8 Lecture 8, October 29, 2014

8.1 Shannon's coding theorem

In order to communicate reliably, Alice and Bob are going to agree in advance on a *codebook*, a set of codewords that are fairly distant from each other (in Hamming distance), with the idea that when a corrupted codeword is received, it will still be closer to the correct codeword than to all others. In this discussion we completely ignore a key computational issue: how are the encoding and decoding maps computed efficiently? In fact it will be enough for us, for a positive result, to demonstrate existence of an encoding map $\mathcal{E} : \{0,1\}^k \to \{0,1\}^n$ and a decoding map $\mathcal{D} : \{0,1\}^n \to \{0,1\}^k$ (we'll call this an (n,k) code) with the desired properties; we won't even explicitly describe what the maps are, let alone specify how to efficiently compute them. We will call k/n the *rate* of such a code. Shannon's great achievement was to realize (and show) that you can simultaneously have positive rate and error probability tending to 0—in fact, exponentially fast.

Theorem 37 (Shannon [60]) *Let $p < 1/2$. For any $\varepsilon > 0$, for all k sufficiently large, there is an (n,k) code with rate $\geq D_2(p\|1/2) - \varepsilon$ and error probability $e^{-\Omega(k)}$ on every message. (The constant in the Ω depends on p and ε.)*

In this theorem statement, "Error" means that Bob decodes to anything different from X, and error probabilities are taken only with respect to the random bit-flips introduced by the channel.

However until almost the end of the proof, the probabilities we will consider are with respect to *both* the channel noise and to the selection of a message X uniformly at random in $\{0,1\}^k$.

Proof: Let

$$n = \frac{k}{D_2(p\|1/2) - \varepsilon} \tag{18}$$

(ignoring rounding). Let $R \in \{0,1\}^n$ denote the error string. So, with Y denoting the received message,

$$Y = \mathcal{E}(X) + R$$

with X uniform in $\{0,1\}^k$, and R consisting of iid Bernoulli rvs which are 1 with probability p. The error event is that $\mathcal{D}(\mathcal{E}(X) + R) \neq X$.

As a first try, let's design $\mathcal{E}$ by simply mapping each $X \in \{0,1\}^k$ to a uniformly, independently chosen string in $\{0,1\}^n$. (This won't be good enough for the theorem.)

To describe the decoding procedure we start with the notion of Hamming distance H. The Hamming distance $H(x,y)$ between two same-length strings over a common alphabet Σ, is the number of indices in which the strings disagree: $H(x,y) = |\{i : x_i \neq y_i\}|$ for $x, y, \in \Sigma^n$.

Define the decoding $\mathcal{D}$ to map Y to a closest codeword in Hamming distance.

In order to analyze how well this works, we pick δ sufficiently small that

$$p + \delta < 1/2 \tag{19}$$

and

$$D_2(p + \delta\|1/2) > D_2(p\|1/2) - \varepsilon/2. \tag{20}$$

If Bob decodes incorrectly then at least one of the following events has to have occurred:

Bad_1: $H(\mathcal{E}(X) + R, \mathcal{E}(X)) \geq (p + \delta)n$

Bad_2: $\exists X' \neq X : H(\mathcal{E}(X) + R, \mathcal{E}(X')) \leq (p + \delta)n$

So let's analyze the probability of failure of either of these clauses.

For Bad_1, applying Lemma 36, we have

$$\begin{aligned}\Pr\left(H(\mathcal{E}(X)+R,\mathcal{E}(X)) \geq (p+\delta)n\right) \\ = \Pr\left(H(R,\bar{0}) \geq (p+\delta)n\right) \\ \leq 2^{-D_2(p+\delta\|p)n}.\end{aligned}$$

For Bad_2, consider that for every $X' \neq X$,

$$\begin{aligned}& \Pr(H(\mathcal{E}(X)+R,\mathcal{E}(X')) \leq (p+\delta)n) \\ = \;& \Pr(H(\bar{0},\mathcal{E}(X')) \leq (p+\delta)n) && \begin{array}{l}\mathcal{E}(X') \text{ is uniform and} \\ \text{indep. of } \mathcal{E}(X)+R\end{array} \\ \leq \;& 2^{-nD_2(p+\delta\|1/2)} && \text{using Eqn 19} \\ \leq \;& 2^{-n(D_2(p\|1/2)-\varepsilon/2)} && \text{using Eqn 20}\end{aligned}$$

Now we allow for both sources of error: that R is heavier than $(p+\delta)n$, or (with a union bound) that one of the 2^k-1 incorrect messages has its codeword land within $(p+\delta)n$ of the corrupted codeword of X.

$$\begin{aligned}& \Pr(\exists X' \neq X \text{ s.t. } \mathcal{D}(\mathcal{E}(X)+R)=X') \\ \leq \;& 2^{-nD_2(p+\delta\|p)} + 2^{k-n(D_2(p\|1/2)-\varepsilon/2)} \\ = \;& 2^{-nD_2(p+\delta\|p)} + 2^{n(D_2(p\|1/2)-\varepsilon)-n(D_2(p\|1/2)-\varepsilon/2)} && \text{by Eqn 18} \\ = \;& 2^{-nD_2(p+\delta\|p)} + 2^{-n\varepsilon/2} \\ \leq \;& 2^{1-cn} && \text{where } c := \min\{D_2(p+\delta\|p), \varepsilon/2\}\end{aligned}$$

Another way of stating this conclusion is by conditioning on the choice of $\mathcal{E}$.

$$\begin{aligned}2^{1-cn} \;&\geq\; \Pr_{\mathcal{E},X,R}(\text{Error}) \\ &=\; E_{\mathcal{E}}(\Pr_{X,R}(\text{Error}|\mathcal{E})) \quad \text{Note the probability is itself an rv}\end{aligned}$$

So there exists some *specific* code $\mathcal{E}^*$ (and corresponding decoding map $\mathcal{D}^*$) achieving $\Pr_{X,R}(\text{Error}|cE^*) \leq 2^{1-cn}$.

There is just one remaining annoyance. The error probability is bounded *on average* over X's, but some messages might suffer bad error rates. There is an easy fix for this. By the Markov inequality, at most half the messages can have error probability greater than twice the given bound. So just don't use those messages. That is to say, if you want to send k bits, design a code as above for $k+1$ bits, then map the k-bit-strings to the good half of the messages. The asymptotic rate is unaffected by this trick, and the error probability $\Pr_R(\text{Error})$ applies to *all* X and is only twice the above bound. (Thus, the error exponent is unaffected.) To be explicit, using $\mathcal{E}^*$ designed for $k+1$ bits and with $n' = \frac{k+1}{D_2(p\|1/2)-\varepsilon}$, we have

$$\begin{aligned}2^{1-cn'} \;&\geq\; \Pr_{X,R}(\text{Error}|\mathcal{E}^*) \\ &=\; E_X \Pr_R(\text{Error}|\mathcal{E}^*, X)\end{aligned}$$

In our actual code we use only the 2^k messages X achieving the smallest values of $\Pr_R(\text{Error}|\mathcal{E}^*, X)$. By Markov's inequality, for these messages X,

$$\Pr_R(\text{Error}|\mathcal{E}^*, X) \leq 2^{2-cn'} \leq 2^{2-cn}$$

Thus no matter what message Alice sends, Bob's probability of error is exponentially small. □

8.2 Central limit theorem

As I mentioned earlier in the course, there are two basic ways in which we express concentration of measure: large deviation bounds, and the central limit theorem. Roughly speaking the former is a weaker conclusion (only upper tail bounds) from weaker assumptions (we don't need full independence—we'll talk about this soon).

The proof of the basic CLT is not hard but relies on a little Fourier analysis and would take us too far out of our way this lecture, so I will just quote it:

Let μ be a probability distribution on $\mathbb{R}$, i.e., for X distributed as μ, $S \subseteq \mathbb{R}$, $\Pr(X \in S) = \mu(S)$.

For $X_1, \ldots, X_n$ sampled independently from μ set $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Theorem 38 *Suppose that μ possesses both first and second moments:*

$$\theta = E[X] = \int x \, d\mu \quad \textit{mean}$$

$$\sigma = E\left[(X-\theta)^2\right] = \int (x-\theta)^2 \, d\mu \quad \textit{variance}$$

Then for all $a < b$,

$$\lim_n \Pr(a\sqrt{\sigma/n} < \overline{X} - \theta < b\sqrt{\sigma/n}) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} \, dt.$$