Lecture 26.

- The purpose of lectures 26 and 27 is to highlight some generalization bounds for which I do not see a clear way to obtain a proof via Rademacher complexity.
- I will not type the details for these lectures, instead only give pointers.

- The source material for this lecture was compiled by Daniel Hsu, who tells me he learned it from Sanjoy Dasgupta. I can recommend these references:
  - The Devroye-Györfi-Lugosi book "A probabilistic theory of pattern classification" has a version of the proof I outlined (search for "Stone's Lemma").
  - There is also some material in the Cover-Thomas Information Theory book.

k-nearest-neighbor.

- Lecture 26 focused on k-nn, indeed just 1-nn. Here is the basic story:
  - On one hand, it seems to go against what we know about generalization: it memorizes the training set, and indeed perfectly labels any data (with distinct x_i's).
  - On the other hand, the algorithm is "simple" and has "geometric regularity". In class, we proved that $\Pr[\lim_{n\to\infty} \|X_1(X) - X\|_2 = 0] = 1$, where $X_1(X)$ is the nearest neighbor to X in a training set of size n: that is, as the training set size $\to \infty$, then for any new data point, asymptotically the nearest neighbor will be arbitrarily close.
  - From here, with some work, we can prove that 1-nn gets error which is roughly twice the optimal error amongst **all possible** classification rules.
  - The proof uses lots of conditioning tricks, and direct geometric reasoning; overall it does *not* look like our other generalization proofs!

- (Continued.)
  - The above material (and the proofs I showed in class) are asymptotic, in expectation, and only for the l_2 metric. For a treatment that is non-asymptotic (finite sample), high probability and works for a variety of metrics, see Chaudhuri-Dasgupta "Rates of Convergence for Nearest Neighbor Classification". This paper focuses on a specific smoothness property of the regression function $\eta(x) = \Pr[Y = 1|X = x]$; specifically,

$$\frac{1}{\mu(B(x,r))}\int_{B(x,r)} \eta(z)\,\mathrm{d}\mu(z) \stackrel{?}{=} \eta(x)$$

    where $B(x,r) = \{z : \|x - z\| \le r\}$? We only discussed $X_1(X) \to X$ above, but we also need to reason about η to make the proof go through. Chaudhuri-Dasgupta quantify this in order to get rates.