Lecture 11: Acceleration via Restarting; Lower Bounds

Yudong Chen

Last week we discussed two variants of Nesterov's accelerated gradient descent (AGD).

Algorithm 1 Nesterov's AGD, smooth and strongly convex

input: initial x_0, strong convexity and smoothness parameters m, L, number of iterations K
initialize: $x_{-1} = x_0$, $\beta = \frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}$.
for $k = 0, 1, \ldots K$
  $y_k = x_k + \beta (x_k - x_{k-1})$
  $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
return x_K

Theorem 1. *For Nesterov's AGD Algorithm 1 applied to m-strongly convex L-smooth f, we have*

$$f(x_k) - f^* \le \left(1 - \sqrt{\frac{m}{L}}\right)^k \cdot \frac{(L+m)\left\|x_0 - x^*\right\|_2^2}{2}.$$

Equivalently, we have $f(x_k) - f^* \le \epsilon$ after at most $k = O\left(\sqrt{\frac{L}{m}} \log \frac{L\|x_0 - x^*\|_2^2}{\epsilon}\right)$ iterations.

Algorithm 2 Nesterov's AGD, smooth convex

input: initial x_0, smoothness parameter L, number of iterations K
initialize: $x_{-1} = x_0$, $\lambda_0 = 0$, $\beta_0 = 0$.
for $k = 0, 1, \ldots K$
  $y_k = x_k + \beta_k (x_k - x_{k-1})$
  $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
  $\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}$, $\beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}$
return x_K

Theorem 2. *For Nesterov's AGD Algorithm 2 applied to L-smooth convex f, we have*

$$f(x_k) - f(x^*) \le \frac{2L\left\|x_0 - x^*\right\|_2^2}{k^2}.$$

In this lecture, we will show that the two types of acceleration above are closely related: we can use one to derive the other. We then show that in a certain precise (but narrow) sense, the convergence rates of AGD are optimal among first-order methods. For this reason, AGD is also known as Nesterov's *optimal* method.

1 Acceleration via regularization

Suppose we only know the AGD method for *strongly* convex functions (Algorithm 1) and its $\left(1-\sqrt{\frac{m}{L}}\right)^k$ guarantee (Theorem 1). Can we use it as a subroutine to develop an accelerated algorithm for (non-strongly) convex functions with a $\frac{1}{k^2}$ convergence rate?

One approach is to add a "regularizer" $\epsilon \|x\|_2^2$ to $f(x)$ and apply Algorithm 1 to the function $f(x)+\epsilon\|x\|_2^2$, which is strongly convex. See HW 3.

2 Acceleration via restarting

In the opposite direction, suppose we only know the AGD method for (non-strongly) convex functions (Algorithm 2) and its $\frac{1}{k^2}$ guarantee (Theorem 2). Can we use it as a subroutine to develop an accelerated algorithm for *strongly* convex functions with a $\left(1-\sqrt{\frac{m}{L}}\right)^k$ convergence rate (equivalently, a $\sqrt{\frac{L}{m}}\log\frac{1}{\epsilon}$ iteration complexity)?

This is possible using a classical and powerful idea in optimization: *restarting*. See Algorithm 3. In each round, we run Algorithm 2 for $\sqrt{\frac{8L}{m}}$ iterations to obtain $\bar{x}_{t+1}$. In the next round, we restart Algorithm 2 using $\bar{x}_{t+1}$ as the initial solution and run for another $\sqrt{\frac{8L}{m}}$ iterations. This is repeated for T rounds.

Algorithm 3 Restarting AGD

input: initial $\bar{x}_0$, strong convexity and smoothness parameters m, L, number of rounds T
for $t = 0, 1, \ldots T$

Run Algorithm 2 with $\bar{x}_t$ (initial solution), L (smoothness parameter), $\sqrt{\frac{8L}{m}}$ (number of iterations) as the input. Let $\bar{x}_{t+1}$ be the output.

return $\bar{x}_T$

Exercise 1. How is Algorithm 3 different from running Algorithm 2 without restarting for $T \times \sqrt{\frac{8L}{m}}$ iterations?

Suppose f is m-strongly convex and L-smooth. By Theorem 2, we know that

$$f(\bar{x}_{t+1}) - f(x^*) \leq \frac{2L\|\bar{x}_t - x^*\|_2^2}{8L/m} = \frac{m\|\bar{x}_t - x^*\|_2^2}{4}.$$

By strong convexity, we have

$$f(\bar{x}_t) \geq f(x^*) + \underbrace{\langle \nabla f(x^*), \bar{x}_t - x^* \rangle}_{=0} + \frac{m}{2}\|\bar{x}_t - x^*\|_2^2,$$

hence $\|\bar{x}_t - x^*\|_2^2 \leq \frac{2}{m}\left(f(\bar{x}_t) - f(x^*)\right)$. Combining, we get

$$f(\bar{x}_{t+1}) - f(x^*) \leq \frac{f(\bar{x}_t) - f(x^*)}{2}.$$

That is, each round of Algorithm 3 halves the optimality gap. It follows that

$$f(\overline{x}_T) - f(x^*) \leq \left(\frac{1}{2}\right)^T (f(\overline{x}_0) - f(x^*)).$$

Therefore, $f(\overline{x}_T) - f(x^*) \leq \epsilon$ can be achieved after at most

$$T = O\left(\log \frac{f(\overline{x}_0) - f(x^*)}{\epsilon}\right) \text{ rounds,}$$

which corresponds to a total of

$$T \times \sqrt{\frac{8L}{m}} = O\left(\sqrt{\frac{L}{m}} \log \frac{f(\overline{x}_0) - f(x^*)}{\epsilon}\right) \text{ AGD iterations.}$$

This iteration complexity is the same as Theorem 1 up to a logarithmic factor.

Remark 1. Note how strong convexity is needed in the above argument.

3 Lower bounds

In this section, we consider an arbitrary iterative algorithm that uses first-order information and satisfies $x_0 = 0$, and

$$x_{k+1} \in \operatorname{Lin}\left\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k)\right\}, \qquad \forall k \geq 0, \tag{1}$$

where the RHS denotes the linear subspace spanned by $\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k)$; in other words, x_{k+1} is a linear combination of the gradients at previous $(k+1)$ iterates.

3.1 Smooth and convex f

Theorem 3. *There exists an L-smooth convex function f such that any first-order method above must satisfy*

$$f(x_k) - f(x^*) \geq \frac{3L \left\|x_0 - x^*\right\|_2^2}{32(k+1)^2}.$$

Comparing with the lower bound above, we see that the $\frac{L}{k^2}$ rate for AGD in Theorem 2 is optimal/unimprovable (up to constants).

Proof of Theorem 3. Let $A \in \mathbb{R}^{d \times d}$ be the matrix given by

$$A_{ij} = \begin{cases} 2, & i = j \\ -1, & j \in \{i-1, i+1\} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Explicitly,

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ & & \ddots & \ddots & \ddots & & \\ 0 & \cdots & & & -1 & 2 & -1 \\ 0 & \cdots & & & & -1 & 2 \end{bmatrix}.$$

Let $e_i \in \mathbb{R}^d$ denote the i-th standard basis vector. Consider the quadratic function

$$f(x) = \frac{L}{8} x^\top A x - \frac{L}{4} x^\top e_1,$$

which is convex and L-smooth since $0 \preccurlyeq A \preccurlyeq 4I$. By induction, we can show that for $k \geq 1$,

$$x_k \in \operatorname{Lin}\{e_1, Ax_1, \ldots, Ax_{k-1}\} \subseteq \operatorname{Lin}\{e_1, \ldots, e_k\}.$$

Therefore, if we let $A_k \in \mathbb{R}^{n \times n}$ denote the matrix obtained by zeroing out the entries of A outside the top-left k-by-k block, then

$$f(x_k) = \frac{L}{8} x_k^\top A_k x_k - \frac{L}{4} x_k^\top e_1 \geq f_k^* := \min_x \left\{ \frac{L}{8} x^\top A_k x - \frac{L}{4} x^\top e_1 \right\}.$$

By setting gradient to zero, we find that the minimum above is attained by

$$x_k^* := \left(1 - \frac{1}{k+1}, 1 - \frac{2}{k+1}, \ldots, 1 - \frac{k}{k+1}, 0, \ldots, 0\right)^\top \in \mathbb{R}^d,$$

with $f_k^* = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right)$. It follows that the global minimizer $x^* = x_d^*$ of f satisfies $f(x^*) = f_d^* = -\frac{L}{8}\left(1 - \frac{1}{d+1}\right)$ and

$$\|x^* - x_0\|_2^2 = \|x_d^*\|_2^2 = \sum_{i=1}^d \left(1 - \frac{i}{d+1}\right)^2 \leq \frac{d+1}{3}.$$

Combining pieces and taking $d = 2k+1$, we have

$$\begin{aligned} f(x_k) - f(x^*) \geq f_k^* - f(x^*) &= \frac{L}{8}\left(\frac{1}{k+1} - \frac{1}{2k+2}\right) \\ &= \frac{L}{16} \frac{k+1}{(k+1)^2} \\ &\geq \frac{3L}{32} \frac{\|x^* - x_0\|_2^2}{(k+1)^2}. \end{aligned}$$

□

3.2 Smooth and strongly convex f

For strongly convex functions, we have the following lower bound, which shows that the $\left(1 - \frac{1}{\sqrt{L/m}}\right)^k$ rate of AGD in Theorem 1 cannot be significantly improved.

Theorem 4. *There exists an m-strongly convex and L-smooth function such that any first-order method must satisfy*

$$f(x_k) - f(x^*) \geq \frac{m}{2}\left(1 - \frac{4}{\sqrt{L/m}}\right)^{k+1} \|x_0 - x^*\|_2^2.$$

Proof. Let $A \in \mathbb{R}^{d \times d}$ be defined in (2) above and consider the function

$$f(x) = \frac{L-m}{8}\left(x^\top A x - 2x^\top e_1\right) + \frac{m}{2}\|x\|_2^2,$$

which is L-smooth and m-strongly convex. Strong convexity implies

$$f(x_k) - f(x^*) \geq \frac{m}{2} \|x_k - x^*\|_2^2.$$

A similar argument as above shows that $x_k \in \mathrm{Lin}\{e_1, \ldots, e_k\}$, hence

$$\|x_k - x^*\|_2^2 \geq \sum_{i=k+1}^{d} x^*(i)^2.$$

For simplicity we take $d \to \infty$ (here we omit the formal limiting argument).[1] The minimizer x^* can be computed by setting the gradient of f to zero, which gives an infinite set of equaitons

$$\begin{aligned} 1 - 2\frac{L/m+1}{L/m-1}x^*(1) + x^*(2) = 0, \\ x^*(k-1) - 2\frac{L/m+1}{L/m-1}x^*(k) + x^*(k+1) = 0, \qquad k = 2, 3, \ldots \end{aligned}$$

Solving these equations gives

$$x^*(i) = \left(\frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}\right)^i, \qquad i = 1, 2, \ldots$$

Combining pieces, we obtain

$$\begin{aligned} f(x_k) - f(x^*) &\geq \frac{m}{2} \sum_{i=k+1}^{\infty} x^*(i)^2 \\ &\geq \frac{m}{2} \left(\frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}\right)^{2(k+1)} \|x_0 - x^*\|_2^2 \\ &= \frac{m}{2} \left(1 - \frac{4}{\sqrt{L/m}+1} + \frac{4}{(\sqrt{L/m}+1)^2}\right)^{k+1} \|x_0 - x^*\|_2^2 \\ &\geq \frac{m}{2} \left(1 - \frac{4}{\sqrt{L/m}}\right)^{k+1} \|x_0 - x^*\|_2^2. \end{aligned}$$

□

Remark 2. The lower bounds in Theorems 3 and 4 are in the worst-case/minimax sense: one cannot find a first-order method that achieves a better convergence rate on *all* smooth convex functions than AGD. This, however, does not prevent better rates to be achieved for a sub class of such functions. It is also possible to achieve better rates by using higher-order information (e.g., the Hessian).

[1] Note that the convergence rates for AGD in Theorems 1 and 2 do not explicitly depend on the dimension d. These results can be generalized to infinite dimensions.