Lecture 11: Acceleration via Regularization and 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 $k = 0, 1, \ldots K$
  $y_k = x_k + \beta\left(x_k - x_{k-1}\right)$
  $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
return x_K

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

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

Equivalently, we have $f(x_k) - f^* \leq \epsilon$ 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 $k = 0, 1, \ldots K$
  $y_k = x_k + \beta_k\left(x_k - x_{k-1}\right)$
  $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$
  $\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}$, $\beta_{k+1} = \frac{\lambda_k - 1}{\lambda_{k+1}}$
return x_K

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

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

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 $(1-\sqrt{\frac{m}{L}})^k$ 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?

The answer is yes (up to logarithmic factors). 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.

Add regularizer $\Rightarrow$ make loss strongly cvx

for cvx f. $g(x) = f(x) + \epsilon\|x\|_2^2$ is 2ϵ-strongly cvx.

Hence. $g(x_k)$, $g(x^*)$ satisfies Thm 1. Note that $x^*=0$ is a minimizer of $g(x)$ because ① $x^*=0$ minimize $f(x)$

② $x^*=0$ minimize $\|x\|_2^2$.

$\underline{\underline{m = 2\epsilon}}$

$$g(x_k) - g(x^*) \le \left(1-\sqrt{\frac{2\epsilon}{L}}\right)^k \frac{(L+2\epsilon)\|x_0-x^*\|^2}{2}$$

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

A common intuition: set $\epsilon \to 0$. Let's say, we require $f(x_k)-f(x^*) \le \delta$.

Set $\epsilon = \delta$. $\left(1-\sqrt{\frac{2\epsilon}{L}}\right)^k = \left(1-\sqrt{\frac{2\delta}{L}}\right)^k \le \frac{\delta}{C}$. $\Rightarrow k \simeq O\left(\sqrt{\frac{L}{\delta}}\log\frac{1}{\delta}\right)$

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 $(1-\sqrt{\frac{m}{L}})^k$ 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?

Ans: Remove previous λ_k, β_k's update. From analysis, $\beta_k \uparrow$ as $k \uparrow$, results (for large k) in increasing momentum. Restarting makes momentum not dramatically large.

2.1 Analysis. (Algo 3).

For m-strongly cvx, smooth f. Apply Thm 2.

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

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

$\Rightarrow \|\bar{x}_t - x^*\|_2^2 \le \frac{2}{m}(f(\bar{x}_t) - f(x^*))$. Plug back.

$f(\bar{x}_{t+1}) - f(x^*) \le \frac{1}{2}(f(\bar{x}_t) - f(x^*))$. Hence $f(\bar{x}_T) - f(x^*) \le (\frac{1}{2})^T (f(\bar{x}_0) - f(x^*))$

To achieve $f(\bar{x}_T) - f(x^*) \le \varepsilon$, only need $T = O\left(\log \frac{f(\bar{x}_0) - f(x^*)}{\varepsilon}\right)$ #.

Total # of iterations: $T \times \sqrt{\frac{8L}{m}} = O\left(\sqrt{\frac{L}{m}} \log \frac{f(\bar{x}_0) - f(x^*)}{\varepsilon}\right)$

通过多次 restarting, 逐步相对抵消 对于 Iteration complexity 的影响可以被消补. (Const factor / remainder influence).

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.

Remark 2. Optional reading: This overview article discusses restarting as a general/meta algorithmic technique.

3 Lower bounds

In this section, we consider a class of first-order iterative algorithms that satisfy $x_0 = 0$, and

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

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, $\boxed{x_{k+1}}$ is an (arbitrary) linear combination of the gradients at the 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 in the sense of (1) must satisfy*

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

Comparing with this lower bound, 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,

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

A对角占优
对角元 ≥ 0.
⇓
A p.s.d.

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

$\nabla f(x) = \frac{L}{4}(Ax - e_1)$

$\nabla^2 f(x) = \frac{L}{4} A \succeq 0.$ $\quad LI - \nabla^2 f(x) = \frac{L}{4}(4I - A) \succeq 0.$

p.s.d.

Hence $\quad 0 \preceq \nabla^2 f(x) \preceq LI.$ $\quad \Rightarrow f(x)$ is cvx, L-smooth.

By induction, we can show $x_k \in \text{Lin}\{e_1, Ax_1, \cdots, Ax_{k-1}\}$
$\subseteq \text{Lin}\{e_1, \cdots, e_k\}.$ At most # k non-zero entries.

Let $A_k \in \mathbb{R}^{d\times d}$.

$$(A_k)_{ij} = \begin{cases} a_{ij} & 1 \le i,j \le k \\ 0 & \text{else.} \end{cases}$$

Then $f(x_k) = \frac{L}{8} x_k^T A x_k - \frac{L}{4} x_k^T e_1 = \frac{L}{8} x_k^T A_k x_k \geq f_k^* = \min_x \{\frac{L}{8} x^T A_k x - \frac{L}{4} x^T e_1\}$

This min can be derived directly. Set $\nabla f_k(x) = 0.$

$$\Rightarrow \quad A_k X_k^* - e_1 = 0.$$

解析解: $X_k^* = \left(1-\frac{1}{k+1}, 1-\frac{2}{k+1}, \cdots, 1-\frac{k}{k+1}, 0, \cdots 0\right)^T$

(Verify this).

$$\Rightarrow \quad f_k^* = -\frac{L}{4}\left(1-\frac{1}{k+1}\right) + \frac{L}{8}\cdot\frac{k}{k+1} = -\frac{L}{8}\frac{k}{k+1}$$

$$\sum_{i,j} a_{ij} x_i x_j$$

$$\begin{cases} 2x_1^2 - x_1x_2 = \frac{2k^2-k^2+k}{(k+1)^2} = \frac{k}{k+1} \\ -x_1x_2 + 2x_2^2 - x_2x_3 = \frac{-k(k-1)+2(k-1)^2-(k-1)(k-2)}{(k+1)^2} = 0. \\ \vdots \qquad \qquad \vdots = 0 \\ 2x_k^2 - x_{k-1}x_k = 0. \qquad (x_{k-1} = 2x_k) \end{cases}$$

Proof.

$$\begin{cases} x_3 = 2x_4 \\ x_2 = 3x_4 \\ x_1 = 6x_4 - 2x_4 = 4x_4. \end{cases}$$

$$k \cdot 2 - (k-1) = k+1.$$

$$kx_k \cdot 2 - (k-1)x_k = 1$$

$$x_k = \frac{1}{k+1}$$

$$x_1 = kx_k = \left(1-\frac{1}{k+1}\right)X$$

$$\vdots$$

$$x_i = \left(1-\frac{i}{k+1}\right)x_k.$$

Then for $f(x) = f_d(x)$, $f^* = f_d^* = -\frac{L}{8}\frac{d}{d+1}$ Set $x_0 = 0$.

In addition, $\|x_d^* - x_0\|_2^2 = \sum_{i=1}^{d}\left(1-\frac{i}{d+1}\right)^2 = \frac{1}{(d+1)^2}\sum_{i=1}^{d}(d+1-i)^2 = \frac{1}{(d+1)^2}\sum_{i=1}^{d} i^2$

$$= \frac{1}{(d+1)^2}\cdot\frac{d(d+1)(2d+1)}{6} = \frac{d(2d+1)}{6(d+1)} \le \frac{2(d+1)^2}{6(d+1)} = \frac{d+1}{3}$$

$$\Rightarrow \quad d+1 \ge 3\|x_0 - x^*\|_2^2.$$

Take $\boxed{d = 2k+1}$

$$f(x_k) - f(x^*) \ge f_k^* - f_d^* = -\frac{L}{8}\left(\frac{k}{k+1} - \frac{d}{d+1}\right) = -\frac{L}{8}\left(\frac{k}{k+1} - \frac{2k+1}{2k+2}\right)$$

$$= -\frac{L}{8}\cdot\frac{-1}{2k+2} = \frac{L}{16(k+1)} = \frac{L(k+1)}{16(k+1)^2} \ge \frac{L(d+1)}{32(k+1)^2} \ge \frac{3L\|x_d^*-x_0\|^2}{32(k+1)^2}$$

$$= \frac{3L\|x^*-x_0\|_2^2}{32(k+1)^2}.$$

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 in the sense of (1) must satisfy*

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

Pf: Let $A \in \mathbb{R}^{d \times d}$ still the same as above.

Consider

$$f(x) = \frac{L-m}{8}(x^T A x - 2x^T e_1) + \frac{m}{2}\|x\|_2^2.$$

Easy to check f $\begin{cases} m\text{-strongly cvx.} \\ L\text{-smooth.} \end{cases}$

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

Similarly we get $x_k \in \mathrm{Lin}\{x_1, \dots, x_k\}$.

Hence $$0 \leq \sum_{i=1}^{k} |x_i - x_i^*|^2 = \|x - x^*\|_2^2 - \sum_{i=k+1}^{d} |x_i - x_i^*|^2$$

$$\|x - x^*\|^2 \geq \sum_{i=k+1}^{d} |\underset{0}{\underset{\|}{x_i}} - x_i^*|^2 = \sum_{i=k+1}^{d} |x_i^*|^2$$

where $x^*(i)$ denotes the ith entry of x^*. For simplicity we take $d \to \infty$ (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 equations

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

Solving these equations gives

$$x^*(i) = \left(\frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}\right)^i, \qquad i = 1, 2, \dots \tag{5}$$

Let $d \to \infty$

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

$$\geq \frac{m}{2}\sum_{i=k+1}^{d} |x_i^*|^2 = \frac{m}{2}\sum_{i=k+1}^{\infty}\left(\frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}\right)^{2i} \geq \frac{m}{2}\left(\frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}\right)^{2(k+1)}\|x^* - x_0\|_2^2$$

$$\frac{\sum_{i=k+1}^{d}|x_i^*|^2}{\|x^* - x_0\|_2^2} \geq \left(\frac{\sqrt{L/m}-1}{\sqrt{L/m}+1}\right)^{2(k+1)}$$

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

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

$\left(1 - \sqrt{\frac{m}{L}}\right)^k$. conv rate.

Remark 3. 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]The convergence rates for AGD in Theorems 1 and 2 do not explicitly depend on the dimension d, hence these results can be generalized to infinite dimensions.