

Lecture 8 (Sep 12)

Scribe?

Last time

- ▷ Better guarantees for convex f
- ▷ Strongly convex

Today

- ▷ Accelerated gradient descent.
- ▷ Lower bounds

Everything we will see today was originally developed by Nesterov.



So far we have seen that GD yields

L -smooth

$$f(x_k) - \min f \leq O\left(\frac{1}{k}\right)$$

L -smooth
 μ -strongly
convex

$$f(x_k) - \min f \leq O\left(\left(\frac{\mu-1}{\mu+1}\right)^{2k}\right)$$

↑ condition number $\frac{L}{\mu}$.

Question: Can we have a faster algorithm that only have access to gradients? Yes! We'll see an alg for L-smooth in HW you'll handle the other case.

In 1983, Nesterov published a paper with a mysterious method.

It updates two sequences:

$$\lambda_{k+1} \leftarrow (1 + \sqrt{1 + 4\lambda_k}) / 2$$

$$y_{k+1} \leftarrow x_k - \frac{1}{L} \nabla f(x_k)$$

$$x_{k+1} \leftarrow y_{k+1} + \frac{(\lambda_k - 1)}{\lambda_{k+1}} (y_{k+1} - y_k).$$

To gain some intuition let's watch a video.

In this class we will analyze this method.

Theorem: Let f be a convex function with L-Lipschitz gradient. Then for any $\min x^*$,

$$f(y_k) - \min f \leq \frac{2L \|x_0 - x^*\|^2}{k^2}.$$

Proof: We start with two Lemmas

Lemma 1: The seq of λ_k 's satisfies

$$\lambda_{k+1}^2 - \lambda_{k+1} = \lambda_k \text{ and for any } k \geq 1$$

$$\lambda_k \geq \frac{k+1}{2}.$$

+

Proof: Identity follows from the formula.

For the second part

$$\begin{aligned} \lambda_{k+1} &= \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2} \geq \frac{1}{2} + \frac{\sqrt{4\lambda_k^2}}{2} \geq \frac{1}{2} + \lambda_k \\ &\geq \frac{k+1}{2} + \lambda_0 \end{aligned}$$

□

Lemma 2: For any u, v

$$\begin{aligned} f(u - \frac{1}{L} \nabla f(u)) - f(v) &\leq -\frac{1}{2L} \|\nabla f(u)\|^2 \\ &\quad + \nabla f(u)^T (u-v). \end{aligned}$$

Proof: Use convexity and DL

$$\begin{aligned} f(u - \frac{1}{L} \nabla f(u)) - f(v) &\leq f(u - \frac{1}{L} \nabla f(u)) - (f(u) + \nabla f(u)^T (v-u)) \\ &\leq -\frac{1}{2L} \|\nabla f(u)\|^2 + \nabla f(u)^T (u-v). \end{aligned}$$

□

Our goal is to use these Lemmas
to find a recursion of $\delta_k = f(y_k) - \min f$.

Apply Lemma 2 with $u = x_k$, $v = y_k$

$$\begin{aligned} \delta_{k+1} - \delta_k &= f(y_{k+1}) - f(y_k) \leq -\frac{1}{2L} \|\nabla f(x_k)\|^2 + \\ &\quad \nabla f(x_k)^T (x_k - y_k) \\ (\heartsuit) \quad &\leq -\frac{L}{2} \|y_{k+1} - x_k\|^2 + L (y_{k+1} - x_k)^T (x_k - y_k). \end{aligned}$$

Apply Lemma 2 with $u = x_k$, $v = x^*$

$$\begin{aligned} \delta_{k+1} &= f(y_{k+1}) - \min f \leq -\frac{1}{2L} \|\nabla f(x_k)\|^2 + \\ &\quad \nabla f(x_k)^T (x_k - x^*) \\ (\heartsuit) \quad &\leq -\frac{L}{2} \|y_{k+1} - x_k\|^2 + L (y_{k+1} - x_k)^T (x_k - x^*). \end{aligned}$$

Adding up $(\lambda_k - 1)(\heartsuit) + (\heartsuit)$ gives

$$\begin{aligned} \lambda_k \delta_{k+1} - (\lambda_k - 1) \delta_k &\leq -\frac{L \lambda_k}{2} \|y_{k+1} - x_k\|^2 \\ &\quad - L (y_{k+1} - x_k)^T (\lambda_k x_k - (\lambda_k - 1)y_k \\ &\quad - x^*) \end{aligned}$$

Multiplying by λ_k gives

$$\begin{aligned}
& \lambda_k^2 \delta_{k+1} - (\lambda_k^2 - \lambda_k) \delta_k \leq \\
& -\frac{L}{2} \left[\|x(y_{k+1} - x_k)\|^2 + 2\lambda_k(y_{k+1} - x_k)^T (\lambda_k x_k - (\lambda_k - 1)y_k - x^*) \right] \\
& = -\frac{L}{2} \left(\underbrace{\| \lambda_k(y_{k+1} - x_k) + \lambda_k x_k - (\lambda_k - 1)y_k - x^* \|^2}_{= \| \lambda_k x_k - (\lambda_k - 1)y_k - x^* \|^2} \right)
\end{aligned}$$

By def

$$x_{k+1} = y_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}} (y_{k+1} - y_k)$$



$$\lambda_{k+1} x_{k+1} - (\lambda_{k+1} - 1) y_{k+1} = \lambda_k y_{k+1} - (\lambda_k - 1) y_k.$$

$$\begin{aligned}
(\lambda_k^2 \delta_{k+1} - \lambda_k^2 \delta_k) &= -\frac{L}{2} \left(\underbrace{\| \lambda_{k+1} x_{k+1} - (\lambda_{k+1} - 1) y_{k+1} - x^* \|^2}_{u_{k+1} :=} \right. \\
&\quad \left. - \underbrace{\| \lambda_k x_k - (\lambda_k - 1) y_k - x^* \|^2}_{u_k :=} \right)
\end{aligned}$$

Summing up from $k=1$ to $k=T-1$ yields

$$\lambda_{T-1}^2 \delta_T - \cancel{\lambda_0^2 \delta_1} \leq -\frac{L}{2} \left(\| \overbrace{u_T}^{z^0} \|^2 - \| u_1 \|^2 \right)$$

$$\begin{aligned}
 &\leq \frac{L}{2} \|\lambda_i x_i - (\lambda_i - 1)y_i - x^*\|^2 \\
 &= \frac{L}{2} \|x_i - x^*\|^2.
 \end{aligned}$$

Then

$$\delta_T \leq \frac{L \|x_i - x^*\|^2}{2 \lambda_{T-1}^2} \leq \frac{2L \|x_i - x^*\|^2}{T^2}$$

□

We just prove that there is an alg. significantly faster than GD!

AGD with 1000 iterations gives the "same" error than GD with 1000.000!

Lower bounds

Question: Can there be a faster algorithm only using gradients?

NO! AGD is the fastest.

Assumption: The given method produces iterates satisfying Subspace spanned by

$$x_k \in x_0 + \text{span} \{ \nabla f(x_0), \dots, \nabla f(x_{k-1}) \}$$

For example for GD we have

$$x_k = x_0 - \sum_{i=0}^{K-1} \alpha_i \nabla f(x_i).$$

This is dimension
dependent

Theorem For any $1 \leq K \leq \frac{1}{2}(d-1)$ and $L \geq 0$, there exists a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ with L -Lips grad such that for any algo. satisfying Assumption 1, we have

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

$$\|x_k - x^*\|^2 \geq \frac{1}{32} \|x_0 - x^*\|^2.$$

+

Proof: Next, we will build "the worst function" in the world.

Let

$$A_K = \left[\begin{array}{cc|c} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & \\ & & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 & \\ \hline & & & & & & 0 \end{array} \right]$$



$$\begin{matrix} 2 & -1 \\ 0 & 3/2 & -1 \\ 0 & 4/3 \end{matrix}$$