Lecture 15. (Sketch.)

- Homework scores out. TA OH next week.
- Project presentations on reading day!

1. Handling approximate gradients.

Suppose we're doing gradient descent over (closed) set S with $D := \sup_{w,w' \in S} \|w - w'\| < \infty$.

- $w_0 \in S$ given.
- Thereafter, $w_i := \Pi_S(w_{i-1} - \eta_i g_i)$, where Π_S denotes orthogonal projection and g_i is an approximate (sub)gradient.

Lemma. Let $((w_i, g_i))_{i=1}^t$ given as above, along with closed convex S, convex f, and any subgradients $s_i \in \partial f(w_{i-1})$. Set $G := \max_i \max\{\|g_i\|, \|s_i\|\}$. Then for any $z \in S$ and constant $\eta_i := \eta > 0$, setting $\hat{w}_t := \sum_{i<t} w_i/t$,

$$f(\hat{w}_t) - f(z) \leq \frac{1}{t}\left(f(w_i) - f(z)\right) \leq \frac{D^2}{2\eta t} + \frac{\eta G^2}{2} + \frac{1}{t}\sum_{i \leq t} \left\langle s_i - g_i, w_{i-1} - z \right\rangle .$$

Lemma gives inequality

$$f(\hat{w}_t) - f(z) \leq \frac{1}{t}\left(f(w_i) - f(z)\right) \leq \frac{D^2}{2\eta t} + \frac{\eta G^2}{2} + \frac{1}{t}\sum_{i \leq t} \left\langle s_i - g_i, w_{i-1} - z \right\rangle .$$

Remarks.

- Set $\eta = D/(G\sqrt{t})$, all but last term is $DG/\sqrt{t}$. ($\eta_i = D/(G\sqrt{i+1})$ only changes constants.)
- Guarantee is on averaged iterate; meanwhile, smooth opt gave bounds for last iterate.
- If $s_i = g_i \in \partial f(w_{i-1})$, last term 0. Otherwise, with no further assumptions,

$$\frac{1}{t}\sum_{i \leq t} \left\langle s_i - g_i, w_{i-1} - z \right\rangle \leq \frac{1}{t}\sum_{i \leq t} 2GD \leq 2GD,$$

  which is useless.

Proof. Following a similar expand-the-square scheme to the smooth case, setting $\epsilon_i := \langle g_i - s_i, z - w_{i-1} \rangle$,

$$\begin{aligned}
\|w_i - z\|^2 = \|\Pi_S(w_{i-1} - \eta g_i) - z\|^2 &\overset{(\star)}{\leq} \|w_{i-1} - \eta g_i - z\|^2 \\
&= \|w_{i-1} - z\|^2 + 2\eta \langle g_i, z - w_{i-1} \rangle + \eta^2 \|g_i\|^2 \\
&= \|w_{i-1} - z\|^2 + 2\eta \langle s_i, z - w_{i-1} \rangle + 2\eta\epsilon_i + \eta^2 \|g_i\|^2 \\
&\leq \|w_{i-1} - z\|^2 + 2\eta(f(z) - f(w_{i-1})) + 2\eta\epsilon_i + \eta^2 G^2,
\end{aligned}$$

where $(\star)$ used Π_S nonexpansive. Rearranging,

$$2\eta(f(w_{i-1}) - f(z)) \leq \|w_{i-1} - z\|^2 - \|w_i - z\|^2 + 2\eta\epsilon_i + \eta^2 G^2.$$

Applying $(2t\eta)^{-1}\sum_{i \leq t}$ to both sides,

$$\frac{1}{t}\sum_{i<t}\left(f(w_i) - f(z)\right) \leq \frac{D^2}{2t\eta} + \frac{\eta G^2}{2t} + \frac{1}{2t}\sum_{i \leq t} \epsilon_i,$$

and the result follows by Jensen's inequality.

Remark.

In the β-smooth case, a step size $\eta \le 2/\beta$ guaranteed the objective function decreases.

Here there is no such guarantee!

2. Stochastic gradients.

We'll usually use the preceding approximate gradient lemma with stochastic gradients; then we can kill off the weird error term with averaging/concentration.

Example. Suppose $f(w) = \mathbb{E}\ell(\langle w, -XY\rangle)$, where ℓ is convex and differentiable. Then $g := -\ell'(\langle w, -xy\rangle)xy$, for (x, y) draw according to the distribution in f, satisfies $\mathbb{E}g = \nabla f(w)$: g is a *stochastic gradient* for f (it is an unbiased estimate of the gradient). We'll come back to the example in the next lecture.

Here is the main bound for stochastic gradients.

Theorem. Suppose closed convex S and convex f given, and $((w_i, g_i))_{i=1}^t$ from subgradient descent with $\mathbb{E}(g_i|w_{i-1}) \in \partial f(w_{i-1})$ and $\eta := D/(G\sqrt{t})$ with $G \ge \max_i \max\{\|g_i\|, \|\mathbb{E}(g_i|w_{i-1})\|\}$. For any $z \in S$,

$$f(\hat{w}_t) - f(z) \le \frac{1}{t}\sum_{i\le t}(f(w_i) - f(z)) \le \frac{DG}{\sqrt{t}},$$

and with probability at least $1 - \delta$ over the stochastic gradients,

$$f(\hat{w}_t) - f(z) \le \frac{1}{t}\sum_{i\le t}(f(w_i) - f(z)) \le \frac{DG\left(1 + \sqrt{8\ln(1/\delta)}\right)}{\sqrt{t}}.$$

Proof. Applying $\mathbb{E}(\cdot)$ to both sides of the earlier lemma with $s_i \in \partial f(w_{i-1})$ arbitrary,

$$\mathbb{E}\left(\frac{1}{t}\sum_{i<t}(f(w_i) - f(z))\right) \le \frac{DG}{\sqrt{t}} + \frac{1}{t}\mathbb{E}\sum_{i\le t}\langle g_i - s_i, z - w_{i-1}\rangle.$$

By the tower property of conditional expectation,

$$\begin{aligned}\mathbb{E}\langle g_i - s_i, z - w_{i-1}\rangle. &= \mathbb{E}\mathbb{E}\left(\langle g_i - s_i, z - w_{i-1}\rangle | w_{i-1}\right) \\ &= \sum_{i\le t}\mathbb{E}\left\langle \mathbb{E}\left(g_i - s_i|w_{i-1}\right), z - w_{i-1}\right\rangle = 0,\end{aligned}$$

which gives the first equality in the theorem, and establishes this error sequence is a Martingale. Consequently, by Azuma's inequality (see next slide), since $\langle g_i - s_i, z - w_i\rangle \le 2GD$, with probability at least $1 - \delta$,

$$\sum_{i\le t}\langle g_i - s_i, z - w_{i-1}\rangle \le 2DG\sqrt{2t\ln(1/\delta)},$$

which finishes the proof.

Remarks.

- The proof had to carefully use conditional expectation because w_i is a random variable that depends on all stochastic gradients coming before it.
- The proof used:
  - **Azuma-Hoeffding inequality.** Suppose $(X_i)_{i=1}^n$ is a martingale difference sequence ($\mathbb{E}(X_i | X_{<i}) = 0$) and $\mathbb{E}|X_i| \leq R$. Then with probability at least $1 - \delta$,

$$\sum_i X_i \leq R\sqrt{2t \ln(1/\delta)}.$$

  In the concentration/generalization part of the course, we will see many inequalities similar to this one.

Remarks.

- In practice, minibatches are often used. To show a benefit, we need to use a more refined martingale inequality that pays attention to variance *[maybe I'll do this in homework 2 or 3...]*.
- In this proof, we work with the averaged iterate. This is okay in the convex case, but in the nonconvex case, it's not clear how to combine parameter vectors.
- The main reason SGD "wins" is iteration time: with n data points, computing $\nabla\widehat{\mathcal{R}} = \nabla n^{-1} \sum_i \ell(-f_w(x_i)y_i)$ takes n times as long as $\nabla\ell(-f_w(x)y)$. For a batch method to be faster, it must somehow recoup this penalty of n. But while some batch solvers have a good dependence on the target error ϵ, it doesn't make sense to solve for $\epsilon \leq 1/\sqrt{n}$ in these statistical applications, therefore even a fast runtime of $n\ln(1/\epsilon) \approx n\ln(n)$ doesn't really outperform SGD's $1/\epsilon^2 \approx n$. Relatedly: problems should be *easier* with more data, not *harder*.