned} \frac{2}{\gamma_{\text{inc}}L_f} \exp\left(2(k-1)\frac{1}{\sqrt{\frac{2\gamma_{\text{inc}}L_f}{\sigma_j} + 1}}\right) &\leq \frac{2(\gamma_{\text{inc}}L_f + \sigma_j)}{\beta^2 \sigma_j^2} \\ \implies k + 1 &\leq 2 + \frac{1}{2} \left(\sqrt{\frac{2\gamma_{\text{inc}}L_f}{\sigma_j} + 1}\right) \log \frac{\gamma_{\text{inc}}L_f(\gamma_{\text{inc}}L_f + \sigma_j)}{\beta^2 \sigma_j^2}. \end{aligned}$$

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

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

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

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 $\min_{x > -1} h(x) = h(0) = 0$.

Theorem 4.2. In AdaAPG, let N be the total number of executions of APGIter. Denote

$$\sigma(x_0, \varepsilon) := \frac{\varepsilon}{(1 + \sqrt{2}\beta) \text{dist}(x_0, X^*)}, \quad (15)$$

where we let $\sigma(x_0, \varepsilon) := +\infty$ when $x_0 \in X^*$. Then, the following assertions hold.

(i) The algorithm terminates at the j -th outer loop whenever $\sigma_j \leq \sigma(x_0, \varepsilon)$. Moreover, we have $\sigma_j \geq \sigma(x_0, \varepsilon)/\gamma_{\text{reg}}$ for all $j > 0$.

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

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

(iii) If $\sigma_0 \leq \sigma(x_0, \varepsilon)$, then $N \leq 2 + \left(\sqrt{\frac{2\gamma_{\text{inc}}L_f}{\sigma_0}} + 1 \right) \log \frac{\gamma_{\text{inc}}L_f + \sigma_0}{\beta\sigma_0}$.

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

$$\begin{aligned} & \frac{\sqrt{2\gamma_{\text{inc}}L_f}}{\sqrt{\gamma_{\text{reg}}} - 1} \left[\gamma_{\text{reg}}\sqrt{\frac{1}{\sigma(x_0, \varepsilon)}} - \sqrt{\frac{1}{\sigma_0}} \right] \log \left(\frac{\gamma_{\text{reg}}\gamma_{\text{inc}}L_f}{\beta\sigma(x_0, \varepsilon)} + \frac{1}{\beta} \right) \\ & + \left(2 + \log_{\gamma_{\text{reg}}} \frac{\sigma_0}{\sigma(x_0, \varepsilon)} \right) \left[2 + \log \left(\frac{\gamma_{\text{reg}}\gamma_{\text{inc}}L_f}{\beta\sigma(x_0, \varepsilon)} + \frac{1}{\beta} \right) \right] \\ & = O \left(\sqrt{\frac{L_f \text{dist}(x_0, X^*)}{\varepsilon}} \log \frac{L_f \text{dist}(x_0, X^*)}{\varepsilon} + \log \frac{\sigma_0 \text{dist}(x_0, X^*)}{\varepsilon} \log \frac{L_f \text{dist}(x_0, X^*)}{\varepsilon} \right). \end{aligned} \quad (16)$$

Proof. (i) By Lemma 4.1, the algorithm must terminate at the j -th loop whenever we have $(1 + \sqrt{2}\beta)\sigma_j \text{dist}(x_0, X^*) \leq \varepsilon$, or equivalently, $\sigma_j \leq \frac{\varepsilon}{(1 + \sqrt{2}\beta) \text{dist}(x_0, X^*)} = \sigma(x_0, \varepsilon)$. Thus, we obtain the former assertion of (i). To see the latter assertion, assume that $\sigma_j < \sigma(x_0, \varepsilon)/\gamma_{\text{reg}}$ 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 latter assertion of (i).

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

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

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

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

Therefore, we see that

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

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

In view of (i), the assertions (iii) and (iv) follow by applying (ii) with $\sigma_\ell = \sigma_0$ and $\sigma_\ell \geq \sigma(x_0, \varepsilon) / \gamma_{\text{reg}}$, respectively. The big- O expression (16) is obtained by substituting the definition of $\sigma(x_0, \varepsilon)$. \square

If σ_0 is chosen appropriately, then the complexity estimates in Theorem 4.2 (iii) and (vi) match the lower complexity bound (7) for the class $\mathcal{F}(x_0, R, L)$, up to a logarithmic factor. Nesterov's regularization technique [12] essentially corresponds to the ideal choice $\sigma_0 = \sigma(x_0, \varepsilon)$, which requires $\text{dist}(x_0, X^*)$ to be known. Here, we show a simple example to choose σ_0 so that **AdaAPG** attains the near optimality without knowing $\text{dist}(x_0, X^*)$.

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

$$\sigma_0 \in \left[\frac{2\varepsilon M}{(1 + \sqrt{2}\beta) \|g_M(x_0)\|}, \frac{2M}{1 + \sqrt{2}\beta} \right]. \quad (17)$$

Then, we have

$$\sigma(x_0, \varepsilon) \leq \sigma_0 \leq \frac{2\gamma_{\text{inc}} L_f}{1 + \sqrt{2}\beta}. \quad (18)$$

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

$$O \left(\sqrt{\frac{L_f \text{dist}(x_0, X^*)}{\varepsilon}} \log \frac{L_f \text{dist}(x_0, X^*)}{\varepsilon} \right). \quad (19)$$

Proof. Since $\varphi(T_M(x_0)) \leq m_M(x_0; T_M(x_0))$ holds by the condition (14) in **PGIter**, Lemma 2.1 (iv) implies

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

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

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

Therefore, (17) implies (18). In particular, we can apply Theorem 4.2 (iv) to obtain the complexity bound (16), whose second term is dominated by the first one due to the upper bound (18) of σ_0 . \square

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 (Notice that ε appears only in the stopping criterion). Then, the choice $\sigma_0 := 2M / (1 + 2\sqrt{\beta})$ given in Corollary 4.3 ensures the nearly optimal complexity bound (19) 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

$$\varphi(x) - \varphi^* \geq \kappa \operatorname{dist}(x, X^*)^\rho, \quad \forall x \in \operatorname{lev}_\varphi(\varphi(x^{(0)})), \quad (20)$$

for some $\kappa > 0$ and $\rho \geq 1$. 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 $\theta \in]0, 1[$:

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

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.

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

Parameters:

- $\gamma_{\text{inc}} > 1, \gamma_{\text{dec}} \geq 1, L_{\text{min}} \in]0, L_f]$ (for backtracking line-search);
- $\gamma_{\text{reg}} > 1$ (the ratio to decrease the regularization parameter);
- $\beta \in]0, 1]$ (controls the timing when the current regularization parameter must be decreased);
- $\theta \in]0, 1[$ (the ratio reducing the residue per iteration).

Input: $x^{(0)} \in \mathbb{E}, L^{(-1)} \in [L_{\text{min}}, \gamma_{\text{inc}} L_f], \sigma^{(0)} > 0, \varepsilon > 0$.

- 1: Compute $\{x_+^{(0)}, M^{(0)}, L^{(0)}\} \leftarrow \text{PGIter}(x^{(0)}, L^{(-1)})$.
 - 2: **for** $t = 0, 1, 2, \dots$ **do**
 - 3: **if** $\|g_{M^{(t)}}(x^{(t)})\| \leq \varepsilon$ **then** output $\{x^{(t)}, M^{(t)}, x_+^{(t)}\}$ and terminate the algorithm.
 - 4: $\varepsilon^{(t)} := \theta \|g_{M^{(t)}}(x^{(t)})\|$.
 - 5: $\{\sigma^{(t+1)}, x^{(t+1)}, x_+^{(t+1)}, M^{(t+1)}, L^{(t+1)}\} \leftarrow \text{AdaAPG}(x_+^{(t)}, L^{(t)}, \sigma^{(t)}, \varepsilon^{(t)})$
 - 6: (with testing the optional condition (11c) of **APGIter**).
 - 7: **end for**
-

4.2.1 Iteration Complexity Results and Near Optimality

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

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

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 $\|g_{M^{(0)}}(x^{(0)})\| \leq \varepsilon$ and otherwise

$$N(\varepsilon, \sigma_*, C) := \left(1 + \log_{1/\theta} \frac{\|g_{M^{(0)}}(x^{(0)})\|}{\varepsilon} + \log_{\gamma_{\text{reg}}} \frac{\sigma^{(0)}}{\sigma_*}\right) \left(2 + \log \frac{\gamma_{\text{inc}} L_f + \sigma_*}{\beta \sigma_*}\right) \\ + \frac{\sqrt{2\gamma_{\text{inc}} L_f}}{\sqrt{\gamma_{\text{reg}} - 1}} \left[\sqrt{\frac{1}{\sigma_*}} - \sqrt{\frac{1}{\sigma^{(0)}}} \right] \log \frac{\gamma_{\text{inc}} L_f + \sigma_*}{\beta \sigma_*} + C \sqrt{2\gamma_{\text{inc}} L_f} \log \frac{\gamma_{\text{inc}} L_f + \sigma_*}{\beta \sigma_*}. \quad (21)$$

Moreover, we define

$$\bar{\sigma} = \begin{cases} \frac{\theta}{1 + \sqrt{2}\beta} \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1\right)^{-\frac{1}{\rho-1}} \varepsilon^{\frac{\rho-2}{\rho-1}} & (\text{if } \rho \geq 2), \\ \frac{\theta}{1 + \sqrt{2}\beta} \kappa^{\frac{2}{\rho}} \left(\frac{L_f}{L_{\min}} + 1\right)^{-1} (\varphi(x^{(0)}) - \varphi^*)^{-\frac{2-\rho}{\rho}} & (\text{if } 1 \leq \rho < 2). \end{cases} \quad (22)$$

The next theorem provides the descriptions of iteration complexity bounds, which have complicated expressions to explicate 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 (20) holds. In rAdaAPG, let N be the total number of executions of APGIter. Denote*

$$\sigma_* := \begin{cases} \sigma^{(0)} & (\text{if } \sigma^{(0)} \leq \bar{\sigma}), \\ \bar{\sigma}/\gamma_{\text{reg}} & (\text{otherwise}), \end{cases} \quad (23)$$

for $\bar{\sigma}$ defined in (22). Also, let $N(\cdot, \cdot, \cdot)$ be defined by (21). Then, the following assertions hold.

(i) If $\rho = 2$, then $N \leq N(\varepsilon, \sigma_*, C)$ holds with $C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\|g_{M^{(0)}}(x^{(0)})\|}{\varepsilon}\right)$.

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

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

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 $\sigma^{(t)} = \sigma^{(0)}$ for all $t \geq 0$ and $N \leq N(\varepsilon, \sigma^{(0)}, C)$ holds with $C = \sqrt{\frac{1}{\sigma^{(0)}}} \left(1 + \log_{1/\theta} \frac{\|g_{M^{(0)}}(x^{(0)})\|}{\varepsilon}\right)$.

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

$$N \leq 1 + N(\max(\varepsilon, \varepsilon_*), \sigma_*, C) + \left(\log \frac{1}{\rho-1}\right)^{-1} \left(\log \log \frac{\varepsilon_*}{\theta^{\frac{\rho-1}{2-\rho}} \min(\varepsilon, \varepsilon_*)} - \log \log \frac{1}{\theta^{\frac{\rho-1}{2-\rho}}}\right), \quad (24)$$

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

$$\varepsilon_* = \left[(1 + \sqrt{2})\theta^{-1}(\gamma_{\text{inc}} L_f + \sigma^{(0)})\right]^{-\frac{\rho-1}{2-\rho}} \kappa^{\frac{1}{2-\rho}} \left(\frac{L_f}{L_{\min}} + 1\right)^{-\frac{1}{2-\rho}}. \quad (25)$$

(iv) If $\rho = 1$, then we have $N \leq 1 + N(\max(\varepsilon, \varepsilon_*), \sigma_*, C)$, where $\varepsilon_* = \kappa \left(\frac{L_f}{L_{\min}} + 1 \right)^{-1}$ and $C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\|g_{M^{(0)}}(x^{(0)})\|}{\max(\varepsilon, \varepsilon_*)} \right)$. Moreover, if $\varepsilon < \varepsilon_*$, then $x_+^{(t)}$ in the output of the algorithm must be an optimal solution to the original problem (1).

Corollary 4.5. Let $x^{(0)} \in \mathbb{E}$ and $L^{(-1)} \in [L_{\min}, \gamma_{\text{inc}} L_f]$. Assume that the Hölderian error bound condition (20) holds. After the initialization step (Line 1) in **rAdaAPG**, determine $\sigma^{(0)}$ by

$$\sigma^{(0)} := \frac{2\varepsilon^{(0)}M}{(1 + \sqrt{2}\beta) \|g_M(x_+^{(0)})\|}, \quad \text{where } \{T_M(x_+^{(0)}), M, L\} \leftarrow \text{PGIter}(x_+^{(0)}, L^{(0)}). \quad (26)$$

Then, **rAdaAPG**($x^{(0)}, L^{(-1)}, \sigma^{(0)}, \varepsilon$) finds $x^{(t)} \in \mathbb{E}$ and $M^{(t)} > 0$ such that $\|g_{M^{(t)}}(x^{(t)})\| \leq \varepsilon$ 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 (25), and we regard $\theta, \gamma_{\text{inc}}, \gamma_{\text{reg}}, \beta, \frac{L_f}{L_{\min}}$ as constants⁴.

$$\left\{ \begin{array}{ll} O \left(\left[\log \frac{\|g_0\|}{\varepsilon} + \sqrt{\frac{L_f}{\kappa^{\frac{1}{\rho-1}} \varepsilon^{\frac{\rho-2}{\rho-1}}}} \right] \log \frac{L_f}{\varepsilon^{\frac{\rho-2}{\rho-1}}} \right) & (\rho > 2), \\ O \left(\sqrt{\frac{L_f}{\kappa}} \log \frac{L_f}{\kappa} \log \frac{\|g_0\|}{\varepsilon} \right) & (\rho = 2), \\ O \left(\sqrt{\frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}}} \log \frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}} \log \frac{\|g_0\| L_f^{\frac{\rho-1}{2-\rho}}}{\kappa^{\frac{1}{2-\rho}}} + (\log \frac{1}{\rho-1})^{-1} \log \log \frac{\kappa^{\frac{1}{2-\rho}}}{L_f^{\frac{\rho-1}{2-\rho}} \varepsilon} \right) & (\rho \in]1, 2[, \varepsilon \leq \varepsilon_*), \\ O \left(\sqrt{\frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}}} \log \frac{L_f \Delta_0^{\frac{2-\rho}{\rho}}}{\kappa^{\frac{2}{\rho}}} \log \frac{\|g_0\|}{\varepsilon} \right) & (\rho \in]1, 2[, \varepsilon \geq \varepsilon_*), \\ O \left(\sqrt{\frac{L_f \Delta_0}{\kappa^2}} \log \frac{L_f \Delta_0}{\kappa^2} \log \frac{\|g_0\|}{\max(\kappa, \varepsilon)} \right) & (\rho = 1). \end{array} \right. \quad (27)$$

Let us observe the consequences of Corollary 4.5. Notice that the iteration complexity bounds (27) in the cases $\rho = 2$ and $\rho > 2$ match the lower bounds (9) up to a logarithmic factor. Therefore, **rAdaAPG** achieves the near optimality for the class $\mathcal{F}(x_0, R, L, \kappa, \rho)$.

Note that the choice (26) of $\sigma^{(0)}$ to obtain the above iteration complexity is independent of the target tolerance ε . Therefore, if we consider **rAdaAPG** with $\varepsilon = 0$ (recall that ε only affects the stopping criterion), it yields a convergent algorithm with near optimality. In summary, to find an approximate solution x with $\|g_L(x)\| \leq \varepsilon$, the proposed method **rAdaAPG** is able to attain the following iteration complexity (and the rate of convergence).

- 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).

⁴This assumption on L_f/L_{\min} means that we know a good estimate L_{\min} close to L_f in advance.

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 (30), the following relation holds if $\rho \neq 1$.

$$\varphi(x_+^{(t)}) - \varphi^* \leq \frac{1}{\kappa^{\frac{1}{\rho-1}}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{\frac{\rho}{\rho-1}} \left\| g_{M^{(t)}}(x^{(t)}) \right\|^{\frac{\rho}{\rho-1}}. \quad (28)$$

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

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

Substituting this ε to our complexity bound (27), we also obtain an iteration complexity bound under the measure $\varphi(\cdot) - \varphi^*$, which is nearly optimal in view of the lower complexity bound (8). 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 (28) 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 (20) holds. In the execution of **rAdaAPG**, the following assertions hold.*

- (i) $L^{(t)}, M^{(t)} \in [L_{\min}, \gamma_{\text{inc}} L_f]$ for all $t \geq 0$.
- (ii) $\varphi(x_+^{(t+1)}) \leq \varphi(x_+^{(t)}) \leq \dots \leq \varphi(x_+^{(0)}) \leq \varphi(x^{(0)})$ for all $t \geq 0$.
- (iii) $t \leq \log_{1/\theta} \frac{\|g_{M^{(0)}}(x^{(0)})\|}{\|g_{M^{(t)}}(x^{(t)})\|}$ for each $t \geq 0$.
- (iv) Whenever $x_+^{(t)} \notin X^*$, we have

$$\text{dist}(x_+^{(t)}, X^*)^{\rho-1} \leq \frac{1}{\kappa} \left(\frac{L_f}{L_{\min}} + 1 \right) \left\| g_{M^{(t)}}(x^{(t)}) \right\|, \quad (29)$$

$$(\varphi(x_+^{(t)}) - \varphi^*)^{\rho-1} \leq \frac{1}{\kappa} \left(\frac{L_f}{L_{\min}} + 1 \right)^{\rho} \left\| g_{M^{(t)}}(x^{(t)}) \right\|^{\rho}. \quad (30)$$

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

(ii) For $t = 0$, we have $\varphi(x_+^{(0)}) = \varphi(T_{M^{(0)}}(x^{(0)})) \leq \varphi(x^{(0)})$ by Lemma 3.7 (i). Moreover, the criterion (11c) and Proposition 3.6 (iv), applied to the subroutine **AdaAPG** in **rAdaAPG**, yield $\varphi(x_+^{(t+1)}) \leq \varphi(x^{(t+1)})$ and $\varphi(x^{(t+1)}) \leq \varphi(x_+^{(t)})$, respectively. This shows $\varphi(x_+^{(t+1)}) \leq \varphi(x_+^{(t)})$.

(iii) Since the recurrence $\varepsilon^{(t)}/\theta = \|g_{M^{(t)}}(x^{(t)})\| \leq \varepsilon^{(t-1)}$ holds for $t \geq 1$, we have $\varepsilon^{(0)} \geq \varepsilon^{(t)}/\theta^t$ for each $t \geq 0$. Therefore, we conclude $t \leq \log_{1/\theta} \frac{\varepsilon^{(0)}}{\varepsilon^{(t)}} = \log_{1/\theta} \frac{\|g_{M^{(0)}}(x^{(0)})\|}{\|g_{M^{(t)}}(x^{(t)})\|}$.

(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 $\text{lev}_{\varphi}(\varphi(x^{(0)})) \setminus X^*$ by (ii))

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

where the second inequality is due to Lemma 2.1 (iii) and the last follows by (i). This shows the assertion (29). Similarly, (30) can be obtained using the latter inequality of (6). \square

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

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

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

Then, under the definition (21) of $N(\cdot, \cdot, \cdot)$, the following inequality holds.

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

where

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

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

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

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

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

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

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

and

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

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

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

Finally, C has the following bound:

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

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

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 (20) 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 (22).*

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

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

where $\sigma(\cdot, \cdot)$ is defined by (15).

(ii) *It follows that $\sigma(x_+, \varepsilon^{(t)}) \geq \bar{\sigma}$ for all $t = 0, \dots, T$. When $\rho \geq 2$, we further obtain*

$$\sigma(x_+, \varepsilon^{(t)}) \geq \frac{\theta^{\frac{1}{\rho-1}}}{1 + \sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-\frac{1}{\rho-1}} (\varepsilon^{(t)})^{\frac{\rho-2}{\rho-1}} \geq \bar{\sigma}, \quad (31)$$

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

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

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

Proof. (i) In the case $\sigma^{(t)} \leq \sigma(x_+, \varepsilon^{(t)})$, 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 $\sigma^{(t)} > \sigma(x_+, \varepsilon^{(t)})$. When the subroutine **AdaAPG** at the t -th outer loop terminates at the first loop $j = 0$, then it is clear that $\sigma^{(t+1)} = \sigma^{(t)} > \sigma(x_+, \varepsilon^{(t)}) \geq \sigma(x_+, \varepsilon^{(t)})/\gamma_{\text{reg}}$. In the other case, the latter assertion of Theorem 4.2 (i) implies $\sigma^{(t+1)} \geq \sigma(x_+, \varepsilon^{(t)})/\gamma_{\text{reg}}$.

(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 (29) implies

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

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 (31).

If $\rho \in [1, 2]$, on the other hand, since $\varphi(x_+^{(t)}) \leq \varphi(x^{(0)})$ by Lemma 4.6 (ii), the inequalities (20) and (30) imply (remark that $\frac{2-\rho}{\rho} > 0$)

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

Therefore, we conclude that

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

This proves the assertion (ii).

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

In the case $\sigma^{(0)} \geq \bar{\sigma}$, on the other hand, let us show $\sigma^{(t)} \geq \bar{\sigma}/\gamma_{\text{reg}}$ ($t = 0, \dots, T+1$) by induction. The assertion for $t = 0$ is clear since $\sigma^{(0)} \geq \bar{\sigma} \geq \bar{\sigma}/\gamma_{\text{reg}}$. Assume that $\sigma^{(t)} \geq \bar{\sigma}/\gamma_{\text{reg}}$ holds for some $t \geq 0$. By (i) and (ii), we have $\sigma^{(t+1)} \geq \min\{\sigma^{(t)}, \sigma(x_+^{(t)}, \varepsilon^{(t)})/\gamma_{\text{reg}}\} \geq \min\{\bar{\sigma}/\gamma_{\text{reg}}, \bar{\sigma}/\gamma_{\text{reg}}\} = \bar{\sigma}/\gamma_{\text{reg}}$. This completes the proof of (iii). \square

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 (20) 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 (22), then there exists $t_0 \in \{0, \dots, T+1\}$ such that the following conditions hold, where $\sigma_*^{(0)} := \frac{\theta}{1 + \sqrt{2}\beta} \cdot \kappa^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1 \right)^{-\frac{1}{\rho-1}} \|g_{M^{(0)}}(x^{(0)})\|^{\frac{\rho-2}{\rho-1}}$.*

$$(i) \quad \sigma^{(0)} = \dots = \sigma^{(t_0)}. \quad (32)$$

$$(ii) \quad \sigma^{(t+1)} \geq \theta^{-\frac{\rho-2}{\rho-1}(T-t)} \bar{\sigma}/\gamma_{\text{reg}}, \quad t = t_0, \dots, T. \quad (33)$$

$$(iii) \quad t_0 \leq 1 + \frac{\rho-1}{\rho-2} \log_{1/\theta} \frac{\sigma_*^{(0)}}{\min(\sigma^{(0)}, \sigma_*^{(0)})}. \quad (34)$$

$$(iv) \quad \theta^{\frac{\rho-2}{\rho-1}(T-t_0)} \geq \bar{\sigma}/\sigma^{(0)}. \quad (35)$$

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

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

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 $\varepsilon^{(t+1)} \leq \theta\varepsilon^{(t)}$ implies

$$\sigma_*^{(t+1)} = c(\varepsilon^{(t+1)})^\omega \leq c\theta^\omega (\varepsilon^{(t)})^\omega = \theta^\omega \sigma_*^{(t)} \leq \sigma_*^{(t)}. \quad (37)$$

Let t_0 be the smallest integer in $\{0, \dots, T+1\}$ such that $\sigma^{(0)} \geq \sigma_*^{(t_0)}$. Remark that, by the definition of T , we have $\varepsilon^{(T+1)} = \theta \|g_{M(T+1)}(x^{(T+1)})\| \leq \theta\varepsilon$. This implies that $\sigma^{(0)} \geq \bar{\sigma} = c(\theta\varepsilon)^\omega \geq c(\varepsilon^{(T+1)})^\omega = \sigma_*^{(T+1)}$. Therefore, t_0 is well-defined.

(i) By the definition of t_0 and (36), we have $\sigma^{(0)} < \sigma_*^{(t)} \leq \sigma(x_+^{(t)}, \varepsilon^{(t)})$ ($0 \leq \forall t < t_0$). Therefore, by induction, we obtain $\sigma^{(0)} = \sigma^{(1)} = \dots = \sigma^{(t_0)}$ due to Lemma 4.8 (i).

(ii) Let us show $\sigma^{(t+1)} \geq \sigma_*^{(t)}/\gamma_{\text{reg}}$ ($t_0 \leq t \leq T$). To prove this, we verify $\sigma^{(t)} \geq \sigma_*^{(t)}/\gamma_{\text{reg}}$ for $t = t_0, \dots, T+1$ by induction. Note that $\sigma^{(t_0)} = \sigma^{(0)} \geq \sigma_*^{(t_0)} > \sigma_*^{(t_0)}/\gamma_{\text{reg}}$ holds by (i) and the definition of t_0 . Now under the hypothesis $\sigma^{(t)} \geq \sigma_*^{(t)}/\gamma_{\text{reg}}$ for t with $t_0 \leq t \leq T$, Lemma 4.8 (i) and (36) imply

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

Therefore, this completes the induction; in addition, the above inequality involves the desired inequality $\sigma^{(t+1)} \geq \sigma_*^{(t)}/\gamma_{\text{reg}}$ ($t_0 \leq t \leq T$). This yields (ii) combined with (36) and (37):

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

(iii) If $t_0 = 0$, then (iii) is trivial since $\sigma^{(0)} \geq \sigma_*^{(t_0)} = \sigma_*^{(0)}$. If $t_0 > 0$, then the definition of t_0 and using (37) imply $\sigma^{(0)} < \sigma_*^{(t_0-1)} \leq \theta^{\omega(t_0-1)} \sigma_*^{(0)}$, which yields $t_0 \leq 1 + \frac{1}{\omega} \log_{1/\theta} \frac{\sigma_*^{(0)}}{\sigma^{(0)}}$.

(iv) By (36), (37), and the definition of t_0 , remark that $\bar{\sigma} \leq \sigma_*^{(T)} \leq \theta^{\omega(T-t_0)} \sigma_*^{(t_0)} \leq \theta^{\omega(T-t_0)} \sigma^{(0)}$. Hence, $\theta^{\omega(T-t_0)} \geq \bar{\sigma}/\sigma^{(0)}$ holds. \square

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

Proof of Theorem 4.4. We may assume that $\|g_{M^{(0)}}(x^{(0)})\| > \varepsilon$ since $N = 0$ on the other case. Suppose that **rAdaAPG** terminated with the stopping criterion at the $(T+1)$ -th outer loop for some $T \geq 0$. Denote by $N^{(t)}$ the number of executions of **APGIter** at the t -th outer loop so that $N = \sum_{t=0}^T N^{(t)}$. Let $\bar{\sigma}$ be defined by (22).

By the definition of T , we have

$$\left\| g_{M^{(t)}}(x^{(t)}) \right\| > \varepsilon, \quad t = 0, \dots, T. \quad (38)$$

Moreover, by the definition of σ_* in (23), using Lemma 4.8 (iii) implies

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

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

$$N \leq N(\varepsilon, \sigma_*, C) \quad \text{with} \quad C = \sqrt{\frac{1}{\sigma_*}} \left(1 + \log_{1/\theta} \frac{\|g_{M^{(0)}}(x^{(0)})\|}{\varepsilon} \right). \quad (39)$$

For the case $\rho = 2$, (39) proves the assertion (i). We discuss the other cases to improve (39).

(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 (39).

Now consider the case $\sigma^{(0)} \geq \bar{\sigma}$. Then, there exists $t_0 \in \{0, \dots, T + 1\}$ satisfying the conditions in Lemma 4.9. By (35), remark that

$$\sum_{t=t_0}^T \sqrt{\theta^{\frac{\rho-2}{\rho-1}(T-t)}} = \sum_{i=0}^{T-t_0} \sqrt{\theta^{\frac{\rho-2}{\rho-1}i}} = \frac{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}T-t_0+1}}}{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}} \leq \frac{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}} \cdot \frac{\bar{\sigma}}{\sigma^{(0)}}}}{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}}. \quad (40)$$

Therefore, we conclude that

$$\begin{aligned} \sum_{t=0}^T \sqrt{\frac{1}{\sigma^{(t+1)}}} &= \sum_{t=0}^{t_0-1} \sqrt{\frac{1}{\sigma^{(t+1)}}} + \sum_{t=t_0}^T \sqrt{\frac{1}{\sigma^{(t+1)}}} \\ &\leq t_0 \sqrt{\frac{1}{\sigma^{(0)}}} + \sqrt{\frac{\gamma_{\text{reg}}}{\bar{\sigma}}} \sum_{t=t_0}^T \sqrt{\theta^{\frac{\rho-2}{\rho-1}(T-t)}} \quad (\text{by (32) and (33)}) \\ &\leq \sqrt{\frac{1}{\sigma^{(0)}}} \left(1 + \frac{\rho-1}{\rho-2} \log_{1/\theta} \frac{\sigma_*^{(0)}}{\min(\sigma_*^{(0)}, \sigma^{(0)})} \right) \\ &\quad + \frac{\sqrt{\gamma_{\text{reg}}}}{1 - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}}} \left(\sqrt{\frac{1}{\bar{\sigma}}} - \sqrt{\theta^{\frac{\rho-2}{\rho-1}}} \sqrt{\frac{1}{\sigma^{(0)}}} \right) \quad (\text{by (34) and (40)}) \\ &=: C. \end{aligned}$$

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

(iii) Case $\rho \in]1, 2[$. If $\varepsilon \geq \varepsilon_*$, then (39) gives our assertion because the second term of (24) vanishes. Suppose, on the other hand, that $\varepsilon < \varepsilon_*$. Denote $\xi_t := \|g_{M^{(t)}}(x^{(t)})\|$ and let $T_* \geq 0$ be the smallest integer such that $\xi_{T_*} \leq \varepsilon_*$. Then, since $\xi_{T_*-1} > \varepsilon_*$ holds, Lemma 4.7 shows that

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

Note that, under the convention $\sum_{t=0}^{-1}(\cdot) = 0$, this inequality also holds if $T_* = 0$ since then $\xi_0 \leq \varepsilon_*$ 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 **AdaAPG**($x_+^{(t)}, L^{(t)}, \sigma^{(t)}, \varepsilon^{(t)}$), which executes

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

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

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

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 $M_0 \leq \gamma_{\text{inc}}L_f$ (Proposition 3.6 (i)), we have

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

Combining them and using (29)⁵, we conclude that

$$\begin{aligned} \|g_{M_0}(x_1)\| &\leq (1 + \sqrt{2})(\gamma_{\text{inc}}L_f + \sigma^{(0)}) \text{dist}(x_+^{(t)}, X^*) \\ &\leq (1 + \sqrt{2})(\gamma_{\text{inc}}L_f + \sigma^{(0)}) \left(\frac{1}{\kappa}\right)^{\frac{1}{\rho-1}} \left(\frac{L_f}{L_{\min}} + 1\right)^{\frac{1}{\rho-1}} \xi_t^{\frac{1}{\rho-1}} \\ &= \theta \varepsilon_*^{\frac{\rho-2}{\rho-1}} \xi_t^{\frac{1}{\rho-1}} \\ &\leq \theta \xi_t^{\frac{\rho-2}{\rho-1}} \xi_t^{\frac{1}{\rho-1}} = \theta \xi_t = \varepsilon^{(t)}, \end{aligned} \tag{41}$$

where the last inequality is due to $\xi_t \leq \varepsilon_*$ for $t \geq T_*$ (and remark $\frac{\rho-2}{\rho-1} < 0$). This shows $N^{(t)} = 1$ and (41) yields the recurrence $\xi_{t+1} \leq \theta \varepsilon_*^{\frac{\rho-2}{\rho-1}} \xi_t^{\frac{1}{\rho-1}}$, for each $t = T_*, \dots, T$. Since $\frac{1}{\rho-1} > 1$, it reduces ξ_t superlinearly. In particular, solving this recurrence implies

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

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

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

Consequently, N is bounded as follows.

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

⁵Although (29) is asserted in the case $x_+^{(t)} \notin X^*$, it trivially holds if $x_+^{(t)} \in X^*$ unless $\rho = 1$.

(iv) Case $\rho = 1$. We have $\sigma^{(t)} \geq \sigma_*$ 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

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

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

$$\left\| g_{M^{(t)}}(x^{(t)}) \right\| \geq \varepsilon_*, \quad 0 \leq t \leq T-1. \quad (42)$$

Now we consider two cases. If $\left\| g_{M^{(T)}}(x^{(T)}) \right\| \geq \varepsilon_*$ holds, then combining with (38) 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 (38) and (42) 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. \square

Proof of Corollary 4.5. The function $N(\cdot, \cdot, \cdot)$ defined in (21) 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 (26) of $\sigma^{(0)}$, we can apply Corollary 4.3 and then the bound (18) becomes

$$\sigma(x_+^{(0)}, \varepsilon^{(0)}) \leq \sigma^{(0)} \leq \frac{2\gamma_{\text{inc}} L_f}{1 + \sqrt{2}\beta}. \quad (43)$$

Then, we have $\sigma^{(0)} \geq \bar{\sigma}$ since $\sigma(x_+^{(0)}, \varepsilon^{(0)}) \geq \bar{\sigma}$ holds by Lemma 4.8 (ii). Therefore, σ_* in (23) becomes $\sigma_* = \Theta(\bar{\sigma})$, which also implies $\sigma_* = O(\sigma^{(0)}) = O(L_f)$ combined with (43). Applying the bounds $\sigma^{(0)} = O(L_f)$, $\sigma_* = O(L_f)$, and $\sigma_* = \Omega(\bar{\sigma})$, we obtain

$$N(\varepsilon, \sigma_*, C) = O \left(\log \frac{L_f}{\bar{\sigma}} \left[\log \frac{\|g_0\|}{\varepsilon} + \sqrt{\frac{L_f}{\bar{\sigma}}} + C\sqrt{L_f} \right] \right). \quad (44)$$

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, (44) yields

$$N(\varepsilon, \sigma_*, C) = O \left(\sqrt{\frac{L_f}{\bar{\sigma}}} \log \frac{L_f}{\bar{\sigma}} \log \frac{\|g_0\|}{\varepsilon} \right). \quad (45)$$

Since $\bar{\sigma} = \Theta(\kappa)$ by (22), we conclude the bound (27) 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 (45) replacing ε by $\max(\varepsilon, \varepsilon_*)$, where $\varepsilon_* = \kappa(L_f/L_{\min} + 1)^{-1} = \Omega(\kappa)$. Since $\bar{\sigma} = \Omega(\kappa^2/\Delta_0)$ by (22), we obtain the bound (27) 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 (27) in this case can be obtained based on the estimate (45) replacing ε by $\max(\varepsilon, \varepsilon_*)$ and applying $\bar{\sigma} = \Omega(\kappa^{\frac{2}{\rho}} \Delta_0^{-\frac{2-\rho}{\rho}})$, $\varepsilon_* = \Theta(\kappa^{\frac{1}{2-\rho}} L_f^{-\frac{\rho-1}{2-\rho}})$.

Finally, consider the case $\rho > 2$. By Lemma 4.8 (ii) and (43), 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, (44) applying $\bar{\sigma} = \Omega(\kappa^{\frac{1}{\rho-1}} \varepsilon^{\frac{\rho-2}{\rho-1}})$ concludes the bound (27) in the case $\rho > 2$. \square

5 Conclusions

In this paper, we proposed two adaptive proximal gradient methods, Algorithms 3 and 4, for composite convex optimization problems. These two algorithms are nearly optimal for the class of problems with smooth convex objective functions and for the class with an additional assumption, the Hölderian error bound condition, respectively. Without this additional assumption, it is unclear whether the latter algorithm also provides the near optimality.

An important question would be the establishment of optimal methods. Remark that, under the gradient norm measure, a first-order method with optimal complexity for smooth convex functions was recently announced in [10], namely, the complexity bound (7) is tight (if $\dim \mathbb{E}$ is large enough). Similarly, it is an important question whether we can improve the complexity (27) to attain the lower bounds (9).

The key idea of this work is the adaptive determination of the regularization parameter used to define the regularized objective function. As proved in Theorem 4.2, our method (Algorithm 3) adapts the unknown and ideal regularization parameter defined in (15). 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 the ideal regularization parameter (in other words, the distance to the solution set) and the “problem structure” (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] Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2**, 183–202 (2009)
- [2] Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math. Dokl.* **27**, 372–376 (1983)

- [3] Nesterov, Y.: Gradient methods for minimizing composite functions. *Math. Program.* **140**, 125–161 (2013)
- [4] Nesterov, Y.: Universal gradient methods for convex optimization problems. *Math. Program.* **152**, 381–404 (2015)
- [5] Renegar, J., Grimmer, B.: A simple nearly-optimal restart scheme for speeding-up first order methods. *ArXiv preprint*, arXiv:1803.00151v1 (2018)
- [6] Roulet, V., d’Aspremont, A.: Sharpness, restart and acceleration. *Adv. Neural. Inf. Process. Syst.* **30** (2017)
- [7] Nesterov, Y.: Smooth minimization of non-smooth functions. *Math. Program.* **103**, 127–152 (2005)
- [8] Nesterov, Y.: Primal-dual subgradient methods for convex problems. *Math. Program.* **120**, 221–259, (2009)
- [9] Lu, Z.: Smooth optimization approach for sparse covariance selection, *SIAM J. Optim.* **19**, 1807–1827 (2009)
- [10] Nesterov, Y., Gasnikov, A., Guminov, S., Dvurechensky, P.: Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optim. Methods Softw.* (2020). <https://doi.org/10.1080/10556788.2020.1731747>
- [11] Kim, D., Fessler, J. A.: Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *ArXiv preprint*, arXiv:1803.06600v3 (2020)
- [12] Nesterov, Y.: How to make the gradients small. *Optima* **88**, 10–11 (2012)
- [13] Kim, D., Fessler, J. A.: Generalizing the optimized gradient method for smooth convex minimization. *SIAM J. Optim.* **28**, 1920–1950 (2018)
- [14] Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. In: *Les Équations aux Dérivées Partielles*, 87–89, Éditions du Centre National de la Recherche Scientifique, Paris (1963)
- [15] Łojasiewicz, S.: Ensembles semi-analytiques. preprint, Institut des Hautes Études Scientifiques (1965)
- [16] Attouch, H., Bolte, J., Svaiter, B. F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.* **137**, 91–129 (2013)
- [17] Bolte, J., Nguyen, T. P., Peypouquet, J., Suter, B. W.: From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program.* **165**, 471–507 (2017)
- [18] Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.* **17**, 1205–1223 (2007)
- [19] Liu, M., Yang, T.: Adaptive accelerated gradient converging methods under Hölderian error bound condition. *Adv. Neural Inf. Process. Syst.* **30** (2017)
- [20] Fercoq, O., Qu, Z.: Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA J. Numer. Anal.* **39**, 2069–2095 (2019)

- [21] Lin, Q., Xiao, L.: An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. *Comput. Optim. Appl.* **60**, 633–674 (2015)
- [22] Nesterov, Y.: *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, Norwell (2004)
- [23] Parikh, N., Boyd, S.: Proximal algorithms. *Foundations and Trends in Optimization* **1**, 123–231 (2014)
- [24] Rockafellar, R. T.: *Convex Analysis*. Princeton University Press, New Jersey (1970)
- [25] Lojasiewicz, S.: Sur le problème de la division. *Studia Math.* **18**, 87–136 (1959)
- [26] Nemirovsky, A.: Information-based complexity of linear operator equations. *J. Complexity* **8**, 153–175 (1992)
- [27] Nemirovsky, A., Nesterov, Y.: Optimal methods for smooth convex minimization. *USSR Comput. Math. Math. Phys.* **25**, 21–30 (1985)