Efficiency of Line Search in Proximal Gradient Methods

Lin Xiao*

November 1, 2019

Abstract

Line search methods are very effective in practice for speeding up first-order methods for minimizing smooth functions. The step size found by a line-search procedure during each iteration can be regarded as the reciprocal of a local Lipschitz constant. We show that the convergence speed of first-order methods equipped with a simple line-search procedure depends on the harmonic mean of the local Lipschitz constants.

1 Introduction

We consider optimization problems of the form

$$\underset{x\in\mathbf{R}^n}{\text{minimize}} \quad F(x) := f(x) + \Psi(x), \tag{1}$$

where $\Psi : \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is convex and lower semi-continuous, and f is differentiable on an open set containing $\operatorname{dom}\Psi$. In addition, we assume that the gradient of f is Lipschitz continuous, i.e., there exists a constant $L_f > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f\|x-y\|, \qquad \forall\, x, y \in \operatorname{dom}\Psi, \tag{2}$$

where $\|\cdot\|$ denotes the standard Euclidean norm. We call L_f the *global* Lipschitz constant of ∇f.

Given an initial point $x_0 \in \operatorname{dom}\Psi$, the proximal gradient method computes a sequence of iterates $x_1, x_2, \ldots$ as follows:

$$x_{k+1} = \underset{x\in\mathbf{R}^n}{\arg\min}\left\{f(x_k) + \langle\nabla f(x_k), x - x_k\rangle + \frac{L_k}{2}\|x-x_k\|^2 + \Psi(x)\right\}, \tag{3}$$

where $L_k > 0$ is a parameter to be chosen at each iteration (see, e.g., [Nes13, Bec17]). This method is often written in the more compact form

$$x_{k+1} = \mathbf{prox}_{\frac{1}{L_k}\Psi}\left(x_k - \frac{1}{L_k}\nabla f(x_k)\right),$$

*Microsoft Research, Redmond, WA 98004, USA. Email: `lin.xiao@microsoft.com`.

where the proximal operator is defined as

$$\mathbf{prox}_{\Psi}(y) = \arg\min_{x} \left\{ \Psi(x) + \frac{1}{2}\|x - y\|^2 \right\}.$$

With the definition of the gradient mapping [Nes13]

$$g_L(x) := L\left(x - \mathbf{prox}_{\frac{1}{L}\Psi}\left(x - \frac{1}{L}\nabla f(x)\right)\right), \tag{4}$$

the proximal gradient method can also be written as

$$x_{k+1} = x_k - \frac{1}{L_k} g_{L_k}(x_k). \tag{5}$$

Notice that if $\Psi(x) = 0$, then $g_{L_k}(x_k) = \nabla f(x_k)$ for any $L_k > 0$. Here it is clear that $1/L_k$ corresponds to the step size.

The proximal gradient method is guaranteed to converge if we choose $L_k \geq L_f$ for all k. In practice, however, it is almost always beneficial to find L_k using a line search procedure during each iteration, even if the global Lipschitz constant L_f is known a priori. A typical line search procedure starts with a relatively small estimate of L_k (a large step size $1/L_k$) and gradually increases it (decreases the step size) until some exit condition is satisfied (see, e.g., [Nes13]). One obvious choice for the exit condition is

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2}\|x_{k+1} - x_k\|^2. \tag{6}$$

We call L_k a *local* Lipschitz constant if it satisfies (6). Under the assumption (2), any $L_k \geq L_f$ would satisfy (6). But L_k often can be much smaller than L_f, which corresponds to a much larger step size $1/L_k$ and faster convergence.

We will show that the convergence speed of the proximal gradient method depends on the harmonic mean of $L_0, L_1, \ldots, L_k$. In other words, we can replace L_f in the standard convergence rate results by the *harmonic mean* $\widetilde{L}_k$, which is defined through

$$\frac{1}{\widetilde{L}_k} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{1}{L_i}. \tag{7}$$

Since the harmonic mean is smaller than the geometric mean and can be much smaller than the arithmetic mean, we obtain tighter bounds on the convergence speed.

2 Non-convex case

Without assuming convexity of f, we measure the quality of the iterates x_k by $\|g_{L_k}(x_k)\|^2$, which is the same as $\|\nabla f(x_k)\|^2$ when $\Psi \equiv 0$. It is shown in [Nes13, Theorem 3] that for all $i \geq 0$,

$$\frac{1}{2L_i}\|g_{L_i}(x_i)\|^2 \leq F(x_i) - F(x_{i+1}). \tag{8}$$

Summing up these inequalities for $i = 0, \ldots, k$, we obtain

$$\sum_{i=0}^{k} \frac{1}{2L_i} \|g_{L_i}(x_i)\|^2 \le F(x_0) - F(x_{k+1}).$$

Assuming F is bounded below by $F_\star$ and using the definition of $\widetilde{L}_k$ in (7), we get

$$\min_{i \in \{0,\ldots,k\}} \|g_{L_i}(x_i)\|^2 \le \frac{2\widetilde{L}_k (F(x_0) - F_\star)}{k+1}.$$

3 Convex case

If the function f is convex, then [XZ14, Lemma 3.7] implies that for any $y \in \operatorname{dom} \Psi$ and any $k \ge 0$,

$$F(y) \ge F(x_{k+1}) + \langle g_{L_k}(x_k), y - x_k \rangle + \frac{1}{2L_k} \|g_{L_k}(x_k)\|^2 + \frac{\mu_f}{2} \|y - x_k\|^2 + \frac{\mu_\Psi}{2} \|y - x_{k+1}\|^2. \quad (9)$$

where μ_f and μ_Ψ are the convexity parameters of f and Ψ respectively. In this section, we do not assume strong convexity, therefore $\mu_f = \mu_\Psi = 0$. Suppose $x_\star$ is a solution to (1), i.e.,

$$x_\star \in \operatorname*{Arg\,min}_x \left\{ f(x) + \Psi(x) \right\}.$$

Then setting $y = x_\star$ in the inequality (9) with $\mu_f = \mu_\Psi = 0$ and rearranging terms, we obtain

$$\begin{aligned} F(x_{k+1}) - F(x_\star) &\le \langle g_{L_k}(x_k), x_k - x_\star \rangle - \frac{1}{2L_k} \|g_{L_k}(x_k)\|^2 \\ &= \frac{L_k}{2} \left(\|x_k - x_\star\|^2 - \left\| x_k - \frac{1}{L_k} g_{L_k}(x_k) - x_\star \right\|^2 \right) \\ &= \frac{L_k}{2} \left(\|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2 \right), \end{aligned}$$

where the last equality is due to (5). Summing up the above inequality for $i = 0, 1, \ldots, k$, we get

$$\sum_{i=0}^{k} \frac{1}{L_i} \big(F(x_{i+1} - F(x_\star) \big) \le \frac{1}{2} \|x_0 - x_\star\|^2 - \frac{1}{2} \|x_{k+1} - x_\star\|^2 \le \frac{1}{2} \|x_0 - x_\star\|^2.$$

From (8), we conclude that $\{F(x_k)\}$ is a decreasing sequence. Therefore,

$$\big(F(x_{k+1} - F(x_\star) \big) \sum_{i=0}^{k} \frac{1}{L_i} \le \sum_{i=0}^{k} \frac{1}{L_i} \big(F(x_{i+1} - F(x_\star) \big) \le \frac{1}{2} \|x_0 - x_\star\|^2,$$

which, combined with the definition of $\widetilde{L}_k$ in (7), yields

$$F(x_{k+1}) - F(x_\star) \le \frac{\widetilde{L}_k \|x_0 - x_\star\|^2}{2(k+1)}.$$

4 Strongly convex case

In this section, we assume $\mu_f + \mu_\Psi > 0$ in (9), i.e., at leas one of f and Ψ is strongly convex. In this case, let $x_\star$ be the unique solution to (1). Using the update formula (5), we have

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x_\star\|^2 &= \frac{1}{2}\left\|x_k - \frac{1}{L_k} g_{L_k}(x_k) - x_\star\right\|^2 \\ &= \frac{1}{2}\|x_k - x_\star\|^2 - \frac{1}{L_k}\left\langle g_{L_k}(x_k),\ x_k - x_\star\right\rangle + \frac{1}{2L_k^2}\|g_{L_k}(x_k)\|^2.\end{aligned}$$

Meanwhile, setting $y = x_\star$ in (9) yields

$$-\langle g_{L_k}(x_k), x_k - x_\star\rangle + \frac{1}{2L_k}\|g_{L_k}(x_k)\|^2 \le F(x_\star) - F(x_{k+1}) - \frac{\mu_f}{2}\|x_k - x_\star\|^2 - \frac{\mu_\Psi}{2}\|x_{k+1} - x_\star\|^2.$$

Combining the two inequalities above, we obtain

$$\frac{1}{2}\|x_{k+1} - x_\star\|^2 \le \frac{1}{2}\|x_k - x_\star\|^2 + \frac{F(x_\star) - F(x_{k+1})}{L_k} - \frac{\mu_f}{2L_k}\|x_k - x_\star\|^2 - \frac{\mu_\Psi}{2L_k}\|x_{k+1} - x_\star\|^2.$$

Multiplying both sides by L_k and rearranging terms, we get

$$F(x_{k+1}) - F(x_\star) + \frac{L_k + \mu_\Psi}{2}\|x_{k+1} - x_\star\|^2 \le \frac{L_k - \mu_f}{2}\|x_k - x_\star\|^2.$$

Since $F(x_{k+1}) - F(x_\star) \ge 0$, we have for all $k \ge 0$,

$$\|x_{k+1} - x_\star\|^2 \le \frac{L_k - \mu_f}{L_k + \mu_\Psi}\|x_k - x_\star\|^2 = \left(\prod_{i=0}^{k} \frac{L_i - \mu_f}{L_i + \mu_\Psi}\right)\|x_0 - x_\star\|^2.$$

From the two inequalities above, we obtain

$$F(x_{k+1}) - F(x_\star) \le \frac{L_k + \mu_\Psi}{2} \cdot \frac{L_k - \mu_f}{L_k + \mu_\Psi}\|x_k - x_\star\|^2 \le \frac{L_k + \mu_\Psi}{2}\left(\prod_{i=0}^{k} \frac{L_i - \mu_f}{L_i + \mu_\Psi}\right)\|x_0 - x_\star\|^2.$$

Using the arithmetic-geometric means inequality, we get

$$\prod_{i=0}^{k} \frac{L_i - \mu_f}{L_i + \mu_\Psi} = \prod_{i=0}^{k}\left(1 - \frac{\mu_f + \mu_\Psi}{L_i + \mu_\Psi}\right) \le \left(1 - \frac{1}{k+1}\sum_{i=0}^{k} \frac{\mu_f + \mu_\Psi}{L_i + \mu_\Psi}\right)^{k+1}$$

Finally, by defining the *shifted* harmonic mean $\widehat{L}_k$ through the equality

$$\frac{1}{\widehat{L}_k + \mu_\Psi} = \frac{1}{k+1}\sum_{i=0}^{k} \frac{1}{L_i + \mu_\Psi},$$

we have

$$F(x_k) - F(x_\star) \le \left(1 - \frac{\mu_f + \mu_\Psi}{\widehat{L}_k + \mu_\Psi}\right)^k \frac{L_f + \mu_\Psi}{2}\|x_0 - x_\star\|^2.$$

Notice that $\widehat{L}_k \ge \widetilde{L}_k$ and the equality holds if $\mu_\Psi = 0$. In any case, it can be much smaller than L_f.

5 Accelerated proximal gradient methods

When f is smooth and convex, we can apply the results of [HRX18, Theorem 5] (which considers the more general setting of relative smoothness) to the Euclidean case, and obtain the following accelerated convergence rate,

$$F(x_k + 1) - F(x_\star) \leq \frac{1}{A_k} \frac{\|x_0 - x_\star\|^2}{2},$$

where A_k satisfies

$$A_k^{1/2} \geq \sum_{i=1}^{k} \frac{1}{2L_i^{1/2}} + \frac{1}{L_0^{1/2}} = \frac{1}{2}\left(\sum_{i=1}^{k} \frac{1}{L_i^{1/2}} + \frac{1}{L_0^{1/2}} + \frac{1}{L_0^{1/2}}\right) = \frac{1}{2}\sum_{i=-1}^{k} \frac{1}{L_i^{1/2}},$$

where we used the definition $L_{-1} = L_0$. Let $\widetilde{L_k^{1/2}}$ be the harmonic mean of $L_{-1}^{1/2}, L_0^{1/2}, \ldots, L_k^{1/2}$, i.e.,

$$\frac{1}{\widetilde{L_k^{1/2}}} = \frac{1}{k+2}\sum_{i=-1}^{k} \frac{1}{L_i^{1/2}}.$$

Then we obtain

$$F(x_{k+1}) - F(x_\star) \leq \frac{4\left(\widetilde{L_k^{1/2}}\right)^2}{(k+2)^2} \frac{\|x_0 - x_\star\|^2}{2}. \tag{10}$$

Notice that $\left(\widetilde{L_k^{1/2}}\right)^2$ is smaller than the geometric and arithmetic means of $L_{-1}, L_0, \ldots, L_k$, i.e.,

$$\left(\widetilde{L_k^{1/2}}\right)^2 \leq \left(\prod_{i=-1}^{k} L_i\right)^{1/(k+2)} \leq \frac{1}{k+2}\sum_{i=-1}^{k} L_i \leq L_f\,.$$

Therefore, the convergence rate in (10) is slightly tighter than the result of [HRX18, Theorem 5], which used the geometric mean.

When f is also strongly convex, similar improvement of accelerated linear convergence rate can also be established. Here we omit the details.

References

[Bec17] Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. SIAM, 2017.

[HRX18] Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. arXiv:1808.03045, 2018.

[Nes13] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[XZ14] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.