

A ON PROXIMAL GRADIENT MAPPING AND ITS MINIMIZATION IN NORM VIA POTENTIAL FUNCTION-BASED ACCELERATION*

Beier Chen and Hui Zhang[†]

Abstract: The proximal gradient descent method, well-known for convex composite optimization, can be completely described by the concept of proximal gradient mapping. In this note, we introduce two new results of proximal gradient mapping–norm monotonicity and refined descent, with which we are able to extend the recently proposed potential function-based framework from gradient descent to proximal gradient descent.

Key words: *potential function-based framework, proximal gradient mapping, acceleration, proximal gradient method, composite optimization*

Mathematics Subject Classification: 90C25, 90C33, 90C47

1 Introduction

First-order methods, which date back to 1847 with the work of Cauchy on the vanilla gradient descent, have recently revived a great deal of research interest due to their low iteration cost as well as low memory storage. How to establish convergence criteria and determine convergence rates for a given first-order method heavily depends on the choice of optimality measures. The standard optimization literature on smooth convex first-order optimization mainly provides guarantees for optimality gap (in terms of function value) and distance gap (between the iterate and the minimizer set). However, these optimality measures only have theoretical value but do not fit practical applications because the optimal function value and the minimizer set are usually unknown before applying first-order methods.

Due to the basic fact that minimizing a smooth convex function is equivalent to minimizing the norm of its gradient, a more practical alternative to the optimality gap and distance gap may be the norm of gradient. This fact was initially exploited by Nesterov in the work [6], which argued that using the norm of gradient as an optimality measure is natural and more practical. There are many different potential function-based frameworks covering broad classes of first-order methods for providing optimality gap and distance gap guarantees, but not for the norm of gradient. This absence motivates the authors of [3] to introduce a novel potential function-based framework, in which they are able to address the problem of minimizing the norm of the gradient of a smooth convex function. As a natural

*This work was supported by the National Science Foundation of China (Nos.11971480).

[†]Corresponding author

development, we wonder whether their potential function-based framework can be extended to other types of first-order methods.

In this paper, we go a small step further along this direction by applying their potential function-based framework to composite convex optimization—minimizing the sum of a smooth convex function and a possibly nonsmooth convex function. To this end, we first revisit the proximal gradient mapping and introduce two new results—norm monotonicity and refined descent; both of them may have an independent interest in their own. Then, based on the newly discovered properties and the potential function-based framework of [3], we establish the sublinear convergence for the norm sequence of proximal gradient mapping. Moreover, we construct a new potential function to obtain faster convergence.

At the time of writing this paper, a closely related work [4], posted on arXiv very recently, also addressed the problem of minimizing the proximal gradient mapping under the name of proximal subgradient norm minimization. Here, we would like to point out three main differences between this work and ours. First, the potential function-based frameworks are different: they followed the discrete Lyapunov function in [7] while we extended that in [3]. Second, the accelerated algorithmic schemes are different: they analyzed the faster iterative shrinkage-thresholding algorithm (FISTA) in [1] while we run two iterative processes for acceleration. At last, the main results are different: they never used the norm monotonicity of proximal gradient mapping so that their result on proximal subgradient norm minimization for ISTA seems suboptimal. Nevertheless, we believe that these two works have their own merits and complement each other.

Now, we summarize the contribution of this paper as follows:

- For minimizing the sum of a smooth strongly convex function and a non-smooth convex function, we establish the exact worst-case convergence rate of the proximal gradient method for any step size and for the norm of proximal gradient mapping; see Theorem 3.4. This result is inspired by the work [8] which built worst-case convergence rates of the proximal gradient method for other standard optimality measures such as distance gap and residual gradient norm. In particular, the residual gradient defined in [8] is some certain subgradient, different from the standard proximal gradient.
- With the help of Theorem 3.4, we establish a refined descent lemma, which improves several existing descent-type lemmas; see Theorem 3.6. We note that a similar discovery was recently independently made by the authors of [9].
- With the help of Theorem 3.6, we extend the recently proposed potential function-based framework in [3] from gradient descent to proximal gradient descent by constructing suitable potential functions; see Theorem 4.1 and Theorem 4.4. Thanks to Theorem 3.4, the monotone convergence of the norm of gradient mapping with tight convergence rates, shown in Theorem 4.1, seems new.

The remainder of the paper is organized as follows. In Section 2, we present the basic notation and preliminary knowledge of different function classes, the proximal gradient method, and the potential function-based framework of [3]. In Section 3, we revisit the proximal gradient mapping and establish two new properties. In Section 4, we study the problem of minimizing the norm of proximal gradient mapping and show the main convergence results. Finally, Section 5 gives some concluding remarks.

2 Preliminaries and Preliminary Results

In this paper, we restrict our attention to an arbitrary finite dimensional space $\mathbb{R}^n$ associated with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\| := \sqrt{\langle \cdot, \cdot \rangle}$. $\mathbb{R}_+$ is the set of nonnegative real numbers. For a closed subset $Q \subseteq \mathbb{R}^n$ and a point $x \in \mathbb{R}^n$, we define by $d(x, Q) := \inf_{y \in Q} \|x - y\|$ the distance function from x to Q , and define the indicator function of Q by

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

2.1 Different classes of functions

In order to introduce the class of smooth convex functions, we first give the definitions of convexity and smoothness. There are several equivalent definitions of convexity; here we present the first-order definition of convexity in the following form:

$$(\forall x, y \in \mathbb{R}^n) : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \quad (2.1)$$

The convexity of f essentially says that the function f can be lower bounded by a linear function; in contrast, the smoothness of f actually says that the function f can be upper bounded by a quadratic function, that is

$$(\forall x, y \in \mathbb{R}^n) : \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \quad (2.2)$$

where $L > 0$ is a constant. A function is called smooth convex if the inequalities (2.1) and (2.2) hold at the same time; the class of smooth convex functions is denoted by $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$. Surprisingly, the convexity inequality (2.1) and the smoothness inequality (2.2) can be equivalently characterized by a single inequality, that is

$$(\forall x, y \in \mathbb{R}^n) : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2, \quad (2.3)$$

from which the convexity is obviously implied. Interestingly, the inequality (2.3) is also equivalent to the cocoercive property of gradient, formulated as

$$(\forall x, y \in \mathbb{R}^n) : \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2. \quad (2.4)$$

The fact of equivalence between the inequalities above was observed in the book [5]. In order to describe a more general fact, we introduce the first-order definition of strong convexity in the form

$$(\forall x, y \in \mathbb{R}^n) : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \quad (2.5)$$

where $\mu \geq 0$ is a constant, called modulus of strong convexity. In particular, for $\mu = 0$ the strong convexity reduces to convexity. In this sense, strong convexity with constant μ is more general than convexity and hence a wider class of functions, denoted by $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ and called L -smooth and μ -strongly convex, follows. As a matter of fact, we have

$$\mathcal{S}_{\mu=0,L}^{1,1}(\mathbb{R}^n) = \mathcal{F}_L^{1,1}(\mathbb{R}^n).$$

Now, the following statement extends the basic fact that convexity and smoothness is equivalent to (2.3) or (2.4); for more details please refer to [10].

Fact 2.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a given real-valued function. Then, $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ if and only if one of the following inequalities holds:

$$(\forall x, y \in \mathbb{R}^n) : \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2, \quad (2.6)$$

$$\begin{aligned} (\forall x, y \in \mathbb{R}^n) : \quad f(x) \geq & f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \\ & + \frac{\mu L}{2(L - \mu)} \|x - y - \frac{1}{L}(\nabla f(x) - \nabla f(y))\|^2, \end{aligned} \quad (2.7)$$

and

$$\mu \|x - y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|. \quad (2.8)$$

At last, we let $\Gamma_0(\mathbb{R}^n)$ be the class of proper closed and convex functions from $\mathbb{R}^n$ to $(-\infty, +\infty]$. For any $g \in \Gamma_0(\mathbb{R}^n)$, its subdifferential at x is given by

$$\partial g(x) := \{y \in \mathbb{R}^n : g(u) \geq g(x) + \langle y, u - x \rangle, \quad \forall u \in \mathbb{R}^n\}.$$

The inequality $g(u) \geq g(x) + \langle y, u - x \rangle$ is called subgradient inequality, and each vector in $\partial g(x)$ is called a subgradient of g at x .

2.2 The proximal gradient method

The proximal gradient method, also called the forward-backward splitting method, is a well-known method for minimizing the sum of a smooth function and a non-smooth function. In the paper, we will be concerned with the following composite optimization

$$\min_{x \in \mathbb{R}^n} \varphi(x) := f(x) + g(x), \quad (2.9)$$

and we make the following blanket assumption.

Assumption 2.1 (Composite model assumption). The component functions f and g satisfy that

- (A) $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, i.e., f is a L -smooth and μ -strongly convex function,
- (B) $g \in \Gamma_0(\mathbb{R}^n)$, i.e., g is a proper closed convex function but it is possibly not smooth, and
- (C) X^* , the set of optimal solutions to (2.9), is nonempty. The optimal value of the problem is denoted by $\bar{\varphi}$.

Before introducing the concrete iterative scheme of the proximal gradient method, we first give the definition of proximal gradient mapping.

Definition 2.2 (PG mapping). Suppose that f and g satisfy properties (A) and (B) of Assumption 2.1. Then the proximal gradient mapping is the operator $\mathcal{G} : \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^n$ defined by

$$\mathcal{G}(x, t) := t^{-1} (x - \mathbf{prox}_{tg}(x - t\nabla f(x))), \quad (2.10)$$

where $\mathbf{prox}_{tg} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the proximal mapping given by

$$\mathbf{prox}_{tg}(x) := \arg \min_{y \in \mathbb{R}^n} \left\{ g(y) + \frac{1}{2t} \|y - x\|^2 \right\}.$$

In particular, when $g = \delta_Q$, the proximal gradient mapping reduces to the gradient mapping in [5].

Now, the proximal gradient method, originally given by

$$x^{k+1} = \mathbf{prox}_{t_k g} (x^k - t_k \cdot \nabla f(x^k)),$$

can be equivalently written into the following form

$$x^{k+1} = x^k - t_k \cdot \mathcal{G}(x^k, t_k). \quad (2.11)$$

2.3 The potential function-based framework

The authors of [3] introduced a novel potential function-based framework to study the convergence of standard gradient-type methods for making the gradients small in smooth convex optimization. In this part, we first review how their method applies to the standard gradient descent for minimizing a smooth convex function $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$. The key ingredient is that they constructed a potential function of the form

$$\mathcal{C}_k = \frac{k}{L} \|\nabla f(x^k)\|^2 + f(x^k),$$

where the sequence $\{x^k\}_{k \geq 0}$ is generated by the standard gradient descent method, i.e.,

$$x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k), \quad \forall k \geq 0.$$

By invoking the inequalities (2.3) or (2.4), they can show that the sequence $\{\mathcal{C}_k\}_{k \geq 0}$ is nonincreasing with respect to k and hence can conclude that $\forall k \geq 0$,

$$\|\nabla f(x^k)\|^2 \leq \frac{2L(f(x^0) - f(x^*))}{2k + 1}, \quad (2.12)$$

where x^0 is an arbitrary initial point and x^* is a minimizer of f . In order to design a faster method than the standard gradient descent, they considered a different potential function of the form

$$\mathcal{C}_k = \sum_{i=0}^{k-1} a_i \|\nabla f(x^i)\|^2 + B_k (f(x^k) - f(x^*)), \quad (2.13)$$

where $a_i > 0$ ($\forall i \geq 0$) the sequence of scalars $B_k > 0$ ($\forall k \geq 0$) is strictly increasing, and the sequence $\{x^k\}_{k \geq 0}$ is generated by the following fast gradient method (FGM)

$$\begin{cases} v^k := v^{k-1} - \frac{b_{k-1}}{L} \cdot \nabla f(x^{k-1}), \\ x^k := \frac{B_{k-1}}{B_k} \left(x^{k-1} - \frac{1}{L} \cdot \nabla f(x^{k-1}) \right) + \frac{b_k}{B_k} v^k, \end{cases} \quad (\text{FGM})$$

with a given arbitrary initial point x^0 and $v^0 = x^0$. Under some restrictions on the parameters a_i and B_k , invoking again the inequalities (2.3) and (2.4) they showed that

$$\mathcal{C}_{k+1} - \mathcal{C}_k \leq \frac{L}{2} (\|x^* - v^k\|^2 - \|x^* - v^{k+1}\|^2), \quad \forall k \geq 0,$$

from which both convergences in function value and in norm of gradient can be obtained. As pointed out in [3], their analysis is the first one that simultaneously leads to both tight convergence guarantees.

3 New Properties on Proximal Gradient Mapping

In this section, we first introduce three basic properties of proximal gradient mapping, whose proofs postpone to Appendix. Then, we highlight two new properties, both of which were discovered in [11] by the second author of this paper and posted on arXiv three years ago, but they have not yet been submitted for publication.

3.1 Basic lemmas

The first lemma is an equivalent characterization of the proximal mapping.

Lemma 3.1. *Let $g \in \Gamma_0(\mathbb{R}^n)$ and $t > 0$. Thus, $z = \mathbf{prox}_{tg}(y)$ if and only if $y \in (I + t \cdot \partial g)(z)$.*

The second lemma provides the relationship between the norm of proximal gradient mapping and the smallest norm of subgradient.

Lemma 3.2. *Suppose that f and g satisfy properties (A) and (B) of Assumption 2.1. For any $x \in \mathbb{R}^n$ and $t > 0$, we have*

$$\|\mathcal{G}(x, t)\| \leq d(0, \partial\varphi(x)). \quad (3.1)$$

The last lemma is a slight modification of the classic descent lemma, originally discovered by Beck and Teboulle in [1]. It also extends Corollary 2.3.2 in [5] from gradient mapping to proximal gradient mapping. When $\mu = 0$, it reduces to the pivotal inequality in the recent work [4].

Lemma 3.3. *Suppose that f and g satisfy properties (A) and (B) of Assumption 2.1. Then, we have*

$$\varphi(x) - \varphi(y - t\mathcal{G}(y, t)) \geq t \left(1 - \frac{L}{2}t\right) \|\mathcal{G}(y, t)\|^2 + \langle \mathcal{G}(y, t), x - y \rangle + \frac{\mu}{2} \|x - y\|^2. \quad (3.2)$$

In particular, the inequality above with $t = \frac{1}{L}$ and $\mu = 0$ in (3.2) yields

$$\varphi(x) - \varphi(y - \frac{1}{L}\mathcal{G}(y, \frac{1}{L})) \geq \frac{1}{2L} \left\| \mathcal{G}(y, \frac{1}{L}) \right\|^2 + \left\langle \mathcal{G}(y, \frac{1}{L}), x - y \right\rangle. \quad (3.3)$$

3.2 New and refined results

For simplicity, we define the updated iterate point by using the superscript "+" as follows:

$$x^+ := \mathbf{prox}_{tg}(x - t\nabla f(x)) = x - t \cdot \mathcal{G}(x, t),$$

where the step size $t > 0$ is clear from the context. Using this notation and Lemma 3.1, we immediately have

$$x - t\nabla f(x) \in x^+ + t\partial g(x^+).$$

Thus, there must exist a subgradient $s^+ \in \partial g(x^+)$ such that

$$x^+ = x - t(\nabla f(x) + s^+). \quad (3.4)$$

Now, we are ready to present the first new property of proximal gradient mapping.

Theorem 3.4 (Norm monotonicity). *Suppose that f and g satisfy properties (A) and (B) of Assumption 2.1. Denote $\rho(t) := \max\{|1 - Lt|, |1 - \mu t|\}$. Then, we have*

$$\|\mathcal{G}(x^+, t)\| \leq d(0, \partial\varphi(x^+)) \leq \rho(t)\|\mathcal{G}(x, t)\| \leq \rho(t)d(0, \partial\varphi(x)). \quad (3.5)$$

In particular, for $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, $g \in \Gamma_0(\mathbb{R}^n)$, and $0 < t \leq \frac{2}{L}$, it holds that

$$\|\mathcal{G}(x^+, t)\| \leq d(0, \partial\varphi(x^+)) \leq \|\mathcal{G}(x, t)\| \leq d(0, \partial\varphi(x)).$$

Proof. The inequalities $\|\mathcal{G}(x^+, t)\| \leq d(0, \partial\varphi(x^+))$ and $\rho(t)\|\mathcal{G}(x, t)\| \leq \rho(t)d(0, \partial\varphi(x))$ directly follow from Lemma 3.2. To show the relationship (3.5), it suffices to show that

$$d(0, \partial\varphi(x^+)) \leq \rho(t)\|\mathcal{G}(x, t)\|. \quad (3.6)$$

Since $s^+ \in \partial g(x^+)$, we have $d(0, \partial\varphi(x^+)) \leq \|\nabla f(x^+) + s^+\|$. Therefore, if we can show that

$$\|\nabla f(x^+) + s^+\|^2 \leq \rho^2(t)\|\mathcal{G}(x, t)\|^2, \quad (3.7)$$

then the desired inequality (3.6) follows immediately. Using the expression $x^+ = x - t(\nabla f(x) + s^+)$ in (3.4), we derive that

$$\begin{aligned} & \|\nabla f(x^+) + s^+\|^2 \\ &= \|\nabla f(x) + s^+ + \nabla f(x^+) - \nabla f(x)\|^2 \\ &= \|\nabla f(x) + s^+\|^2 + 2\langle \nabla f(x) + s^+, \nabla f(x^+) - \nabla f(x) \rangle + \|\nabla f(x^+) - \nabla f(x)\|^2 \\ &= \frac{1}{t^2}\|x^+ - x\|^2 - \frac{2}{t}\langle x^+ - x, \nabla f(x^+) - \nabla f(x) \rangle + \|\nabla f(x^+) - \nabla f(x)\|^2 \\ &\leq \frac{1}{t^2}\|x^+ - x\|^2 - \frac{2}{t}\left(\frac{\mu L}{\mu + L}\|x^+ - x\|^2 + \frac{1}{\mu + L}\|\nabla f(x^+) - \nabla f(x)\|^2\right) + \|\nabla f(x^+) - \nabla f(x)\|^2 \\ &= \frac{1}{t^2}\left[\left(1 - \frac{2t\mu L}{\mu + L}\right)\|x^+ - x\|^2 + t\left(t - \frac{2}{\mu + L}\right)\|\nabla f(x^+) - \nabla f(x)\|^2\right], \end{aligned}$$

where the inequality follows from (2.6) in Fact 2.1. In order to bound $\|\nabla f(x^+) - \nabla f(x)\|^2$ in terms of $\|x^+ - x\|^2$, we use (2.8) in Fact 2.1 to get

$$\mu^2\|x^+ - x\|^2 \leq \|\nabla f(x^+) - \nabla f(x)\|^2 \leq L^2\|x^+ - x\|^2.$$

If $t - \frac{2}{\mu + L} \geq 0$, then we have

$$\left(t - \frac{2}{\mu + L}\right)\|\nabla f(x^+) - \nabla f(x)\|^2 \leq L^2\left(t - \frac{2}{\mu + L}\right)\|x^+ - x\|^2.$$

If $t - \frac{2}{\mu + L} < 0$, then we have

$$\left(t - \frac{2}{\mu + L}\right)\|\nabla f(x^+) - \nabla f(x)\|^2 \leq \mu^2\left(t - \frac{2}{\mu + L}\right)\|x^+ - x\|^2.$$

In both cases, we always have that

$$\left(t - \frac{2}{\mu + L}\right)\|\nabla f(x^+) - \nabla f(x)\|^2 \leq \max\left\{L^2\left(t - \frac{2}{\mu + L}\right), \mu^2\left(t - \frac{2}{\mu + L}\right)\right\}\|x^+ - x\|^2.$$

Therefore, we can continue to derive that

$$\begin{aligned}
& \|\nabla f(x^+) + s^+\|^2 \\
& \leq \frac{1}{t^2} \left[\left(1 - \frac{2t\mu L}{\mu + L}\right) \|x^+ - x\|^2 + t \cdot \max \left\{ L^2 \left(t - \frac{2}{\mu + L}\right), \mu^2 \left(t - \frac{2}{\mu + L}\right) \right\} \|x^+ - x\|^2 \right] \\
& = \frac{1}{t^2} \max \left\{ 1 - \frac{2t\mu L}{\mu + L} + tL^2 \left(t - \frac{2}{\mu + L}\right), 1 - \frac{2t\mu L}{\mu + L} + t\mu^2 \left(t - \frac{2}{\mu + L}\right) \right\} \|x^+ - x\|^2 \\
& = \frac{1}{t^2} \max \{(1 - Lt)^2, (1 - \mu t)^2\} \|x^+ - x\|^2 \\
& = \rho^2(t) \|\mathcal{G}(x, t)\|^2,
\end{aligned}$$

from which the inequality (3.7) follows. This completes the proof. $\square$

Remark 3.5. Here, the factor $\rho(t)$ is optimal; otherwise, it will contradict the following exact worst-case convergence rate for residual gradient norm

$$\|\nabla f(x^+) + s^+\| \leq \rho(t) \|\nabla f(x) + s\|, \quad \forall s \in \partial g(x),$$

which was recently established in [8]. In fact, the inequality above is equivalent to

$$\|\nabla f(x^+) + s^+\| \leq \rho(t) d(0, \partial \varphi(x));$$

whilst in our proof, we have shown $\|\nabla f(x^+) + s^+\| \leq \rho(t) \|\mathcal{G}(x, t)\|$ in (3.7) which is a tighter estimation and hence it is impossible to improve.

Below, we state the second new property of proximal gradient mapping, which may be viewed as a refined descent lemma.

Theorem 3.6 (Refined descent). *Suppose that f and g satisfy properties (A) and (B) of Assumption 2.1. Then, we have*

$$\varphi(x) \geq \varphi(x^+) + \frac{t}{2} \|\mathcal{G}(x, t)\|^2 + \frac{t}{2(1 - \mu t)} \|\mathcal{G}(x^+, t)\|^2, \quad 0 < t \leq \frac{1}{L}. \quad (3.8)$$

In particular,

for $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, $g \in \Gamma_0(\mathbb{R}^n)$, it holds that

$$\varphi(x) \geq \varphi(x^+) + \frac{t}{2} \|\mathcal{G}(x, t)\|^2 + \frac{t}{2} \|\mathcal{G}(x^+, t)\|^2, \quad 0 < t \leq \frac{1}{L}. \quad (3.9)$$

for $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, $g \equiv 0$, it holds that

$$f(x) \geq f(x^+) + \frac{t}{2} \|\nabla f(x)\|^2 + \frac{t}{2} \|\nabla f(x^+)\|^2, \quad 0 < t \leq \frac{1}{L}. \quad (3.10)$$

Proof. Note that $0 < t \leq L^{-1}$ implies $t^{-1} \geq L$ which further implies that the L -smooth function must also be t^{-1} -smooth; hence, we can conclude that

$$\mathcal{S}_{\mu, L}^{1,1}(\mathbb{R}^n) \subset \mathcal{S}_{\mu, t^{-1}}^{1,1}(\mathbb{R}^n).$$

We now use (2.7) in Fact 2.1 with $L = t^{-1}$ and $y = x^+$ to get

$$\begin{aligned} f(x) &\geq f(x^+) + \langle \nabla f(x^+), x - x^+ \rangle + \frac{t}{2} \|\nabla f(x) - \nabla f(x^+)\|^2 \\ &\quad + \frac{\mu}{2(1 - \mu t)} \|x - x^+ - t(\nabla f(x) - \nabla f(x^+))\|^2. \end{aligned}$$

The subgradient inequality of g gives $g(x) \geq g(x^+) + \langle s^+, x - x^+ \rangle$ since $s^+ \in \partial g(x^+)$. Summing up these two inequalities, we derive that

$$\begin{aligned} \varphi(x) &\geq \varphi(x^+) + \langle \nabla f(x^+) + s^+, x - x^+ \rangle + \frac{t}{2} \|\nabla f(x) - \nabla f(x^+)\|^2 \\ &\quad + \frac{\mu}{2(1 - \mu t)} \|x - x^+ - t(\nabla f(x) - \nabla f(x^+))\|^2 \\ &= \varphi(x^+) + \langle \nabla f(x) + s^+, x - x^+ \rangle - \langle \nabla f(x^+) - \nabla f(x), x^+ - x \rangle \\ &\quad + \frac{t}{2} \|\nabla f(x) - \nabla f(x^+)\|^2 + \frac{\mu}{2(1 - \mu t)} \|x - x^+ - t(\nabla f(x) - \nabla f(x^+))\|^2. \end{aligned}$$

Using the expression $x^+ = x - t(\nabla f(x) + s^+)$ in (3.4), we can further derive that

$$\begin{aligned} \varphi(x) &\geq \varphi(x^+) + \frac{1}{t} \|x - x^+\|^2 - \langle \nabla f(x^+) - \nabla f(x), x^+ - x \rangle \\ &\quad + \frac{t}{2} \|\nabla f(x) - \nabla f(x^+)\|^2 + \frac{\mu t^2}{2(1 - \mu t)} \|s^+ + \nabla f(x^+)\|^2 \\ &= \varphi(x^+) + \frac{1}{2t} \|t(\nabla f(x^+) - \nabla f(x)) - x^+ + x\|^2 \\ &\quad + \frac{1}{2t} \|x - x^+\|^2 + \frac{\mu t^2}{2(1 - \mu t)} \|s^+ + \nabla f(x^+)\|^2 \\ &= \varphi(x^+) + \frac{1}{2t} \|x - x^+\|^2 + \frac{t}{2(1 - \mu t)} \|s^+ + \nabla f(x^+)\|^2. \end{aligned}$$

Note that $x - x^+ = t\mathcal{G}(x, t)$ and use the fact that

$$\|s^+ + \nabla f(x^+)\| \geq d(0, \partial\varphi(x^+)) \geq \|\mathcal{G}(x^+, t)\|.$$

We finally obtain

$$\varphi(x) \geq \varphi(x^+) + \frac{t}{2} \|\mathcal{G}(x, t)\|^2 + \frac{t}{2(1 - \mu t)} \|\mathcal{G}(x^+, t)\|^2.$$

This completes the proof. $\square$

Remark 3.7. We make a few remarks:

In [5], for $\varphi = f + g$ with $f \in \mathcal{S}_{\mu, L}^{1,1}(\mathbb{R}^n)$ and g being the indicator function of a set Q , the descent lemma of the projected gradient method can be stated as

$$\varphi(x) \geq \varphi(x^+) + \frac{t}{2} \|g_Q(x, t)\|^2, 0 < t \leq \frac{1}{L}. \quad (3.11)$$

where $g_Q(x, t) := t^{-1}(x - x^+)$ is the gradient mapping of f on Q . In [1], for $\varphi = f + g$ with $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $g \in \Gamma_0(\mathbb{R}^n)$, the corresponding descent lemma of the proximal gradient method is

$$\varphi(x) \geq \varphi(x^+) + \frac{L}{2} \|x^+ - x\|^2. \quad (3.12)$$

It is not hard to see that our result improves these existing descent lemmas.

In time of this paper was under preparation, we noticed that the special case (3.9) was implicitly rediscovered by combining Lemma 9 and Lemma 11 in [9].

4 Small Norm of Proximal Gradient Mapping

In this section, we aim to extend the potential function-based framework previously reviewed from gradient descent to proximal gradient descent and its acceleration, for minimizing the sum of a smooth convex function and a possibly nonsmooth convex function.

4.1 Proximal gradient descent

The following result is a direct extension of Lemma 2.1 in [3]. However, its proof relies on the new properties of proximal gradient mapping in the last section.

Theorem 4.1. *Suppose that Assumption 2.1 with $\mu = 0$ holds. Let x^0 be an arbitrary initial point and assume that $x^{k+1} = x^k - t_k \mathcal{G}(x^k, t_k)$ with constant step sizes $t_k \equiv \frac{\eta}{L}$ for some $0 < \eta \leq 1$. Then*

$$\mathcal{C}_k := \frac{\eta}{L} \cdot k \|\mathcal{G}(x^k, \frac{\eta}{L})\|^2 + \varphi(x^k)$$

is nonincreasing with k , and the norm sequence of proximal gradient mappings converges sublinearly in the sense that for any $k \geq 0$,

$$\|\mathcal{G}(x^k, \frac{\eta}{L})\| \leq \sqrt{\frac{L(\varphi(x^0) - \bar{\varphi})}{\eta k}}. \quad (4.1)$$

Proof. We first show that for any $k \geq 0$,

$$\mathcal{C}_{k+1} \leq \mathcal{C}_k.$$

Using the definition of $\mathcal{C}_k$, we have that

$$\mathcal{C}_{k+1} - \mathcal{C}_k = \frac{\eta}{L} (k+1) \|\mathcal{G}(x^{k+1}, \frac{\eta}{L})\|^2 - \frac{\eta k}{L} \|\mathcal{G}(x^k, \frac{\eta}{L})\|^2 + \varphi(x^{k+1}) - \varphi(x^k).$$

Using Theorem 3.6 yields

$$\varphi(x^k) - \varphi(x^{k+1}) \geq \frac{\eta}{2L} \|\mathcal{G}(x^k, \frac{\eta}{L})\|^2 + \frac{\eta}{2L} \|\mathcal{G}(x^{k+1}, \frac{\eta}{L})\|^2.$$

Thus,

$$\mathcal{C}_{k+1} - \mathcal{C}_k \leq \frac{\eta}{L} \left(k + \frac{1}{2} \right) \left(\|\mathcal{G}(x^{k+1}, \frac{\eta}{L})\|^2 - \|\mathcal{G}(x^k, \frac{\eta}{L})\|^2 \right).$$

In addition, using Theorem 3.4 yields

$$\|\mathcal{G}(x^{k+1}, \frac{\eta}{L})\| \leq \|\mathcal{G}(x^k, \frac{\eta}{L})\|,$$

which leads to the monotonically decreasing $\mathcal{C}_{k+1} \leq \mathcal{C}_k$ and the result

$$\varphi(x^k) - \bar{\varphi} + \frac{\eta k}{L} \cdot \|\mathcal{G}(x^k, \frac{\eta}{L})\|^2 \leq \dots \leq \mathcal{C}_0 = \varphi(x^0) - \bar{\varphi}.$$

Equivalently,

$$\frac{\eta k}{L} \|\mathcal{G}(x^k, \frac{\eta}{L})\|^2 \leq \varphi(x^0) - \varphi(x^k) \leq \varphi(x^0) - \bar{\varphi},$$

from which the conclusion follows. $\square$

Remark 4.2. The authors of the recent paper [4] established the following result for the proximal gradient method:

$$\min_{0 \leq i \leq k} \|\mathcal{G}(x^i, \frac{\eta}{L})\| \leq \sqrt{\frac{2}{3k(k+1)}} \frac{L\|x^0 - x^*\|}{\eta}, \quad (4.2)$$

where $x^* \in X^*$. At the first glance, it seems that our bound (4.1) is worse than the bound (4.2) because the latter improves the order of complexity from $O(\frac{1}{\sqrt{k}})$ to $O(\frac{1}{k})$. But in fact, combining the boundedness of $\mathcal{C}_k$ and the standard convergence result $\varphi(x^k) - \varphi(x^*) \leq O\left(\frac{L\|x^0 - x^*\|^2}{\eta k}\right)$ (see e.g. Theorem 10.21 [2]), we derive that

$$\|\mathcal{G}(x^k, \frac{\eta}{L})\|^2 = O\left(\frac{L(\varphi(x^k) - \varphi(x^*))}{\eta k}\right) = O\left(\frac{L^2\|x^0 - x^*\|^2}{\eta^2 k^2}\right),$$

and hence the better bound of the order $O(\frac{1}{k})$ can be reobtained; similar deduction can be found on page 10 in [3]. The main advantage of our result lies in removing the operation of taking the minimum of proximal gradient. Finally, it should be noted that the bound (4.1) is tight since it generalizes the tight bound (2.12).

4.2 Accelerated norm minimization

We start with the following iterative scheme which is obtained by replacing the gradient in the fast gradient method by the proximal gradient mapping and introducing a new sequence $\{y^k\}$. For any $k \geq 1$,

$$\begin{cases} y^{k-1} := x^{k-1} - \frac{1}{L} \cdot \mathcal{G}(x^{k-1}, \frac{1}{L}), \\ v^k := v^{k-1} - \frac{b_{k-1}}{L} \cdot \mathcal{G}(x^{k-1}, \frac{1}{L}), \\ x^k := \frac{B_{k-1}}{B_k} y^{k-1} + \frac{b_k}{B_k} v^k, \end{cases} \quad (\text{APG})$$

where the sequence of scalars $B_k > 0$ will be determined later and the sequence of scalars b_k is defined by $b_0 = B_0$ and $b_k = B_k - B_{k-1}$ for $k \geq 1$. For simplicity, we let $\mathcal{G}(x^k) \equiv \mathcal{G}(x^k, \frac{1}{L})$ be the proximal gradient mapping when the step size t equals to $\frac{1}{L}$. Our forthcoming analysis mainly relies on the following potential function: for any $k \geq 0$,

$$\mathcal{C}_k := \sum_{i=0}^k a_i \|\mathcal{G}(x^i)\|^2 + B_k(\varphi(y^k) - \bar{\varphi}), \quad (4.3)$$

which is inspired by the potential function (2.13). However, when zooming into the expression more carefully, the reader can find that it is not obtained by simply replacing the gradient in (2.13) by the proximal gradient mapping. Actually, we use the function value at y^k rather than at x^k and the sum is from $i = 0$ to k rather than to $k-1$. These modifications are pivotal to deduce our desired conclusions.

Lemma 4.3. *Suppose that Assumption 2.1 with $\mu = 0$ holds. Let x^0 be an arbitrary initial point with $v^0 = x^0$ and assume that the sequences of $\{x^k\}$, $\{y^k\}$ and $\{v^k\}$ are generated by the algorithm (APG). If the nonnegative scalars a_k, b_k, B_k satisfy that $\forall k \geq 1$,*

$$a_k \leq \frac{B_k - b_k^2}{2L},$$

then we have

$$\mathcal{C}_k - \mathcal{C}_{k-1} \leq \frac{L}{2}(\|x^* - v^k\|^2 - \|x^* - v^{k+1}\|^2), \forall k \geq 1,$$

where $x^* \in X^*$.

Proof. Using the definition of $\mathcal{C}_k$ in (4.3), we have that for any $k \geq 1$,

$$\mathcal{C}_k - \mathcal{C}_{k-1} \leq a_k \|\mathcal{G}(x^k)\|^2 + B_k \varphi(y^k) - B_{k-1} \varphi(y^{k-1}) - b_k \bar{\varphi}. \quad (4.4)$$

Now, we use (3.3) in Lemma 3.3 to bound the unknown optimal function value $\bar{\varphi}$. Actually, the inequality (3.3) with $x = x^*$ and $y = x^k$ gives us

$$\bar{\varphi} = \varphi(x^*) \geq \varphi(y^k) + \frac{1}{2L} \|\mathcal{G}(x^k)\|^2 + \langle \mathcal{G}(x^k), x^* - x^k \rangle. \quad (4.5)$$

Using (3.3) again with $x = y^{k-1}$ and $y = x^k$ leads to

$$\varphi(y^{k-1}) - \varphi(y^k) \geq \frac{1}{2L} \|\mathcal{G}(x^k)\|^2 + \langle \mathcal{G}(x^k), y^{k-1} - x^k \rangle. \quad (4.6)$$

Combining the three inequalities above, we derive that

$$\begin{aligned} \mathcal{C}_k - \mathcal{C}_{k-1} &\leq a_k \|\mathcal{G}(x^k)\|^2 + B_k \varphi(y^k) - B_{k-1} \varphi(y^{k-1}) - b_k \varphi(y^k) - \frac{b_k}{2L} \|\mathcal{G}(x^k)\|^2 \\ &\quad - b_k \langle \mathcal{G}(x^k), x^* - x^k \rangle \\ &= a_k \|\mathcal{G}(x^k)\|^2 + B_{k-1} (\varphi(y^k) - \varphi(y^{k-1})) - \frac{b_k}{2L} \|\mathcal{G}(x^k)\|^2 \\ &\quad - b_k \langle \mathcal{G}(x^k), x^* - x^k \rangle \\ &\leq a_k \|\mathcal{G}(x^k)\|^2 - \frac{B_{k-1}}{2L} \|\mathcal{G}(x^k)\|^2 - B_{k-1} \langle \mathcal{G}(x^k), y^{k-1} - x^k \rangle - \frac{b_k}{2L} \|\mathcal{G}(x^k)\|^2 \\ &\quad + b_k \langle \mathcal{G}(x^k), x^k - x^* \rangle \\ &= \left(a_k - \frac{B_k}{2L} \right) \|\mathcal{G}(x^k)\|^2 + B_{k-1} \left\langle \mathcal{G}(x^k), x^k - x^{k-1} + \frac{1}{L} \mathcal{G}(x^{k-1}) \right\rangle \\ &\quad + b_k \langle \mathcal{G}(x^k), x^k - x^* \rangle. \end{aligned} \quad (4.7)$$

In order to get an acceptable upper bound of $\mathcal{C}_k - \mathcal{C}_{k-1}$, we need to estimate the inner product term $\langle \mathcal{G}(x^k), x^k - x^* \rangle$. This can be done by going through the following arguments which are standard in mirror-descent-type analysis. First, note that

$$\begin{aligned} v^{k+1} &= \arg \min_u \left\{ b_k \langle \mathcal{G}(x^k), u - v^k \rangle + \frac{L}{2} \|u - v^k\|^2 \right\} \\ &= v^k - \frac{b_k}{L} \mathcal{G}(x^k). \end{aligned}$$

Then, we can deduce that

$$\begin{aligned} b_k \langle \mathcal{G}(x^k), x^k - x^* \rangle &= b_k \langle \mathcal{G}(x^k), x^k - v^{k+1} \rangle + L \langle v^k - v^{k+1}, v^{k+1} - x^* \rangle \\ &= b_k \langle \mathcal{G}(x^k), x^k - v^k \rangle + \frac{b_k^2}{L} \|\mathcal{G}(x^k)\|^2 + \frac{L}{2} \|x^* - v^k\|^2 \\ &\quad - \frac{L}{2} \|x^* - v^{k+1}\|^2 - \frac{L}{2} \|v^{k+1} - v^k\|^2 \\ &= b_k \langle \mathcal{G}(x^k), x^k - v^k \rangle + \frac{b_k^2}{2L} \|\mathcal{G}(x^k)\|^2 \\ &\quad + \frac{L}{2} \|x^* - v^k\|^2 - \frac{L}{2} \|x^* - v^{k+1}\|^2, \end{aligned} \quad (4.8)$$

where the relationship $v^{k+1} := v^k - \frac{b_k}{L} \cdot \mathcal{G}(x^k, \frac{1}{L})$ have been repeatedly used. Now, combining (4.8) and (4.7), we can get

$$\begin{aligned} \mathcal{C}_k - \mathcal{C}_{k-1} &\leq \left(a_k - \frac{B_k - b_k^2}{2L} \right) \|\mathcal{G}(x^k)\|^2 + \frac{L}{2} \|x^* - v^k\|^2 - \frac{L}{2} \|x^* - v^{k-1}\|^2 \\ &\quad + \left\langle \mathcal{G}(x^k), B_k x^k - B_{k-1}(x^{k-1} - \frac{1}{L} \mathcal{G}(x^{k-1})) - b_k v^k \right\rangle. \end{aligned} \quad (4.9)$$

Note that

$$B_k x^k - B_{k-1} \left(x^{k-1} - \frac{1}{L} \mathcal{G}(x^{k-1}) \right) - b_k v^k = B_k x^k - B_{k-1} y^{k-1} - b_k v^k = 0.$$

The inner product term (4.9) disappears and hence using the condition $a_k \leq \frac{B_k - b_k^2}{2L}$ we finally obtain

$$\mathcal{C}_k - \mathcal{C}_{k-1} \leq \frac{L}{2} \|x^* - v^k\|^2 - \frac{L}{2} \|x^* - v^{k+1}\|^2.$$

This completes the proof. $\square$

Now, we are ready to present the accelerated convergence of proximal gradient mapping for minimizing the sum of a smooth convex function and a non-smooth convex function.

Theorem 4.4. *Suppose that the assumption in Lemma 4.3 holds. Denote*

$$\tilde{\mathcal{C}} := a_0 \|\mathcal{G}(x^0)\|^2 + b_0 (\varphi(y^0) - \bar{\varphi}) + \frac{L}{2} \|x^* - v^0\|^2.$$

Then, we have

$$\varphi(y^k) - \bar{\varphi} \leq \frac{\tilde{\mathcal{C}}}{B_k}, k \geq 1, \quad (4.10)$$

$$\sum_{i=0}^k a_i \|\mathcal{G}(x^i)\|^2 \leq \tilde{\mathcal{C}}, k \geq 1. \quad (4.11)$$

In particular, if $b_k = \frac{1}{4}(k+1)$, $B_k = \frac{1}{8}(k+1)(k+2)$, $a_k = \frac{1}{32L}(k+1)^2$ for any $k \geq 1$, then

$$\varphi(y^k) - \bar{\varphi} \leq \frac{8\tilde{\mathcal{C}}}{(k+1)(k+2)}, \quad (4.12)$$

and

$$\min_{0 \leq i \leq k} \|\mathcal{G}(x^i)\|^2 \leq \frac{192L\tilde{\mathcal{C}}}{(k+1)(k+2)(k+3)}. \quad (4.13)$$

Proof. Using Lemma 4.3 and the definition $\mathcal{C}_k$, we have

$$\begin{aligned} \mathcal{C}_k &\leq \mathcal{C}_0 + \frac{L}{2} \|x^* - v^0\|^2 - \frac{L}{2} \|x^* - v^{k+1}\|^2 \\ &\leq a_0 \|\mathcal{G}(x^0)\|^2 + B_0 (\varphi(y^0) - \bar{\varphi}) + \frac{L}{2} \|x^* - v^0\|^2 \\ &= \tilde{\mathcal{C}}. \end{aligned} \quad (4.14)$$

Note that each term in $\mathcal{C}_k$ is nonnegative. Thus, for any $k \geq 1$ we can get

$$B_k (\varphi(y^k) - \bar{\varphi}) \leq \mathcal{C}_k \leq \tilde{\mathcal{C}},$$

and

$$\sum_{i=0}^k a_i \|\mathcal{G}(x^i)\|^2 \leq \mathcal{C}_k \leq \tilde{\mathcal{C}},$$

from which the first part follows.

As for the second part, we first show that the condition $a_k \leq \frac{B_k - b_k^2}{2L}$ can be verified by the current setting $b_k = \frac{1}{4}(k+1)$, $B_k = \frac{1}{8}(k+1)(k+2)$, and $a_k = \frac{1}{32L}(k+1)^2$. In fact,

$$\begin{aligned} \frac{B_k - b_k^2}{2L} &= \frac{1}{2L} \left[\frac{1}{8}(k+1)^2 + \frac{1}{8}(k+1) - \frac{1}{16}(k+1)^2 \right] \\ &\geq \frac{1}{2L} \cdot \frac{1}{16}(k+1)^2 \\ &= a_k. \end{aligned}$$

Now, summing a_i from $i = 0$ to $i = k$, we obtain

$$\sum_{i=0}^k a_i = \sum_{i=0}^k \frac{1}{32L} (i+1)^2 = \frac{(k+1)(k+2)(2k+3)}{192L}.$$

Therefore, combining with (4.11) in the first part, we finally get

$$\min_{0 \leq i \leq k} \|\mathcal{G}(x^i)\|^2 \leq \frac{\sum_{i=0}^k a_i \|\mathcal{G}(x^i)\|^2}{\sum_{i=0}^k a_i} \leq \frac{192L \cdot \tilde{\mathcal{C}}}{(k+1)(k+2)(2k+3)}, \quad (4.15)$$

which completes the proof. $\square$

Remark 4.5. Again, we would like to point out that the bound (4.15) is tight since it generalizes the corresponding tight bound in [3].

5 Concluding Remarks

In this paper, we successfully extended the potential function-based framework in [3] from gradient descent to proximal gradient descent, with the help of two newly discovered properties on the proximal gradient mapping. However, the modulus $\mu > 0$ of strong convexity has not yet been exploited in the current potential function-based framework to provide linear convergence guarantees for the norm of gradient or proximal gradient mapping; we would like to leave it as future work.

Acknowledgements

The authors wish to express their thanks to the anonymous referees and the associate editor for valuable suggestions and comments, which allowed us to improve the original presentation.

Appendix: The missing proofs

The proof of Lemma 3.1: Using the definition of the proximal mapping yields

$$z = \mathbf{prox}_{tg}(y) = \arg \min_x \left\{ t \cdot g(x) + \frac{1}{2} \|x - y\|^2 \right\}.$$

Based on the first-order optimality condition, we have

$$0 \in t \cdot \partial g(z) + z - y.$$

Hence, the relationship $y \in (I + t \cdot \partial g)(z)$ follows. This completes the proof.

The proof of Lemma 3.2: Take a subgradient $s \in \partial \varphi(x) = \partial g(x) + \nabla f(x)$; then, it holds that

$$x - t \nabla f(x) + ts \in (I + t \cdot \partial g)(x).$$

Hence, from Lemma 3.1 we have

$$x = \mathbf{prox}_{tg}(x - t \nabla f(x) + ts).$$

Using the nonexpansive property of proximal mapping, for any $s \in \partial \varphi(x)$ we have

$$\begin{aligned} t \|\mathcal{G}(x, t)\| &= \|x - \mathbf{prox}_{tg}(x - t \nabla f(x))\| \\ &= \|\mathbf{prox}_{tg}(x - t \nabla f(x) + ts) - \mathbf{prox}_{tg}(x - t \nabla f(x))\| \\ &\leq t \|s\|, \quad \forall s \in \partial \varphi(x), \end{aligned}$$

from which the upper bound (3.1) follows. This completes the proof.

The proof of Lemma 3.3: First of all, we define the following auxiliary function

$$h(x, y) := g(x) + f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2t} \|x - y\|^2.$$

Denote

$$y^+ := \arg \min_x h(x, y). \tag{A.1}$$

Then, one can verify that

$$y^+ = y - t \mathcal{G}(y, t).$$

Applying the L -smoothness of f in (2.2), we obtain

$$\varphi(x) = f(x) + g(x) \leq h(x, y) + \left(\frac{L}{2} - \frac{1}{2t}\right) \|x - y\|^2.$$

Plugging $x = y^+$ in the above equation, we get

$$\varphi(y^+) \leq h(y^+, y) + \left(\frac{L}{2} - \frac{1}{2t}\right) \|y^+ - y\|^2,$$

or equivalently,

$$\varphi(x) - \varphi(y^+) \geq \varphi(x) - h(y^+, y) - \left(\frac{L}{2} - \frac{1}{2t}\right) \|y^+ - y\|^2. \tag{A.2}$$

Due to the optimality condition of (A.1), there must exist a subgradient $g_s \in \partial g(y^+)$ such that

$$0 = g_s + \nabla f(y) + \frac{1}{t}(y^+ - y). \quad (\text{A.3})$$

Invoking the subgradient inequality for g and the μ -strong convexity for f , we have

$$\begin{aligned} f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2, \\ g(x) &\geq g(y^+) + \langle g_s, x - y^+ \rangle. \end{aligned}$$

Adding these two inequalities together, we get

$$\varphi(x) \geq f(y) + g(y^+) + \langle \nabla f(y), x - y \rangle + \langle g_s, x - y^+ \rangle + \frac{\mu}{2} \|x - y\|^2. \quad (\text{A.4})$$

On the other hand,

$$h(y^+, y) = g(y^+) + f(y) + \langle y^+ - y, \nabla f(y) \rangle + \frac{1}{2t} \|y^+ - y\|^2.$$

Combining the preceding equation with (A.2) and (A.4), we finally get

$$\begin{aligned} \varphi(x) - \varphi(y^+) &\geq \varphi(x) - h(y^+, y) - \left(\frac{L}{2} - \frac{1}{2t}\right) \|y^+ - y\|^2 \\ &\geq -\frac{1}{2t} \|y^+ - y\|^2 + \langle x - y^+, \nabla f(y) + g_s \rangle - \left(\frac{L}{2} - \frac{1}{2t}\right) \|y^+ - y\|^2 + \frac{\mu}{2} \|x - y\|^2 \\ &\stackrel{(\text{A.3})}{=} -\frac{1}{2t} \|y^+ - y\|^2 + \frac{1}{t} \langle y - y^+, x - y^+ \rangle - \left(\frac{L}{2} - \frac{1}{2t}\right) \|y^+ - y\|^2 + \frac{\mu}{2} \|x - y\|^2. \end{aligned}$$

The desired conclusion follows by substituting $\mathcal{G}(y, t) = t^{-1}(y - y^+)$ into the above relationship. This completes the proof.

References

- [1] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* 2 (2009), 183–202.
- [2] A. Beck, *First-order methods in optimization*, SIAM, Philadelphia, 2017.
- [3] J. Diakonikolas, P. Wang, Potential function-based framework for minimizing gradients in convex and min-max optimization, *SIAM J. Optim.* 32 (2022), 1668–1697.
- [4] B. Li, B. Shi, Y.X. Yuan, Proximal subgradient norm minimization of ISTA and FISTA, *arXiv preprint arXiv:2211.01610*, 2022.
- [5] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Springer, New York, 2004.
- [6] Y. Nesterov, How to make the gradients small, *Optimization* 88 (2012), 10–11.
- [7] B. Shi, S. Du, M. Jordan, W. Su, Understanding the acceleration phenomenon via high-resolution differential equations, *Math. Program.* 195 (2022), 79–148.
- [8] A. Taylor, J. Hendrickx, F. Glineur, Exact worst-case convergence rates of the proximal gradient method for composite convex minimization, *J. Optim. Theory Appl.* 178 (2018), 455–476.

- [9] M. Teboulle, Y. Vaisbourd, An elementary approach to tight worst case complexity analysis of gradient based methods, *Math. Program.* 201 (2022), 63–96.
- [10] L. Zhang, J. Wang, H. Zhang, New insights in smoothness and strong convexity with improved convergence of gradient descent, *arXiv preprint arXiv:2110.15470*, 2021.
- [11] X. Zhang, H. Zhang, A new exact worst-case linear convergence rate of the proximal gradient method, *arXiv preprint arXiv:1902.09181*, 2019.

Manuscript received 2 June 2023
revised 26 September 2023
accepted for publication 29 October 2023

BEIER CHEN

Department of Mathematics, National University of Defense Technology
Changsha, Hunan 410073, China

HUI ZHANG

Department of Mathematics, National University of Defense Technology
Changsha, Hunan 410073, China
E-mail address: h.zhang1984@163.com