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:

\[ R(\hat{f}) - R(\bar{g}) = R(\hat{f}) - R(\bar{\hat{f}}) \text{ generalization} \]
\[ + R(\bar{\hat{f}}) - R(\bar{\bar{f}}) \text{ optimization} \]
\[ + R(\bar{\bar{f}}) - R(\bar{\bar{g}}) \text{ concentration} \]
\[ = R(\bar{\bar{f}}) - R(\bar{\bar{g}}) \text{ approximation}. \]

In this final statistical part of the course,

\[ R(f) = \mathbb{E} \ell(-f(X)Y), \quad \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(-f(x_i)y_i), \]

where \(((x_i, y_i))_{i=1}^{n}\) are drawn iid from the same distribution as the \(\mathbb{E}\) in \(R\); 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 \(\sum Z_i\) lies in an interval of radius \(O(\sqrt{n})\) rather than \(O(n)\) 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 \(R(f) = n^{-1} \sum Z_i\) should concentrate around \(\hat{R}(f)\)!

- “\(f\) 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 \(X_1, X_2, \ldots:\)

\[ \frac{1}{t} \sum_{i=1}^{t} X_i \overset{a.s.}{\rightarrow} \mathbb{E}X_1 \quad \text{(SLLN)}, \]
\[ \frac{1}{\sigma \sqrt{t}} \sum_{i=1}^{t} X_i \overset{d}{\rightarrow} \mathcal{N}(\mathbb{E}X_1, 1) \quad \text{(CLT)}, \]
\[ \limsup_{t} \frac{1}{\sigma \sqrt{2t \ln \ln t}} \sum_{i=1}^{t} X_i \overset{a.s.}{\rightarrow} 1 \quad \text{(LiL)}. \]

- 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 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.

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

\[
\Pr[\mathcal{N}(0,\sigma^2) \geq \epsilon] = \frac{1}{\sigma\sqrt{2\pi}} \int_{\epsilon}^{\infty} e^{-x^2/(2\sigma^2)} \, dx
\]

\[
= \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{\infty} e^{-(x+\epsilon)^2/(2\sigma^2)} \, dx
\]

\[
= \frac{e^{-\epsilon^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \int_{0}^{\infty} e^{-x^2/(2\sigma^2)} e^{-x\epsilon/\sigma^2} \, dx
\]

\[
\leq e^{-\epsilon^2/(2\sigma^2)}/2,
\]

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

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

**Remark (concentration via Markov and moments).** Define \( A_n = n^{-1} \sum_i (X_i - \mathbb{E}X_i) \). 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 \( n \):

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

- Similarly, for even integer \( p \geq 2 \),

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

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

**Question:** what is the right dependence on \( n \)?

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

\[
A_n := \sum_i (X_i - \mathbb{E}X_i)/n; \text{ if } (X_1, \ldots, X_n) \text{ iid,}
\]

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

(This is still very abstract...)

"..."
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 $(X_1, \ldots, X_n)$ with $X_i \in [a_i, b_i]$ a.s.,

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \geq \epsilon \right] \leq \exp \left( - \frac{2n^2 \epsilon^2}{\sum_{i}(b_i - a_i)^2} \right).$$

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

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

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