Lecture 8

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

↑ 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:

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

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

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

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^*,

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

Proof: We start with two Lemmas

Lemma 1: The seq $\{\lambda_k\}$ satisfies $\lambda_{k+1}^2 - \lambda_{k+1} = \lambda_k^2$ and for any $k \ge 1$

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

Proof: Identity follows from the formula.
For the second part

$$\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2} \ge \frac{1}{2} + \frac{\sqrt{4\lambda_k^2}}{2} \ge \frac{1}{2} + \lambda_k$$
$$\ge \frac{k+1}{2} + \lambda_0 \qquad \square$$

Lemma 2: For any u, v

$$f\left(u - \frac{1}{L}\nabla f(u)\right) - f(v) \le -\frac{1}{2L}\|\nabla f(u)\|^2 + \nabla f(u)^T(u-v).$$

Proof: Use convexity and DL

$$f\left(u - \frac{1}{L}\nabla f(u)\right) - f(v) \le f\left(u - \frac{1}{L}\nabla f(u)\right) - \left(f(u) + \nabla f(u)^T(v-u)\right)$$
$$\le -\frac{1}{2L}\|\nabla f(u)\|^2 + \nabla f(u)^T(u-v).$$
$\square$

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

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

$$\delta_{k+1} - \delta_k = f(y_{k+1}) - f(y_k) \le -\frac{1}{2L}\|\nabla f(x_k)\|^2 + \nabla f(x_k)^T(x_k - y_k)$$

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

(☺) $$\le -\frac{L}{2}\|y_{k+1} - x_k\|^2 - L(y_{k+1} - x_k)^T(x_k - y_k).$$

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

$$\delta_{k+1} = f(y_{k+1}) - \min f \le -\frac{1}{2L}\|\nabla f(x_k)\|^2 + \nabla f(x_k)^T(x_k - x^*)$$

(♡) $$\le -\frac{L}{2}\|y_{k+1} - x_k\|^2 - L(y_{k+1} - x_k)^T(x_k - x^*).$$

Adding up $(\lambda_k - 1)$ (☺) + (♡) gives

$$\lambda_k \delta_{k+1} - (\lambda_k - 1)\delta_k \le -\frac{L\lambda_k}{2}\|y_{k+1} - x_k\|^2 - L(y_{k+1} - x_k)^T(\lambda_k x_k - (\lambda_k - 1)y_k - x^*)$$

Multiplying by λ_k gives

$$\lambda_k^2 \delta_{k+1} - (\overbrace{\lambda_k^2 - \lambda_k}^{\lambda_{k-1}^2})\delta_k \le$$

$$-\frac{L}{2}\left[\|\lambda_k(y_{k+1}-x_k)\|^2 + 2\lambda_k(y_{k+1}-x_k)^T(\lambda_k x_k - (\lambda_k-1)y_k - x^*)\right]$$

$$= -\frac{L}{2}\left(\|\overbrace{\lambda_k(y_{k+1}-\cancel{x_k}) + \lambda_k \cancel{x_k} - (\lambda_k-1)y_k - x^*}^{\lambda_k y_{k+1} - (\lambda_k-1)y_k - x^*}\|^2 - \|\lambda_k x_k - (\lambda_k-1)y_k - x^*\|^2\right)$$

By def

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

$$\Updownarrow$$

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

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

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

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

$$\leq \frac{L}{2} \left\| \lambda_1 x_1 - \cancelto{0}{(\lambda_1 - 1)} y_1 - x^* \right\|^2$$

$$= \frac{L}{2} \| x_1 - x^* \|^2 .$$

Then

$$\delta_T \leq \frac{L \| x_1 - x^* \|^2}{2 \lambda_{T-1}^2} \leq \frac{2L \| x_1 - x^* \|^2}{T^2} \qquad \square$$

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!