Operations Research Letters 45 (2017) 561–564

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Convergence of first-order methods via the convex conjugate

Javier Peña

Tepper School of Business, Carnegie Mellon University, USA

ARTICLE INFO

Article history:
Received 17 April 2017
Received in revised form 27 July 2017
Accepted 21 August 2017
Available online 18 September 2017

Keywords:
First-order methods
Conjugate
Acceleration

ABSTRACT

This paper gives a unified and succinct approach to 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 convex minimization. In the three cases the proof of convergence follows from a generic bound defined by the convex conjugate of the objective function.

1. Introduction

The subgradient, gradient, and accelerated gradient methods are icons in the class of first-order algorithms for convex optimization. Under a suitable Lipschitz continuity assumption on the objective function and a judicious choice of step-sizes, the subgradient method yields a point whose objective value is within $\mathcal{O}(1/\sqrt{k})$ of the optimal value after k iterations. In a similar vein, under a suitable Lipschitz continuity assumption on the gradient of the objective function and a judicious choice of step-sizes, the gradient and accelerated gradient methods yield points whose objective values are within $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ of the optimal value respectively after k iterations.

Although the proofs of the $\mathcal{O}(1/\sqrt{k})$, $\mathcal{O}(1/k)$, and $\mathcal{O}(1/k^2)$ convergence rates for these three algorithms share some common ideas, they are traditionally treated separately. In particular, the known proofs of the $\mathcal{O}(1/k^2)$ convergence rate of the accelerated gradient method, first established by Nesterov in a landmark paper [13], are notoriously less intuitive than those of the $\mathcal{O}(1/\sqrt{k})$ and $\mathcal{O}(1/k)$ convergence rates of the subgradient and gradient methods. Nesterov's accelerated gradient method has had a profound influence in optimization and has led to a vast range of developments. See, e.g., [4,5,14,17,19] and the many references therein.

Several recent articles [1,7,9,12,15,18] have proposed novel approaches that add insight and explain how the accelerated gradient method and some variants achieve a faster convergence rate. This paper makes a contribution of similar spirit. It provides a unified and succinct approach for deriving the convergence rates of the subgradient, gradient, and accelerated gradient algorithms. The crux of the approach is a generic upper bound via the convex conjugate of the objective function. (See Lemma 1 in Section 2.) The construction of the upper bound captures key common features and differences among the three algorithms.

The paper is self-contained and relies only on the basic convex analysis background recalled next. (For further details see [6,11,16].) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function. Endow $\mathbb{R}^n$ with an inner product $\langle\cdot,\cdot\rangle$ and let $\|\cdot\|$ denote the corresponding Euclidean norm. Given a constant $G > 0$, the function f is G-Lipschitz if for all $x, y \in \operatorname{dom}(f) := \{x \in \mathbb{R}^n : f(x) < \infty\}$

$$f(x) - f(y) \le G\|x - y\|.$$

Observe that if f is convex and G-Lipschitz then for all $x \in \operatorname{int}(\operatorname{dom}(f))$ and $g \in \partial f(x)$

$$g \in \partial f(x) \Rightarrow \|g\| \le G. \tag{1}$$

Suppose f is differentiable on $\operatorname{dom}(f)$. Given a constant $L > 0$, the gradient ∇f is L-Lipschitz if for all $x, y \in \operatorname{dom}(f)$

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|.$$

Observe that if f is differentiable and ∇f is L-Lipschitz then for all $x, y \in \operatorname{dom}(f)$

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2.$$

In particular, if $x \in \operatorname{dom}(f)$ is such that $x - \frac{1}{L}\nabla f(x) \in \operatorname{dom}(f)$ then

$$f\left(x - \frac{1}{L}\nabla f(x)\right) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|^2. \tag{2}$$

Let $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ denote the convex conjugate of f, that is,

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

E-mail address: jfp@andrew.cmu.edu.

http://dx.doi.org/10.1016/j.orl.2017.08.013
0167-6377/

The construction of the conjugate readily yields the following property known as *Fenchel's inequality*. For all $z, x \in \mathbb{R}^n$

$$f^*(z) + f(x) \geq \langle z, x \rangle$$

and equality holds if $z \in \partial f(x)$.

2. First-order methods for unconstrained convex optimization

Throughout the sequel assume $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and consider the problem

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

Let $\bar{f}$ and $\bar{X}$ respectively denote the optimal value and set of optimal solutions to (3).

Algorithms 1 and 2 describe respectively the subgradient method and accelerated gradient method for (3). The subgradient method becomes the gradient method when f is differentiable. Algorithm 2 is a variant of Nesterov's original accelerated gradient method [13]. This version has been discussed in [4,14,19].

Algorithm 1 Subgradient/gradient method

```
input: $x_0 \in \mathbb{R}^n$ and a convex function $f : \mathbb{R}^n \to \mathbb{R}$
for $k = 0, 1, 2, \ldots$ do
    pick $g_k \in \partial f(x_k)$ and $t_k > 0$
    $x_{k+1} := x_k - t_k g_k$
end for
```

Algorithm 2 Accelerated gradient method

```
input: $x_0 \in \mathbb{R}^n$ and a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$
$y_0 := x_0$, $\theta_0 := 1$
for $k = 0, 1, 2, \ldots$ do
    pick $t_k > 0$
    $x_{k+1} := y_k - t_k \nabla f(y_k)$
    let $\theta_{k+1} \in (0, 1)$ be such that $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$
    $y_{k+1} := x_{k+1} + \frac{\theta_{k+1}(1-\theta_k)}{\theta_k}(x_{k+1} - x_k)$
end for
```

Theorems 1, 2, and 3 state well-known convergence properties of Algorithms 1 and 2.

Theorem 1. *Suppose f is G-Lipschitz. Then the sequence of iterates $x_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ generated by Algorithm 1 satisfies*

$$\frac{\sum_{i=0}^k t_i f(x_i) - \frac{G^2}{2} \sum_{i=0}^k t_i^2}{\sum_{i=0}^k t_i} \leq f(x) + \frac{\|x_0 - x\|^2}{2 \sum_{i=0}^k t_i} \tag{4}$$

for all $x \in \mathbb{R}^n$. In particular, if $\bar{X} \neq \emptyset$ then $\min_{i=0,1,\ldots,k} f(x_i) - \bar{f} \leq \frac{\mathrm{dist}(x_0,\bar{X})^2 + G^2 \sum_{i=0}^k t_i^2}{2 \sum_{i=0}^k t_i}$, and $\min_{i=0,1,\ldots,k} f(x_i) - \bar{f} \leq \frac{\mathrm{dist}(x_0,\bar{X})^2 + G^2}{2\sqrt{k+1}}$ for $t_i = \frac{1}{\sqrt{k+1}}$, $i = 0, 1, \ldots, k$.

Theorem 2. *Suppose ∇f is L-Lipschitz and $t_k = \frac{1}{L}$, $k = 0, 1, \ldots$. Then the sequence of iterates $x_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ generated by Algorithm 1 satisfies*

$$\frac{f(x_1) + \cdots + f(x_k)}{k} \leq f(x) + \frac{L\|x_0 - x\|^2}{2k} \tag{5}$$

for all $x \in \mathbb{R}^n$. In particular, if $\bar{X} \neq \emptyset$ then $f(x_k) - \bar{f} \leq \frac{L\,\mathrm{dist}(x_0,\bar{X})^2}{2k}$.

Theorem 3. *Suppose f is differentiable, ∇f is L-Lipschitz, and $t_k = \frac{1}{L}$ for $k = 0, 1, \ldots$. Then the sequence of iterates $x_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ generated by Algorithm 2 satisfies*

$$f(x_k) \leq f(x) + \frac{L\theta_{k-1}^2 \|x_0 - x\|^2}{2} \tag{6}$$

for all $x \in \mathbb{R}^n$. In particular, if $\bar{X} \neq \emptyset$ then $f(x_k) - \bar{f} \leq \frac{2L\,\mathrm{dist}(x_0,\bar{X})^2}{(k+1)^2}$.

The central contribution of this paper is a unified approach to the proofs of Theorems 1, 2, and 3. The crux of the approach is the following lemma.

Lemma 1. *There exists a sequence $z_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ such that for $k = 1, \ldots$ and $\mu_k = \frac{1}{\sum_{i=0}^k t_i}$ the left-hand side LHS_k of (4) in Theorem 1 satisfies*

$$\begin{aligned} \mathsf{LHS}_k &\leq -f^*(z_k) + \langle z_k, x_0 \rangle - \frac{\|z_k\|^2}{2\mu_k} \\ &= -f^*(z_k) + \min_{u \in \mathbb{R}^n} \left\{ \langle z_k, u \rangle + \frac{\mu_k}{2}\|u - x_0\|^2 \right\}. \end{aligned} \tag{7}$$

There also exist sequences $z_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ such that (7) holds for $\mu_k = \frac{L}{k}$ and the left-hand side LHS_k of (5) in Theorem 2, as well as for $\mu_k = L\theta_{k-1}^2$ and the left-hand side LHS_k of (6) in Theorem 3.

Lemma 1 captures some key common features and differences among the subgradient, gradient, and accelerated gradient algorithms. The right-hand side in (7) has the same form in all cases and has the same kind of dependence on the initial point x_0. Furthermore, as Section 3 details, the construction of the sequences z_k, μ_k, $k = 1, 2 \ldots$ follows the same template for the three algorithms. However, some details of the construction for these sequences need to be carefully tailored to each of the three algorithms.

Proof of Theorems 1, 2, and 3. Lemma 1 and Fenchel's inequality imply that for some $z_k \in \mathbb{R}^n$, $k = 1, 2, \ldots$ and all $x \in \mathbb{R}^n$ the left-hand-sides LHS_k of (4), (5), and (6) satisfy

$$\begin{aligned} \mathsf{LHS}_k &\leq -f^*(z_k) + \min_{u \in \mathbb{R}^n} \left\{ \langle z_k, u \rangle + \frac{\mu_k}{2}\|u - x_0\|^2 \right\} \\ &\leq -f^*(z_k) + \langle z_k, x \rangle + \frac{\mu_k \cdot \|x - x_0\|^2}{2} \\ &\leq f(x) + \frac{\mu_k \cdot \|x - x_0\|^2}{2}. \end{aligned}$$

To finish, recall that $\mu_k = \frac{1}{\sum_{i=0}^k t_i}$ for (4), $\mu_k = \frac{L}{k}$ for (5), and $\mu_k = L\theta_{k-1}^2$ for (6). For the second part of Theorem 2 observe that $f(x_k) \leq \frac{f(x_1) + \cdots + f(x_k)}{k}$ because (2) implies that $f(x_{i+1}) \leq f(x_i) - \frac{1}{2L}\|\nabla f(x_i)\|^2 \leq f(x_i)$, $i = 0, 1, \ldots$. For the second part of Theorem 3 observe that a straightforward induction shows that the conditions $\theta_{k+1} \in (0, 1)$, $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$, and $\theta_0 = 1$ imply $\theta_{k-1} \leq \frac{2}{k+1}$. □

3. Proof of Lemma 1

Construct the sequences $\mu_k \in \mathbb{R}$, $z_k \in \mathbb{R}^n$, $k = 1, 2 \ldots$ as follows. First, choose sequences $\theta_k \in (0, 1), y_k \in \mathbb{R}^n, g_k \in \partial f(y_k)$, $k = 1, 2, \ldots$, and two initial values $\mu_0 \in \mathbb{R}_+, z_0 \in \mathbb{R}^n$ or $\mu_1 \in \mathbb{R}_+, z_1 \in \mathbb{R}^n$. Second, let $\mu_k \in \mathbb{R}$, $z_k \in \mathbb{R}^n$, $k = 1, 2 \ldots$ be defined by the rules

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

This construction readily implies

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