Convergence rates of proximal gradient methods via the convex conjugate

David H. Gutman* Javier F. Peña†

January 8, 2018

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.

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 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 $\varphi : \mathbb{R}^n \to \mathbb{R}$ is differentiable and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a closed convex function such that for $t > 0$ the proximal map

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

is computable.

The significance of Nesterov's and Beck and Teboulle's breakthroughs has prompted interest in new approaches to explain how acceleration is achieved in first-order methods [1,3–5,7,12,13]. Some of these approaches are based on geometric [3,4], control [7], and differential

*Department of Mathematical Sciences, Carnegie Mellon University, USA, dgutman@andrew.cmu.edu
†Tepper School of Business, Carnegie Mellon University, USA, jfp@andrew.cmu.edu

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 convex conjugate of the objective function.

We extend the main construction in [12] to give a unified derivation of the convergence rates of the proximal gradient and accelerated proximal gradient algorithms for the composite convex minimization problem (1). As in [12], the central result of this paper (Theorem 1) is an upper bound on the iterates generated by both the non-accelerated and the accelerated proximal gradient methods. This bound is constructed via the convex conjugate of the objective function. Theorem 1 readily yields 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. Theorem 1 highlights some key similarities and differences between the non-accelerated and the accelerated algorithms. It is noteworthy that Theorem 1 and its variant, Theorem 2, hold under certain conditions on the step sizes and momentum used in the algorithm but do not require any Lipschitz assumption. The convex conjugate approach underlying Theorem 1 also extends to a *proximal subgradient algorithm* when the component φ is merely convex but not necessarily smooth. (See Algorithm 2 and Proposition 1.) 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. The latter derivations are similar to the derivation of the convergence rates for the proximal gradient and accelerated proximal gradient algorithms.

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 $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function such that the proximal map (2) is computable. Let $f := \varphi + \psi$ and consider the problem (1) that 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 (non-accelerated) 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 $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$. In this case $\theta_k \in (0, 1)$ for $k \geq 1$ and there is a non-trivial momentum term in Step 7.

Algorithm 1 Template for proximal gradient method

1: **input:** $x_0 \in \mathbb{R}^n$
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**

The main result in this paper is Theorem 1 below which subsumes 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 suitable 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

$$x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla\varphi(y_k)). \tag{4}$$

The optimality conditions for (4) 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 $g_k^{\psi} \in \partial\psi(x_{k+1})$.

Step 5 and Step 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 it 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 $\theta_k \in (0,1],\ k = 0,1,2,\dots$ and the step sizes $t_k > 0,\ k = 0,1,2,\dots$ are such that (6) holds. Let $x_k \in \mathbb{R}^n,\ k = 1,2,\dots$ be the iterates generated by Algorithm 1. Let $z_k \in \mathbb{R}^n,\ k = 1,2\dots$ 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

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

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

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

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

(b) When $t_k > 0$ and $\theta_k \in (0,1],\ k = 0,1,\dots$ are such that $\sum_{i=0}^{k-1} \frac{t_i}{\theta_i} = (1-\theta_k) \sum_{i=0}^{k} \frac{t_i}{\theta_i}$ let

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

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

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

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 $t_k \geq \frac{1}{L},\ k = 0,1,2,\dots$ for some constant $L > 0$. This holds in particular if $\nabla\varphi$ is Lipschitz and t_k is chosen via a standard backtracking procedure. 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 $t_k = \frac{1}{L},\ k = 0, 1, 2, \ldots$ 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

$$\sum_{i=0}^{k-1} \frac{t_i}{\theta_i} = (1 - \theta_k) \sum_{i=0}^{k} \frac{t_i}{\theta_i} = \frac{1}{L\theta_{k-1}^2} \geq \frac{(k+1)^2}{4L}.$$

Thus case (b) in Theorem 1 applies and inequality (9) yields 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 non-increasing 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 $k = 1, 2, \ldots$

$$f(x_k) - \bar{f} \leq -(R_k \cdot (f - \bar{f}))^*(z_k) + \langle z_k, x_0 \rangle - \frac{t_{k-1}}{2\theta_{k-1}^2} \|z_k\|^2, \tag{10}$$

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

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

Suppose the step sizes $t_k,\ k = 0, 1, 2, \ldots$ are non-increasing, satisfy (6), and $t_k \geq \frac{1}{L},\ k = 0, 1, 2, \ldots$ for some constant $L > 0$. This holds in particular when $\nabla\varphi$ is Lipschitz and t_k is chosen via a suitable backtracking procedure 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

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

Hence if $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$ or $\theta_k = \frac{2}{k+2}$ for $k = 0, 1, \ldots$ then

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

3 Proof of Theorem 1 and Theorem 2

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

$$h^*(z) + h(x) \geq \langle z, x \rangle \tag{11}$$

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

$$f^*(z^\varphi + z^\psi) \leq \varphi^*(z^\varphi) + \psi^*(z^\psi) \text{ for all } z^\varphi, z^\psi \in \mathbb{R}^n. \tag{12}$$

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

$$(R \cdot f)^*(Rz) = R \cdot (f^*(z)), \tag{13}$$

and

$$(R \cdot f)^*(z) \leq f^*(z). \tag{14}$$

3.1 Proof of Theorem 1

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

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

The first step follows from (6). The third step follows from (11) and $g_0^\varphi = \nabla\varphi(x_0),\ g_0^\psi \in \partial\psi(x_1)$. The last step follows from (12) and the choice of $z_1 = g_0 = g_0^\varphi + g_0^\psi$ and $\mu_1 = \frac{1}{t_0}$.

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,

$$\langle z_{k+1}, x_0 \rangle - \frac{\|z_{k+1}\|^2}{2\mu_{k+1}} = (1-\gamma_k)\left(\langle z_k, x_0 \rangle - \frac{\|z_k\|^2}{2\mu_k}\right) + \gamma_k \left(\left\langle g_k, x_0 - \frac{z_k}{\mu_k} \right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2\right). \tag{15}$$

In addition, the convexity of f^*, properties (11), (12), 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

$$\begin{aligned} -f^*(z_{k+1}) &\geq -(1-\gamma_k)f^*(z_k) - \gamma_k f^*(g_k) \\ &\geq -(1-\gamma_k)f^*(z_k) - \gamma_k(\varphi^*(g_k^\varphi) + \psi^*(g_k^\psi)) \\ &= -(1-\gamma_k)f^*(z_k) - \gamma_k\left(\langle g_k^\varphi, y_k\rangle - \varphi(y_k) + \left\langle g_k^\psi, x_{k+1}\right\rangle - \psi(x_{k+1})\right). \end{aligned} \tag{16}$$

Let RHS_k denote the right-hand side in (8). From (15) and (16) it follows that

$$\begin{aligned} &\mathsf{RHS}_{k+1} - (1-\gamma_k)\mathsf{RHS}_k \\ &\quad\geq \gamma_k\left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k}\right\rangle + \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^\psi, y_k - x_{k+1}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2\right). \end{aligned} \tag{17}$$

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

$$\begin{aligned} &\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k \\ &\quad\leq \gamma_k\left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k}\right\rangle + \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^\psi, y_k - x_{k+1}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2\right). \end{aligned} \tag{18}$$

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

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

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 (18) holds in case (a).

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

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

The second step follows from (6) and the convexity of φ and ψ. The last step follows from $\theta_k = \gamma_k$, equation (5), and $\frac{\gamma_k^2}{(1-\gamma_k)\mu_k} = t_k$. Thus (18) holds in case (b) as well.

3.2 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 (10) by induction. To ease notation, let $\mu_k := \frac{\theta_{k-1}^2}{t_{k-1}}$ throughout this proof. For $k = 1$ inequality (10) 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 (10) 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 $\rho_k := \frac{R_{k+1}}{R_k} = \frac{t_{k-1}}{t_k} \cdot \frac{\theta_k^2}{\theta_{k-1}^2(1-\theta_k)} = \frac{\mu_{k+1}}{\mu_k(1-\theta_k)} \geq 1$. Next, proceed as in the proof of Theorem 1. First,

$$\begin{aligned} \langle z_{k+1}, x_0\rangle - \frac{\|z_{k+1}\|^2}{2\mu_{k+1}} &= \rho_k(1-\theta_k)\left(\langle z_k, x_0\rangle - \frac{\|z_k\|^2}{2\mu_k}\right) + \theta_k \cdot \left\langle g_k, x_0 - \frac{z_k}{\mu_k}\right\rangle - \frac{\theta_k^2}{2\mu_{k+1}}\|g_k\|^2 \\ &= \rho_k(1-\theta_k)\left(\langle z_k, x_0\rangle - \frac{\|z_k\|^2}{2\mu_k}\right) + \theta_k \cdot \left\langle g_k, x_0 - \frac{z_k}{\mu_k}\right\rangle - \frac{t_k}{2}\|g_k\|^2. \quad (19) \end{aligned}$$

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

$$\begin{aligned} -(R_{k+1}\cdot f)^*(z_{k+1}) &\geq -(1-\theta_k)(R_{k+1}\cdot f)^*(\rho_k \cdot z_k) - \theta_k(R_{k+1}\cdot f)^*(g_k) \\ &\geq -(1-\theta_k)(\rho_k \cdot R_k \cdot f)^*(\rho_k \cdot z_k) - \theta_k \cdot f^*(g_k) \quad (20) \\ &\geq -\rho_k(1-\theta_k)(R_k \cdot f)^*(z_k) - \theta_k(\varphi^*(g_k^\varphi) + \psi^*(g_k^\psi)) \\ &= -\rho_k(1-\theta_k)(R_k \cdot f)^*(z_k) - \theta_k\left(\langle g_k^\varphi, y_k\rangle - \varphi(y_k) + \left\langle g_k^\psi, x_{k+1}\right\rangle - \psi(x_{k+1})\right). \end{aligned}$$

The first step follows from the convexity of f^*. The second step follows from (14). The third step follows from (12) and (13). The last step follows from (11) 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 (10). The induction hypothesis implies that $\mathsf{RHS}_k \geq f(x_k) \geq 0$. Thus from (19), (20), and $\rho_k \geq 1$ it follows that

$$\begin{aligned} &\mathsf{RHS}_{k+1} - (1-\theta_k)\mathsf{RHS}_k \\ &\quad \geq \mathsf{RHS}_{k+1} - \rho_k(1-\theta_k)\mathsf{RHS}_k \quad (21) \\ &\quad \geq \theta_k\left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k}\right\rangle + \varphi(y_k) + \psi(x_{k+1}) + \left\langle g_k^\psi, y_k - x_{k+1}\right\rangle\right) - \frac{t_k}{2}\|g_k\|^2. \end{aligned}$$

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 (21). This completes the proof by induction.

4 Proximal subgradient method

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

Algorithm 2 Proximal subgradient method

1: **input:** $x_0 \in \mathbb{R}^n$
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**

When ψ is the indicator function I_C of a closed convex set C, 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

$$\min_{x \in C} \varphi(x). \tag{22}$$

The classical convergence rate for the projected gradient is an immediate consequence of Proposition 1 as we detail below. Proposition 1 in turn is obtained via a minor tweak on the construction and proof of Theorem 1. 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

$$z_k = \frac{\sum_{i=0}^k t_i g_i}{\sum_{i=0}^k t_i}. \tag{23}$$

Proposition 1. *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 (23). Then for $k = 0, 1, 2, \ldots$*

$$\begin{aligned} \frac{\sum_{i=0}^k t_i (\varphi(x_i) + \psi(x_{i+1})) - \frac{1}{2} \sum_{i=0}^k t_i^2 \|g_i^{\varphi}\|^2}{\sum_{i=0}^k t_i} &\leq -f^*(z_k) + \langle z_k, x_0 \rangle - \frac{\sum_{i=0}^k t_i}{2} \|z_k\|^2 \qquad (24) \\ &\leq \min_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2 \sum_{i=0}^k t_i} \|u - x_0\|^2 \right\}. \end{aligned}$$

In particular,

$$\frac{\sum_{i=0}^k t_i(\varphi(x_i)+\psi(x_{i+1})) - \frac{1}{2}\sum_{i=0}^k t_i^2\|g_i^\varphi\|^2}{\sum_{i=0}^k t_i} \le f(x) + \frac{\|x_0-x\|^2}{2\sum_{i=0}^k t_i}$$

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

Proof. Let LHS_k and RHS_k denote respectively the left-hand and right-hand sides in (24). We proceed by induction. For $k=0$ we have

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

The second step follows from (11) and $g_0^\varphi \in \partial\varphi(x_0),\ g_0^\psi \in \partial\psi(x_1)$. The third step follows from (12) 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

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

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

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

□

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 (22). We next show that in this case Proposition 1 yields the classical convergence rates (26) and (27), as well and the modern and more general one (28) 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 Proposition 1 it follows that

$$\sum_{i=0}^k t_i(\varphi(x_i) - \bar\varphi) \le \frac{\sum_{i=0}^k t_i^2\|g_i^\varphi\|^2 + \operatorname{dist}(x_0,\bar X)^2}{2}. \tag{25}$$

In particular, if $\|g\| \leq L$ for all $x \in C$ and $g \in \partial\varphi(x)$ then (25) implies

$$\min_{i=0,\ldots,k}(\varphi(x_i) - \bar{\varphi}) \leq \frac{\sum_{i=0}^k t_i^2 L^2 + \operatorname{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^k t_i}. \tag{26}$$

Let $\alpha_i := t_i\|g_i^\varphi\|$, $i = 0, 1, \ldots$. Then Step 4 in Algorithm 2 can be rewritten as $x_{k+1} = \Pi_C\left(x_k - \alpha_k \cdot \frac{g_k^\varphi}{\|g_k^\varphi\|}\right)$ provided $\|g_k^\varphi\| > 0$, which occurs as long as x_k is not an optimal solution to (22). If $\|g_i^\varphi\| > 0$ for $i = 0, 1, \ldots, k$ then (25) implies

$$\min_{i=0,\ldots,k}(\varphi(x_i) - \bar{\varphi}) \leq L \cdot \frac{\sum_{i=0}^k \alpha_i^2 + \operatorname{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^k \alpha_i}. \tag{27}$$

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 $g \in \partial\varphi(x)$

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

As discussed by Grimmer [6], $\mathcal{L}$-steepness is a more general and weaker condition than the traditional bound $\|g\| \leq L$ for all $x \in C$ and $g \in \partial\varphi(x)$. 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.

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 $i = 0, 1, \ldots, k$ then (25) implies

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

and thus

$$\min_{i=0,\ldots,k}(\varphi(x_i) - \bar{\varphi}) \leq \sup\left\{t : \frac{t}{\mathcal{L}(t)} \leq \frac{\sum_{i=0}^k \alpha_i^2 + \operatorname{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^k \alpha_i}\right\}. \tag{28}$$

Acknowledgements

This research has been funded by NSF grant CMMI-1534850.

References

[1] Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

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

[3] S. Bubeck, Y. Lee, and M. Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[4] D. Drusvyatskiy, M. Fazel, and S. Roy. An optimal first order method based on optimal quadratic averaging. *arXiv preprint arXiv:1604.06543*, 2016.

[5] N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *COLT*, pages 658–695, 2015.

[6] B. Grimmer. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *arXiv preprint arXiv:1712.04104*, 2017.

[7] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[8] P. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[9] Y. Nesterov. A method for unconstrained convex minimization problem with rate of convergence $\mathcal{O}(1/k^2)$. Doklady AN SSSR (in Russian). *(English translation. Soviet Math. Dokl.)*, 269:543–547, 1983.

[10] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Kluwer Academic Publishers, 2004.

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

[12] J. Peña. Convergence of first-order methods via the convex conjugate. *Operations Research Letters*, 45:561–564, 2017.

[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*, pages 2510–2518, 2014.