Lecture 27.

- Today I talked about a second thing that is painful with Rademacher complexity: getting generalization bounds that have $1/n$ rather than $1/\sqrt{n}$. For neural nets this is not currently a huge deal (because many of the bounds are ≥ 1 so squaring doesn't help), but:
  - eventually we'll have good bounds and it will matter, in particular
  - there are situations where the *lower bound* is $1/n$, and
  - it's important to point out that maybe Rademacher complexity isn't the ultimate end-all be-all of generalization tools due to its awkwardness with these $1/n$ style bounds (which in the literature are often called "fast rates").
- In class I said our concentration bounds are by analogy to Gaussians. But for instance with classification, our distribution of errors is a Binomial, which is only well approximated by a gaussian if the bias term p is not too small relative to $1/n$.

- Here are some examples I discussed in class:
  - A direct VC argument; see chapter 12 of Devroye-Györfi-Lugosi "A probabilistic theory of pattern classification".
  - If $\mathcal{F}$ is convex and the loss ℓ is strongly convex, we can roughly get what we want, but we have to replace $\widehat{\mathcal{R}}_\ell(f)$ in the RHS with something like $(1+\mathcal{O}(1))(\widehat{\mathcal{R}}_\ell(f) - \inf_{g\in\mathcal{F}} \widehat{\mathcal{R}}_\ell(g))$. We can use Rademacher techniques for this.
  - In the case of ordinary least squares, using matrix concentration (to say the pseudoinverse and thus the ordinary least squares estimator are similar on the training set and on the distribution) we can get optimal rates; see John Duchi's lecture notes for a nice treatment of at least the lower bound.
  - We can use variance-sensitive versions of Hoeffding (namely, Bernstein's inequality) to prove classification generalization for *fixed* functions that have both a $1/\sqrt{n}$ and a $1/n$ term, where first vanishes as the classification error improves. We can also use a "relativization" argument to get this for whole function classes; see the two surveys by Boucheron-Bousquet-Lugosi.
  - See also "Talagrand's inequality for empirical processes".