SIAM J. OPTIM.
Vol. 29, No. 1, pp. 162–174

CONVERGENCE RATES OF PROXIMAL GRADIENT METHODS VIA THE CONVEX CONJUGATE*

DAVID H. GUTMAN† AND JAVIER F. PEÑA‡

Abstract. We give a novel proof of the $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ convergence rates of the proximal gradient and accelerated proximal gradient methods for composite convex minimization. The crux of the new proof is an upper bound constructed via the convex conjugate of the objective function.

Key words. convex conjugate, proximal gradient, acceleration

AMS subject classifications. 90C25, 90C46, 90C52

DOI. 10.1137/18M1164329

1. Introduction.

The development of accelerated versions of first-order methods has had a profound influence in convex optimization. In his seminal paper [9] Nesterov devised a first-order algorithm with optimal $\mathcal{O}(1/k^2)$ rate of convergence for unconstrained convex optimization via a modification of the standard gradient descent algorithm that includes *momentum* steps. A later breakthrough was the acceleration of the *proximal gradient method* independently developed by Beck and Teboulle [2] and by Nesterov [11]. The proximal gradient method, also known as the forward-backward splitting method [8], is an extension of the gradient descent method to solve the composite minimization problem

$$\min_{x\in\mathbb{R}^n} \varphi(x) + \psi(x), \tag{1}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is differentiable on $\mathrm{dom}(\varphi)$ and $\psi : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is a closed convex function such that $\mathrm{dom}(\psi) \subseteq \mathrm{dom}(\varphi)$ and such that for $t > 0$ the proximal map $\mathsf{Prox}_t : \mathbb{R}^n \to \mathrm{dom}(\psi)$ defined by

$$\mathsf{Prox}_t(x) := \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \left\{ \psi(y) + \frac{1}{2t}\|y - x\|^2 \right\} \tag{2}$$

is computable.

The enormous significance of Nesterov's and Beck and Teboulle's breakthroughs has prompted interest in new approaches to explaining how acceleration is achieved in first-order methods [1, 3, 4, 5, 7, 12, 13]. Some of these approaches are based on geometric [3, 4], control [7], and differential equations [13] techniques. The recent article [12] relies on the convex conjugate to give a unified and succinct derivation of the $\mathcal{O}(1/\sqrt{k}), \mathcal{O}(1/k)$, and $\mathcal{O}(1/k^2)$ convergence rates of the subgradient, gradient, and accelerated gradient methods for unconstrained smooth convex minimization. The crux of the approach in [12] is a generic upper bound on the iterates generated by the subgradient, gradient, and accelerated gradient algorithms constructed via the

*Received by the editors January 9, 2018; accepted for publication (in revised form) November 2, 2018; published electronically January 17, 2019.
http://www.siam.org/journals/siopt/29-1/M116432.html
Funding: This research was funded by NSF grant CMMI-1534850.
†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (dgutman@andrew.cmu.edu).
‡Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213 (jfp@andrew.cmu.edu).

convex conjugate of the objective function. A natural open question posed in [12] is whether a similar convex conjugate approach can be developed for the broader class of proximal gradient methods. The main contribution of this paper is an affirmative answer to this question.

We extend the main construction in [12] to give a unified derivation of the convergence rates of the proximal gradient, accelerated proximal gradient, and proximal subgradient algorithms for the composite convex minimization problem (1). As in [12], the central results of this paper (Theorems 1, 2, and 3) are upper bounds on the iterates generated by the nonaccelerated proximal gradient, accelerated proximal gradient, and proximal subgradient methods. The expressions in the three upper bounds (see (8), (11), and (16)) as well as their proofs (see section 4) are strikingly similar. They highlight the commonalities and differences of the three methods. The upper bounds are constructed via the convex conjugate of the objective function. Theorems 1 and 2 readily yield the widely known $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ convergence rates of the proximal gradient and accelerated proximal gradient algorithms for (1) when the smooth component φ has Lipschitz gradient and the step sizes are chosen judiciously. The convex conjugate approach underlying Theorems 1 and 2 also extends to a *proximal subgradient algorithm* when the component φ is merely convex but not necessarily smooth. (See Algorithm 2 and Theorem 3.) This extension automatically yields a novel derivation of both classical [10, Theorem 3.2.2] as well as modern convergence rates [6, Theorem 5] for the projected subgradient algorithm.

We should note that in contrast to the classical proofs of the iconic convergence rates $\mathcal{O}(1/k)$ for proximal gradient, $\mathcal{O}(1/k^2)$ for accelerated proximal gradient, and $\mathcal{O}(1/\sqrt{k})$ for projected subgradient algorithms, our central results, namely, Theorems 1, 2, and 3, require substantially weaker assumptions. More precisely, Theorems 1, 2, and 3 hold under suitable assumptions on the step sizes and momentum steps but do not require any Lipschitz condition on the components of the objective function or on their gradients. As a consequence, for the proximal gradient method Theorem 1 guarantees convergence of the iterates' objective values to optimality in the absence of Lipschitz continuity provided the step sizes are not summable. Similarly, for the accelerated proximal gradient Theorem 2 guarantees the same type of convergence under an even milder boundedness condition. Finally, Theorem 3 yields convergence results of similar flavor for the projected subgradient method provided the subgradient oracle satisfies a fairly mild and general steepness condition.

Throughout the paper we assume that $\mathbb{R}^n$ is endowed with an inner product $\langle \cdot, \cdot \rangle$ and that $\| \cdot \|$ denotes the corresponding Euclidean norm.

2. Proximal gradient and accelerated proximal gradient methods. Let $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function such that the proximal map (2) is computable, and let $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a differentiable convex function such that $\mathrm{dom}(\psi) \subseteq \mathrm{dom}(\varphi)$. Let $f := \varphi + \psi$, and consider the problem (1), which can be rewritten as

$$\min_{x \in \mathbb{R}^n} f(x). \tag{3}$$

Algorithm 1 describes a template of a proximal gradient algorithm for (3).

Step 7 of Algorithm 1 incorporates a momentum step. The nonaccelerated proximal gradient method is obtained by choosing $\theta_{k+1} = 1$ in Step 6. In this case Step 7 simply sets $y_{k+1} = x_{k+1}$ and does not incorporate any momentum. Other choices of $\theta_{k+1} \in (0, 1]$ yield accelerated versions of the proximal gradient method. In particular, the FISTA algorithm in [2] is obtained by choosing $\theta_{k+1} \in (0, 1]$ via the rule

Algorithm 1 Template for proximal gradient method

1: **input:** $x_0 \in \mathrm{dom}(\varphi)$
2: $y_0 := x_0$; $\theta_0 := 1$
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: pick $t_k > 0$
5: $x_{k+1} := \mathsf{Prox}_{t_k}(y_k - t_k \nabla\varphi(y_k))$
6: pick $\theta_{k+1} \in (0, 1]$
7: $y_{k+1} := x_{k+1} + \frac{\theta_{k+1}(1-\theta_k)}{\theta_k}(x_{k+1} - x_k)$
8: **end for**

$\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$. In this case $\theta_k \in (0, 1)$ for $k \geq 1$ and there is a nontrivial momentum term in Step 7. Algorithm 1 implicitly assumes that the choice of θ_{k+1} in Step 6 is so that the point y_{k+1} in Step 7 satisfies $y_{k+1} \in \mathrm{dom}(\varphi)$. This holds provided $\mathrm{dom}(\varphi)$ is sufficiently larger than $\mathrm{dom}(\psi)$.

The main results in this paper are Theorem 1 and its variant Theorem 2 below which subsume the widely known convergence rates $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ of the proximal gradient and accelerated proximal gradient algorithms under suitable choices of $t_k, \theta_k,\ k = 0, 1, \ldots$.

Theorem 1 relies on a suitably constructed sequence $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$. The construction of $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$, in turn is motivated by the identity (5) below.

Consider Step 5 in Algorithm 1, namely

$$x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla\varphi(y_k)) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ \psi(x) + \frac{1}{2t_k}\|x - (y_k - t_k \nabla\varphi(y_k))\|^2 \right\}. \tag{4}$$

The optimality conditions for (4) can be written as

$$g_k^\psi + \frac{1}{t_k}(x_{k+1} - (y_k - t_k \nabla\varphi(y_k))) = 0$$

for some $g_k^\psi \in \partial\psi(x_{k+1})$. These conditions imply that

$$x_{k+1} = y_k - t_k \cdot g_k,$$

where $g_k := g_k^\varphi + g_k^\psi$ for $g_k^\varphi := \nabla\varphi(y_k)$ and for some $g_k^\psi \in \partial\psi(x_{k+1})$. Thus Steps 5 and 7 of Algorithm 1 imply that for $k = 0, 1, \ldots$

$$\frac{y_{k+1} - (1 - \theta_{k+1})x_{k+1}}{\theta_{k+1}} = \frac{x_{k+1} - (1 - \theta_k)x_k}{\theta_k} = \frac{y_k - (1 - \theta_k)x_k}{\theta_k} - \frac{t_k}{\theta_k} g_k.$$

Since $\theta_0 = 1$ and $y_0 = x_0$, it follows that for $k = 1, 2, \ldots$

$$\frac{y_k - (1 - \theta_k)x_k}{\theta_k} = x_0 - \sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i \Leftrightarrow (1 - \theta_k)(y_k - x_k) = \theta_k \left(x_0 - y_k - \sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i \right). \tag{5}$$

As is customary, we will assume that the step sizes t_k chosen at Step 4 in Algorithm 1 satisfy the following decrease condition:

$$\begin{aligned} f(x_{k+1}) &\leq \min_{x \in \mathbb{R}^n} \left\{ \varphi(y_k) + \langle \nabla\varphi(y_k), x - y_k \rangle + \frac{1}{2t_k}\|x - y_k\|^2 + \psi(x) \right\} \\ &= \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^\psi, y_k - x_{k+1} \right\rangle - \frac{t_k}{2}\|g_k\|^2. \end{aligned} \tag{6}$$

The condition (6) holds in particular when $\nabla\varphi$ is Lipschitz and $t_k,\ k = 0, 1, \ldots,$ are chosen via a standard backtracking procedure. Observe that (6) implies $f(x_{k+1}) \leq f(y_k)$.

Theorem 1 also relies on the convex conjugate function. Recall that if $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex function, then its *convex conjugate* $h^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as

$$h^*(z) = \sup_{x \in \mathbb{R}^n} \{\langle z, x\rangle - h(x)\}.$$

THEOREM 1. *Suppose $\theta_k \in (0, 1],\ k = 0, 1, 2, \ldots,$ and the step sizes $t_k > 0,\ k = 0, 1, 2, \ldots,$ are such that (6) holds. Let $x_k \in \mathbb{R}^n,\ k = 1, 2, \ldots,$ be the iterates generated by Algorithm 1. Let $z_k \in \mathbb{R}^n,\ k = 1, 2 \ldots,$ be as follows:*

$$z_k := \frac{\sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i}{\sum_{i=0}^{k-1} \frac{t_i}{\theta_i}}. \tag{7}$$

Then for $k = 1, 2, \ldots$

$$\mathsf{LHS}_k \leq -f^*(z_k) + \langle z_k, x_0\rangle - \frac{\sum_{i=0}^{k-1} \frac{t_i}{\theta_i}}{2} \|z_k\|^2, \tag{8}$$

where LHS_k is as follows depending on the choice of $\theta_k \in (0, 1]$ and $t_k > 0$:

(a) *When $\theta_k = 1,\ k = 0, 1, \ldots,$ let*

$$\mathsf{LHS}_k := \frac{\sum_{i=0}^{k} t_i f(x_{i+1})}{\sum_{i=0}^{k} t_i}.$$

(b) *When $t_k = 1/L,\ k = 0, 1, \ldots,$ for some positive constant L and $\theta_k,\ k = 0, 1, 2, \ldots,$ are chosen via $\theta_0 = 1$ and $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1}),\ k = 0, 1, \ldots,$ let*

$$\mathsf{LHS}_k = f(x_k).$$

Theorem 1 readily implies that in both case (a) and case (b)

$$\begin{aligned}
\mathsf{LHS}_k &\leq \inf_{u \in \mathbb{R}^n} \{f(u) - \langle z_k, u\rangle\} + \min_{u \in \mathbb{R}^n} \left\{ \langle z_k, u\rangle + \frac{1}{2 \cdot \sum_{i=0}^{k-1} \frac{t_i}{\theta_i}} \|u - x_0\|^2 \right\} \\
&\leq \inf_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2 \cdot \sum_{i=0}^{k-1} \frac{t_i}{\theta_i}} \|u - x_0\|^2 \right\} \\
&\leq f(x) + \frac{1}{2 \cdot \sum_{i=0}^{k-1} \frac{t_i}{\theta_i}} \|x - x_0\|^2
\end{aligned}$$

for all $x \in \mathbb{R}^n$.

Let $\bar{f}$ and $\bar{X}$ respectively denote the optimal value and set of optimal solutions to (3). If $\bar{f}$ is finite and $\bar{X}$ is nonempty, then in both case (a) and case (b) of Theorem 1 we get

$$f(x_k) - \bar{f} \leq \frac{\operatorname{dist}(x_0, \bar{X})^2}{2 \cdot \sum_{i=0}^{k-1} \frac{t_i}{\theta_i}}. \tag{9}$$

Suppose $t_k \geq 1/L,\ k = 0, 1, 2, \ldots$, for some constant $L > 0$. This holds in particular for $L := \max\{L_0, L_\varphi/\alpha\}$ if $\nabla\varphi$ is L_φ-Lipschitz and t_k is chosen via the following standard type of backtracking procedure: pick $t_k = 1/L_0$ for some $L_0 > 0$, and scale t_k by $\alpha \in (0, 1)$ until (6) holds for $x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k\nabla\varphi(y_k))$. Then inequality (9) yields the following known convergence bound for the proximal gradient method:

$$f(x_k) - \bar{f} \leq \frac{L \cdot \operatorname{dist}(x_0, \bar{X})^2}{2k}.$$

On the other hand, suppose $t_k = 1/L,\ k = 0, 1, 2, \ldots$, for some constant $L > 0$ and $\theta_k,\ k = 0, 1, 2, \ldots$, are chosen via $\theta_0 = 1$ and $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$. Then a straightforward induction shows that

$$\sum_{i=0}^{k-1} \frac{t_i}{\theta_i} = (1 - \theta_k) \sum_{i=0}^{k} \frac{t_i}{\theta_i} = \frac{1}{L\theta_{k-1}^2}. \tag{10}$$

The conditions $\theta_0 = 1$, $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$ and an additional induction show that $\theta_{k-1}^2 \leq 4/(k+1)^2,\ k = 1, 2, \ldots$. Thus Theorem 1(b), inequality (9), and equation (10) yield the following known convergence bound for the accelerated proximal gradient method:

$$f(x_k) - \bar{f} \leq \frac{2L \cdot \operatorname{dist}(x_0, \bar{X})^2}{(k+1)^2}.$$

Although Theorem 1 yields the iconic $\mathcal{O}(1/k^2)$ convergence rate of the accelerated proximal gradient algorithm, it applies under the somewhat restrictive conditions stated in case (b) above. In particular, case (b) does not cover the more general case when $t_k,\ k = 0, 1, \ldots$, are chosen via backtracking as in the FISTA with backtracking algorithm in [2]. The convergence rate in this case, namely [2, Theorem 4.4], is a consequence of Theorem 2 below. Theorem 2 is a variant of Theorem 1(b) that applies to more flexible choices of $t_k, \theta_k,\ k = 0, 1, \ldots$. In particular, Theorem 2 applies to the popular choice $\theta_k = \frac{2}{k+2},\ k = 0, 1, \ldots$.

THEOREM 2. *Suppose $\bar{f} = \min_{x\in\mathbb{R}^n} f(x)$ is finite, $\theta_k \in (0, 1],\ k = 0, 1, 2, \ldots$, satisfy $\theta_0 = 1$ and $\theta_{k+1}^2 \geq \theta_k^2(1 - \theta_{k+1})$, and the step sizes $t_k > 0,\ k = 0, 1, 2, \ldots$, are nonincreasing and such that (6) holds. Let $x_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$, be the iterates generated by Algorithm 1. Let $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$, be as follows:*

$$z_k = \frac{\theta_{k-1}^2}{t_{k-1}} \cdot \sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i.$$

Then for $k = 1, 2, \ldots$

$$f(x_k) - \bar{f} \leq -(R_k \cdot (f - \bar{f}))^*(z_k) + \langle z_k, x_0\rangle - \frac{t_{k-1}}{2\theta_{k-1}^2}\|z_k\|^2, \tag{11}$$

where $R_1 = 1$ and $R_{k+1} = \frac{t_{k-1}}{t_k} \cdot \frac{\theta_k^2}{\theta_{k-1}^2(1-\theta_k)} \cdot R_k \geq 1,\ k = 1, 2, \ldots$. In particular, if $\bar{X} = \{x \in \mathbb{R}^n : f(x) = \bar{f}\}$ is nonempty, then

$$f(x_k) - \bar{f} \leq \inf_{u\in\mathbb{R}^n} \left\{ R_k \cdot (f(u) - \bar{f}) + \frac{\theta_{k-1}^2}{2t_{k-1}}\|u - x_0\|^2 \right\} \leq \frac{\theta_{k-1}^2 \cdot \operatorname{dist}(x_0, \bar{X})^2}{2t_{k-1}}.$$

Suppose the step sizes $t_k,\ k = 0, 1, 2, \ldots$, are nonincreasing and satisfy (6) and $t_k \geq 1/L,\ k = 0, 1, 2, \ldots$, for some constant $L > 0$. This holds in particular when $\nabla\varphi$ is Lipschitz and t_k is chosen via a suitable backtracking procedure such as the one in [2]. If $\theta_0 = 1$ and $\theta_{k+1}^2 \geq \theta_k^2(1 - \theta_{k+1}),\ k = 0, 1, \ldots$, then Theorem 2 implies that

$$f(x_k) - \bar{f} \leq \frac{L\theta_{k-1}^2 \cdot \operatorname{dist}(x_0, \bar{X})^2}{2}.$$

If $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1}),\ k = 0, 1, \ldots$, or $\theta_k = 2/(k + 2),\ k = 0, 1, \ldots$, then $\theta_{k-1}^2 \leq 4/(k + 1)^2$, and so

$$f(x_k) - \bar{f} \leq \frac{2L \cdot \operatorname{dist}(x_0, \bar{X})^2}{(k + 1)^2}.$$

We conclude this section by noting other immediate and interesting consequences of Theorems 1 and 2. Observe that these two theorems rely only on some assumptions on the step sizes $t_k,\ k = 0, 1, 2, \ldots$, and on the momentum steps $\theta_k,\ k = 0, 1, 2, \ldots$. Unlike classical proofs of convergence for the proximal gradient and accelerated proximal gradient algorithms, Theorems 1 and 2 do not require $\nabla\varphi$ to be Lipschitz continuous. As a consequence, the iterates generated by Algorithm 1 satisfy $f(x_k) \to \bar{f}$ for a broader class of functions. In particular, consider the special case when $\psi = 0$ and $\nabla\varphi$ satisfies the following type of Hölder continuity: there exist constants L and $v \in (0, 1]$ such that for all $x, y \in \mathbb{R}^n$

$$\|\nabla\varphi(y) - \nabla\varphi(x)\| \leq L\|x - y\|^v.$$

In this case $f = \varphi$ and some straightforward calculations show that for all $x, y \in \mathbb{R}^n$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{1 + v}\|y - x\|^{1+v}.$$

Thus a standard backtracking procedure guarantees that the stepsize t_k at each main iteration of Algorithm 1 can be chosen so that (6) holds and

$$t_k \geq C \cdot \|\nabla f(y_k)\|^{\frac{1-v}{v}} \tag{12}$$

for some constant $C > 0$. When $\theta_k = 1,\ k = 1, 2, \ldots$, inequality (6) implies that the sequences $f(x_k),\ k = 0, 1, 2, \ldots$, and $\operatorname{dist}(x_k, \bar{X}),\ k = 0, 1, 2, \ldots$, are nonincreasing. Thus in that case the convexity of f implies that

$$f(x_k) - \bar{f} \leq \|\nabla f(x_k)\| \cdot \operatorname{dist}(x_k, \bar{X}) \leq \|\nabla f(x_k)\| \cdot \operatorname{dist}(x_0, \bar{X}). \tag{13}$$

Combining (12), (13), Theorem 1, and the fact that $f(x_k),\ k = 0, 1, 2, \ldots$, is nonincreasing, it follows that $f(x_k) \to \bar{f}$ and $\nabla f(x_k) \to 0$ when $\theta_k = 1,\ k = 1, 2, \ldots$.

Theorem 1 also implies that the iterates generated by Algorithm 1 satisfy $f(x_k) \to \bar{f}$ when $\theta_k = 1,\ k = 1, 2, \ldots$, provided $\sum_{i=0}^k t_i \to \infty$. Similarly, Theorem 2 implies that the iterates generated by Algorithm 1 satisfy $f(x_k) \to \bar{f}$ when $\theta_{k+1}^2 \geq \theta_k^2(1-\theta_{k+1})$ provided $t_k/\theta_k^2 \to \infty$. We note that the condition $\sum_{i=0}^k t_i \to \infty$ is implied by and therefore weaker than the popular Lipschitz continuity assumption on $\nabla\varphi$. The same is true for the condition $t_k/\theta_k^2 \to \infty$ when $\theta_k = 2/(k + 2),\ k = 0, 1, 2, \ldots$, or $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1}),\ k = 0, 1, 2, \ldots$.

Algorithm 2 Proximal subgradient method

1: **input:** $x_0 \in \mathrm{dom}(\varphi)$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: pick $g_k^\varphi \in \partial\varphi(x_k)$ and $t_k > 0$
4: $x_{k+1} := \mathsf{Prox}_{t_k}(x_k - t_k g_k^\varphi)$
5: **end for**

3. Proximal subgradient method. Algorithm 2 describes a variant of Algorithm 1 for the case when $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is merely convex.

When ψ is the indicator function I_C of a closed convex set $C \subseteq \mathrm{dom}(\varphi)$, Step 4 in Algorithm 2 can be rewritten as $x_{k+1} = \arg\min_{x \in C} \|x_k - t_k \cdot g_k^\varphi - x\| = \Pi_C(x_k - t_k \cdot g_k^\varphi)$. Hence when $\psi = I_C$, Algorithm 2 becomes the projected subgradient method for

$$\min_{x \in C} \varphi(x). \tag{14}$$

The classical convergence rate for the projected subgradient method is an immediate consequence of Theorem 3, as we detail below. Observe that

$$x_{k+1} = \mathsf{Prox}_{t_k}(x_k - t_k g_k^\varphi) \Leftrightarrow x_{k+1} = x_k - t_k \cdot g_k,$$

where $g_k = g_k^\varphi + g_k^\psi$ for some $g_k^\psi \in \partial\psi(x_{k+1})$. Next, let $z_k \in \mathbb{R}^n$, $k = 0, 1, 2 \ldots$, be as follows:

$$z_k = \frac{\sum_{i=0}^k t_i g_i}{\sum_{i=0}^k t_i}. \tag{15}$$

THEOREM 3. *Let $x_k \in \mathbb{R}^n$, $k = 0, 1, 2, \ldots$, be the sequence of iterates generated by Algorithm 2, and let $z_k \in \mathbb{R}^n$, $k = 0, 1, 2 \ldots$, be defined by (15). Then for $k = 0, 1, 2, \ldots$*

$$\frac{\sum_{i=0}^k t_i(\varphi(x_i) + \psi(x_{i+1})) + \frac{1}{2}\sum_{i=0}^k t_i^2(\|g_i^\psi\|^2 - \|g_i^\varphi\|^2)}{\sum_{i=0}^k t_i} \leq -f^*(z_k) + \langle z_k, x_0\rangle - \frac{\sum_{i=0}^k t_i}{2}\|z_k\|^2. \tag{16}$$

Let $C \subseteq \mathbb{R}^n$ be a nonempty closed convex set and $\psi = I_C$. As noted above, in this case Algorithm 2 becomes the projected subgradient algorithm for problem (14). We next show that in this case Theorem 3 yields the classical convergence rates (18) and (19) as well as the modern and more general one (20) recently established by Grimmer [6, Theorem 5].

Suppose $\bar\varphi = \min_{x \in C} \varphi(x)$ is finite and $\bar X := \{x \in C : \varphi(x) = \bar\varphi\}$ is nonempty. From Theorem 3 it follows that

$$\frac{\sum_{i=0}^k t_i\varphi(x_i) + \frac{1}{2}\sum_{i=0}^k t_i^2(\|g_i^\psi\|^2 - \|g_i^\varphi\|^2)}{\sum_{i=0}^k t_i} \leq \inf_{u \in C}\{\varphi(u) - \langle z_k, u\rangle\} + \min_u \left\{\langle z_k, u\rangle + \frac{1}{2\sum_{i=0}^k t_i}\|u - x_0\|^2\right\} \leq \bar\varphi + \frac{\mathrm{dist}(x_0, \bar X)^2}{2\sum_{i=0}^k t_i}.$$

Therefore,

$$\sum_{i=0}^{k} t_i(\varphi(x_i) - \bar{\varphi}) \le \frac{\sum_{i=0}^{k} t_i^2(\|g_i^{\varphi}\|^2 - \|g_i^{\psi}\|^2) + \operatorname{dist}(x_0, \bar{X})^2}{2}. \tag{17}$$

In particular, if $\|g\| \le L$ for all $x \in C$ and $g \in \partial\varphi(x)$, then (17) implies

$$\min_{i=0,\dots,k} (\varphi(x_i) - \bar{\varphi}) \le \frac{\sum_{i=0}^{k} t_i^2 L^2 + \operatorname{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^{k} t_i}. \tag{18}$$

Let $\alpha_i := t_i\|g_i^{\varphi}\|$, $i = 0, 1, \dots$. Then Step 4 in Algorithm 2 can be rewritten as $x_{k+1} = \Pi_C\left(x_k - \alpha_k \cdot g_k^{\varphi}/\|g_k^{\varphi}\|\right)$ provided $\|g_k^{\varphi}\| > 0$, which occurs as long as x_k is not an optimal solution to (14). If $\|g_i^{\varphi}\| > 0$ for $i = 0, 1, \dots, k$ and $\|g\| \le L$ for all $x \in C$ and $g \in \partial\varphi(x)$, then (17) implies

$$\min_{i=0,\dots,k} (\varphi(x_i) - \bar{\varphi}) \le L \cdot \frac{\sum_{i=0}^{k} \alpha_i^2 + \operatorname{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^{k} \alpha_i}. \tag{19}$$

Let $\mathcal{L} : \mathbb{R}_+ \to \mathbb{R}_+$. Following Grimmer [6], the subgradient oracle for φ is $\mathcal{L}$-steep on C if for all $x \in C$ and $g \in \partial\varphi(x)$

$$\|g\| \le \mathcal{L}(\varphi(x) - \bar{\varphi}).$$

As discussed by Grimmer [6], $\mathcal{L}$-steepness is a more general condition than the traditional bound $\|g\| \le L$ for $x \in C$ and $g \in \partial\varphi(x)$ used above. Indeed, the latter bound is precisely $\mathcal{L}$-steepness for the constant function $\mathcal{L}(t) = L$ and holds when φ is L-Lipschitz on C. For another example of $\mathcal{L}$-steepness, consider the case when $C = \mathbb{R}^n$ and φ is differentiable on $\mathbb{R}^n$ and such that $\nabla\varphi$ is L-Lipschitz. In this case it readily follows that

$$\bar{\varphi} \le \min_y \left\{\varphi(x) + \langle\nabla\varphi(x), y - x\rangle + \frac{L}{2}\|y - x\|^2\right\} = \varphi(x) - \frac{1}{2L}\|\nabla\varphi(x)\|^2.$$

Thus the subgradient oracle for φ is $\mathcal{L}$-steep for $\mathcal{L}(t) = \sqrt{2Lt}$. More generally, if $\nabla\varphi$ is Hölder-continuous, that is, if there exist L and $v > 0$ such that for all $x, y \in \mathbb{R}^n$

$$\|\nabla\varphi(y) - \nabla\varphi(x)\| \le L\|x - y\|^v,$$

then

$$\begin{aligned}\bar{\varphi} &\le \min_y \left\{\varphi(x) + \langle\nabla\varphi(x), y - x\rangle + \frac{L}{1+v}\|y - x\|^{1+v}\right\} \\ &= \varphi(x) - \frac{v}{1+v} \cdot \frac{1}{L^{\frac{1}{v}}}\|\nabla\varphi(x)\|^{\frac{1+v}{v}}.\end{aligned}$$

Thus the subgradient oracle for φ is $\mathcal{L}$-steep for $\mathcal{L}(t) = ((1+v)^v L t^v/v^v)^{1/(1+v)}$.

Suppose the subgradient oracle for φ is $\mathcal{L}$-steep for some $\mathcal{L} : \mathbb{R}_+ \to \mathbb{R}_+$. If $\alpha_i := t_i\|g_i^{\varphi}\| > 0$ for $i = 0, 1, \dots, k$, then (17) implies

$$\sum_{i=0}^{k} \alpha_i \cdot \frac{\varphi(x_i) - \bar{\varphi}}{\mathcal{L}(\varphi(x_i) - \bar{\varphi})} \le \frac{\sum_{i=0}^{k} \alpha_i^2 + \operatorname{dist}(x_0, \bar{X})^2}{2},$$

and thus

$$\min_{i=0,\dots,k}(\varphi(x_i)-\bar{\varphi}) \le \sup\left\{t : \frac{t}{\mathcal{L}(t)} \le \frac{\sum_{i=0}^k \alpha_i^2 + \operatorname{dist}(x_0,\bar{X})^2}{2\sum_{i=0}^k \alpha_i}\right\}. \tag{20}$$

For $\alpha_i = a/\sqrt{k+1},\ i=0,\dots,k$, with $a>0$ inequality (20) yields

$$\min_{i=0,\dots,k}(\varphi(x_i)-\bar{\varphi}) \le \sup\left\{t : \frac{t}{\mathcal{L}(t)} \le \frac{1}{2\sqrt{k+1}}\left(a + \frac{\operatorname{dist}(x_0,\bar{X})^2}{a}\right)\right\}. \tag{21}$$

As we discussed above, when φ is L-Lipschitz on C, the subgradient oracle is $\mathcal{L}$-steep for $\mathcal{L}(t)=L$. Hence inequality (21) yields the classical $\mathcal{O}(1/\sqrt{k})$ convergence rate of the projected subgradient method. Furthermore, when $C=\mathbb{R}^n$ and φ is differentiable and $\nabla\varphi$ is L-Lipschitz, the subgradient oracle is $\mathcal{L}$-steep for $\mathcal{L}(t)=\sqrt{2Lt}$. Hence inequality (21) yields

$$\min_{i=0,\dots,k}(\varphi(x_i)-\bar{\varphi}) = \mathcal{O}(1/k), \tag{22}$$

which matches the dependence on k of the classical convergence rate of the gradient method. As noted by Grimmer [6], it is striking that (22) holds for Algorithm 2, which relies only on the availability of a subgradient oracle for φ. However, we should note that the $\mathcal{O}(1/k)$ rate attained by Algorithm 2 depends on the choice $\alpha_i = a/\sqrt{k+1},\ i=0,\dots,k$. In particular, (22) holds for a prescribed number k of iterations, and the constant in the $\mathcal{O}(1/k)$ expression in (22) depends on how closely a approximates $\operatorname{dist}(x_0,\bar{X})$.

4. Proofs of Theorems 1, 2, and 3. We will use the following properties of the convex conjugate.

Suppose $h:\mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is a convex function. Then

$$h^*(z)+h(x) \ge \langle z,x\rangle \tag{23}$$

for all $z,x\in\mathbb{R}^n$, and equality holds if $z\in\partial h(x)$.

Suppose $f,\varphi,\psi:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$ are convex functions and $f=\varphi+\psi$. Then

$$f^*(z^\varphi+z^\psi) \le \varphi^*(z^\varphi)+\psi^*(z^\psi) \ \text{ for all } z^\varphi, z^\psi\in\mathbb{R}^n. \tag{24}$$

Suppose $f:\mathbb{R}^n\to\mathbb{R}_+\cup\{\infty\}$ is a convex function and $R\ge 1$. Then

$$(R\cdot f)^*(Rz) = R\cdot(f^*(z)), \tag{25}$$

and

$$(R\cdot f)^*(z) \le f^*(z). \tag{26}$$

4.1. Proof of Theorem 1. We prove (8) by induction. To ease notation, let $\mu_k := \frac{1}{\sum_{i=0}^{k-1} t_i/\theta_i}$ throughout this proof. For $k=1$ we have

$$\begin{aligned}
\mathsf{LHS}_1 = f(x_1) &\le \varphi(x_0)+\psi(x_1)+\left\langle g_0^\psi, x_0-x_1\right\rangle - \frac{t_0}{2}\|g_0\|^2\\
&= \varphi(x_0)-\langle g_0^\varphi, x_0\rangle + \psi(x_1) - \left\langle g_0^\psi, x_1\right\rangle + \langle g_0, x_0\rangle - \frac{t_0}{2}\|g_0\|^2\\
&= -\varphi^*(g_0^\varphi) - \psi^*(g_0^\psi) + \langle g_0, x_0\rangle - \frac{t_0}{2}\|g_0\|^2\\
&\le -f^*(z_1) + \langle z_1, x_0\rangle - \frac{\|z_1\|^2}{2\mu_1}.
\end{aligned}$$

The first step follows from (6). The third step follows from (23) and $g_0^\varphi = \nabla\varphi(x_0)$, $g_0^\psi \in \partial\psi(x_1)$. The last step follows from (24), the choice of $z_1 = g_0 = g_0^\varphi + g_0^\psi$, and $\mu_1 = 1/t_0$.

Suppose (8) holds for k, and let $\gamma_k = \frac{t_k/\theta_k}{\sum_{i=0}^k t_i/\theta_i}$. The construction (7) implies that

$$\begin{aligned} z_{k+1} &= (1-\gamma_k)z_k + \gamma_k g_k, \\ \mu_{k+1} &= (1-\gamma_k)\mu_k. \end{aligned}$$

Therefore,

$$\begin{aligned} \langle z_{k+1}, x_0\rangle - \frac{\|z_{k+1}\|^2}{2\mu_{k+1}} &= (1-\gamma_k)\left(\langle z_k, x_0\rangle - \frac{\|z_k\|^2}{2\mu_k}\right) \\ &\quad + \gamma_k\left(\left\langle g_k, x_0 - \frac{z_k}{\mu_k}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2\right). \end{aligned} \tag{27}$$

In addition, the convexity of f^*, properties (23), (24), and $g_k^\varphi = \nabla\varphi(y_k)$, $g_k^\psi \in \partial\psi(x_{k+1})$, $g_k = g_k^\varphi + g_k^\psi$ imply

(28)

$$\begin{aligned} -f^*(z_{k+1}) &\geq -(1-\gamma_k)f^*(z_k) - \gamma_k f^*(g_k) \\ &\geq -(1-\gamma_k)f^*(z_k) - \gamma_k(\varphi^*(g_k^\varphi) + \psi^*(g_k^\psi)) \\ &= -(1-\gamma_k)f^*(z_k) - \gamma_k\left(\langle g_k^\varphi, y_k\rangle - \varphi(y_k) + \left\langle g_k^\psi, x_{k+1}\right\rangle - \psi(x_{k+1})\right). \end{aligned}$$

Let RHS_k denote the right-hand side in (8). From (27) and (28) it follows that

$$\mathsf{RHS}_{k+1} - (1-\gamma_k)\mathsf{RHS}_k \geq \gamma_k \cdot D_k,$$

where

$$D_k := \left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k}\right\rangle + \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^\psi, y_k - x_{k+1}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2.$$

Hence to complete the proof of (8) by induction it suffices to show that

$$\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k \leq \gamma_k \cdot D_k. \tag{29}$$

To that end, we consider case (a) and case (b) separately.

Case (a). In this case $\gamma_k = \frac{t_k}{\sum_{i=0}^k t_i}$ and $y_k = x_k$. Thus $\mu_k = \frac{1}{\sum_{i=0}^{k-1} t_i}$, $\frac{\gamma_k}{(1-\gamma_k)\mu_k} = t_k$, and $x_0 - y_k - \frac{z_k}{\mu_k} = 0$. Therefore,

$$\begin{aligned} &\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k \\ &= \gamma_k \cdot f(x_{k+1}) \\ &\leq \gamma_k\left(\varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^\psi, y_k - x_{k+1}\right\rangle - \frac{t_k}{2}\|g_k\|^2\right) \\ &= \gamma_k\left(\varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^\psi, y_k - x_{k+1}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2\right) \\ &= \gamma_k \cdot D_k. \end{aligned}$$

The second step follows from (6). The third and fourth steps follow from $\frac{\gamma_k}{(1-\gamma_k)\mu_k} = t_k$ and $x_0 - y_k - \frac{z_k}{\mu_k} = 0$, respectively. Thus (29) holds in case (a).

Case (b). In this case (10) yields $\gamma_k = \theta_k$ and $\frac{\gamma_k^2}{(1-\gamma_k)\mu_k} = t_k$. Therefore,

$$\begin{aligned}
&\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k \\
&= f(x_{k+1}) - (1-\gamma_k)(\varphi(x_k) + \psi(x_k)) \\
&\le \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^{\psi}, y_k - x_{k+1} \right\rangle - \frac{t_k}{2}\|g_k\|^2 \\
&\quad - (1-\gamma_k)\left(\varphi(y_k) + \langle g_k^{\varphi}, x_k - y_k\rangle + \psi(x_{k+1}) + \left\langle g_k^{\psi}, x_k - x_{k+1} \right\rangle\right) \\
&= \gamma_k\left(\varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^{\psi}, y_k - x_{k+1} \right\rangle\right) + (1-\gamma_k)\langle g_k, y_k - x_k\rangle - \frac{t_k}{2}\|g_k\|^2 \\
&= \gamma_k \cdot D_k.
\end{aligned}$$

The second step follows from (6) and the convexity of φ and ψ. The last step follows from $\theta_k = \gamma_k$, (5), and $\frac{\gamma_k^2}{(1-\gamma_k)\mu_k} = t_k$. Thus (29) holds in case (b) as well.

4.2. Proof of Theorem 2. The proof of Theorem 2 is a modification of the proof of Theorem 1. Without loss of generality assume $\bar{f} = 0$, as otherwise we can work with $f - \bar{f}$ in place of f. Again we prove (11) by induction. To ease notation, let $\mu_k := \theta_{k-1}^2/t_{k-1}$ throughout this proof. For $k = 1$ inequality (11) is identical to (8) since $R_1 = 1$ and $\theta_0 = 1$. Hence this case follows from the proof of Theorem 1 for $k = 1$. Suppose (11) holds for k. Observe that

$$\begin{aligned}
z_{k+1} &= \rho_k(1-\theta_k)z_k + \theta_k g_k, \\
\mu_{k+1} &= \rho_k(1-\theta_k)\mu_k
\end{aligned}$$

for $\rho_k := \frac{R_{k+1}}{R_k} = \frac{t_{k-1}}{t_k} \cdot \frac{\theta_k^2}{\theta_{k-1}^2(1-\theta_k)} = \frac{\mu_{k+1}}{\mu_k(1-\theta_k)} \ge 1$. Next, proceed as in the proof of Theorem 1. First,

$$\begin{aligned}
&\langle z_{k+1}, x_0\rangle - \frac{\|z_{k+1}\|^2}{2\mu_{k+1}} \\
&= \rho_k(1-\theta_k)\left(\langle z_k, x_0\rangle - \frac{\|z_k\|^2}{2\mu_k}\right) + \theta_k \cdot \left\langle g_k, x_0 - \frac{z_k}{\mu_k} \right\rangle - \frac{\theta_k^2}{2\mu_{k+1}}\|g_k\|^2 \\
&= \rho_k(1-\theta_k)\left(\langle z_k, x_0\rangle - \frac{\|z_k\|^2}{2\mu_k}\right) + \theta_k \cdot \left\langle g_k, x_0 - \frac{z_k}{\mu_k} \right\rangle - \frac{t_k}{2}\|g_k\|^2.
\end{aligned} \tag{30}$$

Second, the convexity of f^* and the fact that $f \ge \bar{f} = 0$ imply

$$\begin{aligned}
-(R_{k+1}\cdot f)^*(z_{k+1}) &\ge -(1-\theta_k)(R_{k+1}\cdot f)^*(\rho_k \cdot z_k) - \theta_k(R_{k+1}\cdot f)^*(g_k) \\
&\ge -(1-\theta_k)(\rho_k \cdot R_k \cdot f)^*(\rho_k \cdot z_k) - \theta_k \cdot f^*(g_k) \\
&\ge -\rho_k(1-\theta_k)(R_k \cdot f)^*(z_k) - \theta_k(\varphi^*(g_k^{\varphi}) + \psi^*(g_k^{\psi})) \\
&= -\rho_k(1-\theta_k)(R_k \cdot f)^*(z_k) \\
&\quad - \theta_k\left(\langle g_k^{\varphi}, y_k\rangle - \varphi(y_k) + \left\langle g_k^{\psi}, x_{k+1} \right\rangle - \psi(x_{k+1})\right).
\end{aligned} \tag{31}$$

The first step follows from the convexity of f^*. The second step follows from (26). The third step follows from (24) and (25). The last step follows from (23) and $g_k^{\varphi} = \nabla\varphi(y_k)$, $g_k^{\psi} \in \partial\psi(x_{k+1})$.

Let RHS_k denote the right-hand side in (11). The induction hypothesis implies that $\mathsf{RHS}_k \geq f(x_k) \geq 0$. Thus from (30), (31), and $\rho_k \geq 1$ it follows that

(32)

$$\begin{aligned}
&\mathsf{RHS}_{k+1} - (1-\theta_k)\mathsf{RHS}_k \\
&\geq \mathsf{RHS}_{k+1} - \rho_k(1-\theta_k)\mathsf{RHS}_k \\
&\geq \theta_k \left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k} \right\rangle + \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^{\psi}, y_k - x_{k+1} \right\rangle \right) - \frac{t_k}{2}\|g_k\|^2.
\end{aligned}$$

Finally, proceeding exactly as in case (b) in the proof of Theorem 1, we get

$$\begin{aligned}
&f(x_{k+1}) - (1-\theta_k)f(x_k) \\
&\leq \theta_k \left(\varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^{\psi}, y_k - x_{k+1} \right\rangle \right) + (1-\theta_k)\left\langle g_k, y_k - x_k \right\rangle - \frac{t_k}{2}\|g_k\|^2 \\
&= \theta_k \left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k} \right\rangle + \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^{\psi}, y_k - x_{k+1} \right\rangle \right) - \frac{t_k}{2}\|g_k\|^2 \\
&\leq \mathsf{RHS}_{k+1} - (1-\theta_k)\mathsf{RHS}_k.
\end{aligned}$$

The second step follows from (5). The third step follows from (32). This completes the proof by induction.

4.3. Proof of Theorem 3. Let LHS_k and RHS_k denote, respectively, the left-hand and right-hand sides in (16). We proceed by induction. For $k = 0$ we have

$$\begin{aligned}
\mathsf{LHS}_0 &= \varphi(x_0) + \psi(x_1) + \frac{t_0(\|g_0^{\psi}\|^2 - \|g_0^{\varphi}\|^2)}{2} \\
&= -\varphi^*(g_0^{\varphi}) + \langle g_0^{\varphi}, x_0 \rangle - \psi^*(g_0^{\psi}) + \left\langle g_0^{\psi}, x_1 \right\rangle + \frac{t_0(\|g_0^{\psi}\|^2 - \|g_0^{\varphi}\|^2)}{2} \\
&\leq -f^*(g_0) + \langle g_0, x_0 \rangle - \frac{t_0\|g_0\|^2}{2} \\
&= \mathsf{RHS}_0.
\end{aligned}$$

The second step follows from (23) and $g_0^{\varphi} \in \partial\varphi(x_0),\ g_0^{\psi} \in \partial\psi(x_1)$. The third step follows from (24) and $g_0 = g_0^{\varphi} + g_0^{\psi},\ x_1 = x_0 - t_0 \cdot g_0$.

Next we show the main inductive step k to $k+1$. Observe that $z_{k+1} = (1-\gamma_k)z_k + \gamma_k g_{k+1}$ for $k = 0, 1, \ldots$, where $\gamma_k = \frac{t_{k+1}}{\sum_{i=0}^{k+1} t_i} \in (0,1)$. Proceeding exactly as in the proof of Theorem 1, we get

$$\begin{aligned}
&\mathsf{RHS}_{k+1} - (1-\gamma_k)\mathsf{RHS}_k \\
&\qquad \geq \gamma_k \left(\varphi(x_{k+1}) + \psi(x_{k+2}) + \left\langle g_{k+1}^{\psi}, x_{k+1} - x_{k+2} \right\rangle - \frac{t_{k+1}\|g_{k+1}\|^2}{2} \right) \\
&\qquad = \gamma_k \left(\varphi(x_{k+1}) + \psi(x_{k+2}) + \frac{t_{k+1}\|g_{k+1}^{\psi}\|^2}{2} - \frac{t_{k+1}\|g_{k+1}^{\varphi}\|^2}{2} \right).
\end{aligned}$$

The second step follows because $g_{k+1} = g_{k+1}^{\varphi} + g_{k+1}^{\psi}$ and $x_{k+2} = x_{k+1} - t_{k+1} \cdot g_{k+1}$. The proof is thus completed by observing that

$$\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k = \gamma_k \left(\varphi(x_{k+1}) + \psi(x_{k+2}) + \frac{t_{k+1}\|g_{k+1}^{\psi}\|^2}{2} - \frac{t_{k+1}\|g_{k+1}^{\varphi}\|^2}{2} \right).$$

REFERENCES

[1] Z. Allen-Zhu and L. Orecchia, *Linear coupling: An ultimate unification of gradient and mirror descent*, in Innovations in Theoretical Computer Science Conference, Schloss Dagstuhl, Dagstuhl, Germany, 2017, 3.

[2] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542.

[3] S. Bubeck, Y. Lee, and M. Singh, *A Geometric Alternative to Nesterov's Accelerated Gradient Descent*, preprint, https://arxiv.org/abs/1506.08187, 2015.

[4] D. Drusvyatskiy, M. Fazel, and S. Roy, *An optimal first order method based on optimal quadratic averaging*, SIAM J. Optim., 28 (2018), pp. 251–271, https://doi.org/10.1137/16M1072528.

[5] N. Flammarion and F. Bach, *From averaging to acceleration, there is only a step-size*, in Proceedings of the 28th Conference on Learning Theory (COLT), Paris, France, 2015, pp. 658–695.

[6] B. Grimmer, *Convergence Rates for Deterministic and Stochastic Subgradient Methods without Lipschitz Continuity*, preprint, https://arxiv.org/abs/1712.04104, 2017.

[7] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM J. Optim., 26 (2016), pp. 57–95, https://doi.org/10.1137/15M1009597.

[8] P. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979, https://doi.org/10.1137/0716071.

[9] Y. Nesterov, *A method for solving the convex programming problem with convergence rate* $\mathcal{O}(1/k^2)$, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547 (in Russian).

[10] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic Publishers, Boston, MA, 2004.

[11] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.

[12] J. Peña, *Convergence of first-order methods via the convex conjugate*, Oper. Res. Lett., 45 (2017), pp. 561–564.

[13] W. Su, S. Boyd, and E. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 2510–2518.