Lecture 11. (Sketch.)

- Today we'll cover gradient descent of smooth objectives.
- We'll introduce some convexity along the way.
- Some good references:
  - Optimization: "Convex optimization: algorithms & complexity" by Sebastien Bubeck; "Introductory lectures on convex optimization", Yurii Nesterov; "Fundamentals of Convex Analysis", Claude Lemarechal and Jean-Baptiste Hiriart-Urruty.
- **Note:** I've added a homework problem.

1. Smooth objectives in ML.

- We say "f is β-smooth" to mean β-Lipschitz gradients:

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|.$$

(The math community says "smooth" for C^∞.)

- We primarily invoke smoothness via the key inequality

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2.$$

The right hand side is a quadratic which upper bounds f, and shares function values and gradients with f at x. In words: for any point x, there exists a quadratic function

Proof of smoothness inequality.

$$\begin{aligned}
&\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| \\
&= \left| \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle \,\mathrm{d}t - \langle \nabla f(x), y - x \rangle \right| \\
&\leq \int_0^1 \left| \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \right| \mathrm{d}t \\
&\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\| \,\mathrm{d}t \\
&\leq \int_0^1 t\beta \|y - x\|^2 \,\mathrm{d}t \\
&= \frac{\beta}{2}\|y - x\|^2.
\end{aligned}$$

Example: least squares.

Define $f(w) := \frac{1}{2}\|Xw - y\|^2$, and note $\nabla f(w) = X^\top(Xw - y)$. For any w, w',

$$\begin{aligned}
f(w') &= \frac{1}{2}\|Xw' - Xw + Xw - y\|^2 \\
&= \frac{1}{2}\|Xw' - Xw\|^2 + \left\langle Xw' - Xw, Xw - y \right\rangle + \frac{1}{2}\|Xw - y\|^2 \\
&= \frac{1}{2}\|Xw' - Xw\|^2 + \left\langle w' - w, \nabla f(w) \right\rangle + f(w).
\end{aligned}$$

- Since $\frac{\sigma_{\min}(X)}{2}\|w' - w\|^2 \leq \frac{1}{2}\|Xw' - Xw\|^2 \leq \frac{\sigma_{\max}(X)}{2}\|w' - w\|^2$, thus f is $\sigma_{\max}(X)$-smooth (and $\sigma_{\min}$-strongly-convex, as we'll discuss).
- The smoothness bound holds **with equality** if we use the seminorm $\|v\|_X = \|Xv\|$. We'll discuss smoothness wrt other norms in homework.

2. Convergence of gradient descent to critical points.

Define the *gradient iteration*

$$w' := w - \eta \nabla f(w),$$

where $\eta \geq 0$ is the step size. When f is β smooth but not necessarily convex, the smoothness inequality directly gives

$$\begin{aligned} f(w') &\leq f(w) + \left\langle \nabla f(w), w' - w \right\rangle + \frac{\beta}{2}\|w' - w\|^2 \\ &= f(w) - \eta\|\nabla f(w)\|^2 + \frac{\beta\eta^2}{2}\|\nabla f(w)\|^2 \\ &= f(w) - \eta\left(1 - \frac{\beta\eta}{2}\right)\|\nabla f(w)\|^2. \end{aligned}$$

If we choose η appropriately ($\eta \leq 2/\beta$) then: either we are near a critical point ($\nabla f(w) \approx 0$), or we can decrease f.

Let's refine our notation to tell iterates apart:

1. Let w_0 be given.
2. Recurse: $w_i := w_{i-1} - \eta_i \nabla f(w_{i-1})$.

Rearranging our iteration inequality and summing over $i < t$,

$$\begin{aligned} \sum_{i<t} \eta_{i+1}\left(1 - \frac{\beta\eta_{i+1}}{2}\right)\|\nabla f(w_i)\|^2 &\leq \sum_{i<t}\left(f(w_i) - f(w_{i+1})\right) \\ &= \left(f(w_0) - f(w_t)\right) \end{aligned}$$

We can summarize these observations in the following theorem.

Theorem. Let $(w_i)_{i\geq 0}$ be given by gradient descent on β-smooth f.

- If $\eta \in [0, 2/\beta]$, then $f(w_{i+1}) \leq f(w_i)$.
- If $\eta := 1/\beta$, then

$$\begin{aligned} \min_{i<t}\|\nabla f(w)\|^2 \leq \frac{1}{t}\sum_{i<t}\|\nabla f(w)\|^2 &\leq \frac{2\beta}{t}\left(f(w_0) - f(w_t)\right) \\ &\leq \frac{2\beta}{t}\left(f(w_0) - \inf_w f(w)\right). \end{aligned}$$

Remarks.

- We have no guarantee about the last iterate $\|\nabla f(w_t)\|$: we may get near a flat region at some $i < t$, but thereafter bounce out.
- This derivation is at the core of many papers with a "local optimization" (critical point) guarantee for gradient descent.
- The gradient iterate with step size $1/\beta$ is the result of minimizing the quadratic provided by smoothness:

$$w - \frac{1}{\beta}\nabla f(w) = \arg\min_{w'}\left(f(w) + \left\langle \nabla f(w), w' - w \right\rangle + \frac{\beta}{2}\|w' - w\|^2\right)$$

Remarks (continued).

- In t iterations, we found a point w with $\|\nabla f(w)\| \leq \sqrt{2\beta/t}$. We can do better with Nesterov-Polyak cubic regularization: by choosing the next iterate according to

$$\arg\min_{w'} \left(f(w) + \left\langle \nabla f(w), w' - w \right\rangle + \frac{1}{2} \left\langle \nabla^2 f(w)^{-1}(w' - w), w' - w \right\rangle + \frac{L}{6}\|w' - w\|^3 \right)$$

where $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$, then after t iterations, some iterate w satisfies

$$\|\nabla f(w)\| \leq \frac{\mathcal{O}(1)}{t^{2/3}}, \qquad \nabla^2 f(w) \succeq -\frac{\mathcal{O}(1)}{t^{1/3}}.$$

Note: it is not obvious that the above cubic can be solved efficiently, but indeed there are various ways. If we go up a few higher derivatives, it becomes NP-hard.

Remarks (continued).

- Gradient descent alone is known to avoid saddle points, see "Gradient Descent Only Converges to Minimizers" by Jason Lee, Max Simchowitz, Michael I Jordan, Ben Recht.

3. Convergence rate for smooth & convex.

Theorem. Suppose f is β-smooth and convex, and $(w_i)_{\geq 0}$ given by GD with $\eta_i := 1/\beta$. Then for any z,

$$f(w_t) - f(z) \leq \frac{\beta}{2t} \left(\|w_0 - z\|^2 - \|w_t - z\|^2 \right).$$

Remark. We only invoke convexity via the inequality

$$f(w') \geq f(w) + \left\langle \nabla f(w), w' - w \right\rangle,$$

meaning f lies above all tangents.

Proof. By convexity and the earlier smoothness inequality $\|\nabla f(w)^2\|^2 \leq 2\beta(f(w) - f(w'))$,

$$\begin{aligned}
\|w' - z\|^2 &= \|w - z\|^2 - \frac{2}{\beta}\left\langle \nabla f(w), w - z \right\rangle + \frac{1}{\beta^2}\|\nabla f(w)\|^2 \\
&\leq \|w - z\|^2 + \frac{2}{\beta}(f(z) - f(w)) + \frac{2}{\beta}(f(w) - f(w')) \\
&= \|w - z\|^2 + \frac{2}{\beta}(f(z) - f(w')).
\end{aligned}$$

Rearranging and applying $\sum_{i<t}$,

$$\frac{2}{\beta}\sum_{i<t}(f(w_{i+1}) - f(z)) \leq \sum_{i<t}\left(\|w_i - z\|^2 - \|w_{i+1} - z\|^2 \right)$$

The final bound follows by noting $f(w_i) \leq f(w_t)$, and since the right hand side telescopes.