Lecture 17. (Sketch.)

- No class November 7.
- Project proposals are up; meetings the week before Thanksgiving.

Concentration and generalization.

Error decomposition from start of course:

$$\begin{aligned} \mathcal{R}(\hat f) - \mathcal{R}(\bar g) &= \mathcal{R}(\hat f) - \widehat{\mathcal{R}}(\hat f) && \text{generalization} \\ &+ \widehat{\mathcal{R}}(\hat f) - \widehat{\mathcal{R}}(\bar f) && \text{optimization} \\ &+ \widehat{\mathcal{R}}(\bar f) - \mathcal{R}(\bar f) && \text{concentration} \\ &= \mathcal{R}(\bar f) - \mathcal{R}(\bar g) && \text{approximation.} \end{aligned}$$

In this final statistical part of the course,

$$\mathcal{R}(f) = \mathbb{E}\ell(-f(X)Y), \qquad \widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(-f(x_i)y_i),$$

where $((x_i, y_i))_{i=1}^n$ are drawn iid from the same distribution as the $\mathbb{E}$ in $\mathcal{R}$; this provides the needed *coherence* between past and future.

In this final part of the course, we'll handle the generalization and concentration terms.

Concentration?

- Concentration of measure is the study of distributions clumping up ("concentrating") when some operations are performed on them.
- We have already seen that averages cause this behavior: we know (from hw0 and from the "approximate gradients" lecture) that $\sum_i Z_i$ lies in an interval of radius $\mathcal{O}(\sqrt{n})$ rather than $\mathcal{O}(n)$ when Z_i are iid (or a Martingale).
- $((x_i, y_i))_{i=1}^n$ are iid, thus $(Z_i)_{i=1}^n$ with $Z_i := \ell(-f(x_i)y_i)$ are iid **(for f fixed a priori)**, thus $\widehat{\mathcal{R}}(f) = n^{-1}\sum_i Z_i$ should concentrate around $\mathcal{R}(f)$!
- "f fixed a priori" is crucial and we'll return to it next lecture. (It's the difference between "generalization" and "concentration".)
- Concentration also appears in geometry; look up "isoperimetry" (**Project idea**!).

Sums of random variables.

- Classical statistical asymptotics for iid $X_1, X_2, \ldots$:

$$\frac{1}{t}\sum_{i=1}^{t} X_i \xrightarrow{\text{a.s.}} \mathbb{E}X_1 \qquad \text{(SLLN)},$$

$$\frac{1}{\sigma\sqrt{t}}\sum_{i=1}^{t} X_i \xrightarrow{\text{d}} \mathcal{N}(\mathbb{E}X_1, 1) \qquad \text{(CLT)},$$

$$\limsup_{t} \frac{1}{\sigma\sqrt{2t\ln\ln t}}\sum_{i=1}^{t} X_i \overset{\text{a.s.}}{=} 1 \qquad \text{(LiL)}.$$

- In machine learning, care about finite time! Easy cases:
  1. An easy case: an average of n $\mathcal{N}(0,1)$ random variables is $\mathcal{N}(0, 1/n)$!
  2. Bernoulli X_i: average of n is $\mathrm{Binom}(n,p)/n$ with expectation p and variance $p(1-p)/n$.

  Not just concentrated: *anti-concentrated*. (**Project idea:** learn more about this.)

2. Markov's inequality.

Let's get something for general random variables.

Theorem (Markov). For any nonnegative r.v. X and $\epsilon > 0$,

$$\Pr[X \geq \epsilon] \leq \frac{\mathbb{E}X}{\epsilon}.$$

Proof. Apply $\mathbb{E}$ to both sides of $\epsilon \mathbb{1}[X \geq \epsilon] \leq X$.

Corollary. For any nonnegative, nondecreasing $f \geq 0$ and $f(\epsilon) > 0$,

$$\Pr[X \geq \epsilon] \leq \frac{\mathbb{E}f(X)}{f(\epsilon)}.$$

Proof. Note $\Pr[X \geq \epsilon] \leq \Pr[f(X) \geq f(\epsilon)]$ and apply Markov.

Remark (concentration via Markov and moments). Define $A_n = n^{-1}\sum_i (X_i - \mathbb{E}X_1)$. For an inequality to verify concentration, the simplest thing it can report is $\Pr[|A_n| > \epsilon]$ goes to 0 as n increases.

- Markov doesn't suffice:

$$\Pr[|A_n| \geq \epsilon] \leq \frac{\mathbb{E}|A_n|}{\epsilon} = \frac{\mathbb{E}|X_1|}{\epsilon}.$$

- Second moment gives a quantity which goes to 0 with n:

$$\Pr[|A_n| \geq \epsilon] \leq \frac{\mathbb{E}A_n^2}{\epsilon^2} = \frac{\mathrm{Var}(X_1)}{n\epsilon^2}.$$

- Similarly, for even integer $p \geq 2$,

$$\Pr[|A_n| \geq \epsilon] \leq \frac{\mathbb{E}|\sum_i X_i - \mathbb{E}X_1|^p}{(n\epsilon)^p}.$$

  With some bloord, tears, and assumptions on $\max_{i \leq p} \mathbb{E}|X|^p$, get $\Pr[A_n \geq \epsilon] \leq \mathcal{O}(1)/(\epsilon\sqrt{n})^p$.

Question: what is the right dependence on n?

3. Chernoff bounds and moment generating functions.

For many problems in ML, we'll be able to mimic the behavior of Gaussians. What do Gaussians do?

- Since $\sum_i X_i/n$ is $\mathcal{N}(0, 1/n)$, and

$$\begin{aligned}
\Pr[\mathcal{N}(0,\sigma^2) \geq \epsilon] &= \frac{1}{\sigma\sqrt{2\pi}} \int_\epsilon^\infty e^{-x^2/(2\sigma^2)}\, dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty e^{-(x+\epsilon)^2/(2\sigma^2)}\, dx \\
&= \frac{e^{-\epsilon^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \int_0^\infty e^{-x^2/(2\sigma^2)} e^{-x\epsilon/\sigma^2}\, dx \\
&\leq e^{-\epsilon^2/(2\sigma^2)}/2,
\end{aligned}$$

  thus $\Pr[\sum_i X_i/n \geq \epsilon] \leq \exp(-n\epsilon^2/2)/2$!

Remark. For pth moment bounded random variables, we got RHS $(\epsilon\sqrt{n})^{-p}$; Gaussians, we got $\exp(-(\epsilon\sqrt{n})^2)$.

Let's try to get this for other random variables.

Given r.v. X, define **moment generating function** $t \mapsto \mathbb{E}\exp(tX)$.

- Not always finite! Consider $e^{tX} = \sum_{i \geq 0} \frac{(tX)^i}{i!}$ and X symmetric: need all even moments finite!

By Markov, since $r \mapsto \exp(tr)$ is nondecreasing for $t \geq 0$,

$$\Pr[X \geq \epsilon] = \inf_{t \geq 0} \Pr[\exp(tX) \geq \exp(t\epsilon)] \leq \inf_{t \geq 0} \frac{\mathbb{E}\exp(tX)}{\exp(t\epsilon)}.$$

The **Chernoff bounding technique** applies this to $A_n := \sum_i (X_i - \mathbb{E}X_i)/n$; if $(X_1, \ldots, X_n)$ iid,

$$\Pr[A_n \geq \epsilon] \leq \inf_{t \geq 0} \frac{\mathbb{E}\exp(tA_n)}{\exp(t\epsilon)} = \inf_{t \geq 0} \frac{(\mathbb{E}\exp((t/n)(X_1 - \mathbb{E}X_1)))^n}{\exp(t\epsilon)}.$$

(This is still very abstract...)

To get mileage out of this, let's consider X **subgaussian with variance proxy** σ^2:

$$\mathbb{E}\exp(tX) \le \exp(t^2\sigma^2/2).$$

Remark. Might seem abstract for now, but we'll show this holds often in ML; e.g., for boudned random variables.

Lemma. If $(X_1,\dots,X_n)$ respectively σ_i^2-subgaussian, indepedent, then $S_n := \sum_i X_i/n$ is σ^2-subgaussian with $\sigma^2 = \sum_i \sigma_i^2/n^2$.

Proof. For any t,

$$\mathbb{E}\exp(tS_n) = \prod_i \mathbb{E}\exp(tX_i/n) \le \prod_i \mathbb{E}\exp(t^2\sigma_i^2/(2n^2))$$
$$= \mathbb{E}\exp((t^2/2)\sum_i \sigma_i^2/n^2).$$

Remark. Quick sanity check: "variance proxy" is scaling with averages in the same way as a variance.

Theorem (Chernoff bound for subgaussian r.v.'s). Suppose $(X_1,\dots,X_n)$ independent and respectively σ_i^2-subgaussian. Then

$$\Pr\left[\frac{1}{n}\sum_i X_i \ge \epsilon\right] \le \exp\left(-\frac{n^2\epsilon^2}{2\sum_i \sigma_i^2}\right).$$

Proof. $S_n := \sum_i X_i/n$ is σ^2-subgaussian with $\sigma^2 = \sum_i \sigma_i^2/n^2$, so

$$\Pr[S_n \ge \epsilon] \le \inf_{t\ge 0} \mathbb{E}\exp(tZ)/\exp(t\epsilon) \le \inf_{t\ge 0}\exp\left(t^2\sigma^2/2 - t\epsilon\right)$$
$$\overset{(\star)}{=} \exp\left(\frac{\epsilon^2}{\sigma^4}\left(\frac{\sigma^2}{2}\right) - \frac{\epsilon^2}{\sigma^2}\right) = \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right),$$

where $(\star)$ took the minimum $t = \epsilon/\sigma^2 \ge 0$ to the convex quadratic.

Remarks.

- (Sanity check.) This bound agrees with our earlier Gaussian back-of-envelope calculation up to the multiplicative factor $1/2$ ($\mathcal{N}(0,\sigma^2)$ is σ^2-subgaussian).
- ("Inverting" concentration/deviation inequalities). In learning theory we often set the bound to δ and solve for ϵ, giving

$$\Pr\left[S_n \le \sqrt{\frac{2\sum_i \sigma_i^2}{n^2}\ln\left(\frac{1}{\delta}\right)}\right] \ge 1-\delta.$$

- The $\ln(1/\delta)$ in this inverted bound is important. Later we will union bound over many (functions of) r.v.'s, getting a bound with $\ln(k/\delta)$ (for k union bounds).

4. Hoeffding's inequality.

Lemma (Hoeffding). If $X \in [a,b]$ a.s., then $X - \mathbb{E}X$ is $(b-a)^2/4$-subgaussian.

Proof. Omitted.

Theorem (Hoeffding inequality). Given iid $(X_1,\dots,X_n)$ with $X_i \in [a_i,b_i]$ a.s.,

$$\Pr\left[\frac{1}{n}\sum_i (X_i - \mathbb{E}X_i) \ge \epsilon\right] \le \exp\left(-\frac{2n^2\epsilon^2}{\sum_i (b_i-a_i)^2}\right).$$

Proof. Suffices to plug the Hoeffding Lemma into the subgaussian Chernoff bound.

Remark. For classification, setting $Z_i := \mathbb{1}[f(X_i) \ne Y_i]$: with probability at least $1-\delta$,

$$\mathcal{R}_z(f) - \widehat{\mathcal{R}}_z(f) = \mathbb{E}Z_1 - \frac{1}{n}\sum_{i=1}^n Z_i \le \sqrt{\frac{1}{2n}\ln\left(\frac{1}{\delta}\right)}.$$

Remarks.

- There are many other standard Chernoff bounds
  - "Bernstein's inequality" is like Hoeffding, but has a variance term.
  - Azuma and Freedman are Hoeffding and Bernstein for Martingales; the Chernoff bounding technique is still used. (Some people use many of these names interchangeably.)
  - "McDiarmid's inequality" will be used in the next few lectures; it replaces $\sum_i X_i/n$ with any "stable" function of $(X_1, \ldots, X_n)$.
  - For Gaussian random variables, there are nice bounds.
- There are also interesting more sophisticated bounds for things like matrices (doing better than union bound on all coordinates), heavy-tailed distributions (changing the estimator), ...