3.3 Lecture 15 (5/Nov): Application of large deviation bounds: Shannon's coding theorem. Central limit theorem

3.3.1 Shannon's block coding theorem. A probabilistic existence argument.

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 achievement was to realize (and show) that you can simultaneously have positive rate and error probability tending to 0—in fact, exponentially fast.

Theorem 49 (Shannon [90]) *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.

Proof: Let

$$n = \frac{k}{D_2(p\|1/2) - \varepsilon} \tag{3.2}$$

(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.)

So (for now) when we speak of error probability, we have two sources of randomness: channel noise R, and code design $\mathcal{E}$.

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.

For most of the remainder of the proof (in particular until after the lemma), we fix a particular message X, and analyze the probability that it is decoded incorrectly.

In order to speak separately about the two sources of error, we define M_X to be the rv (which is a function of $\mathcal{E}$) $M_X = \Pr_R(\text{Error on } X|\mathcal{E})$. So for any $\mathcal{E}$, $0 \leq M_X \leq 1$.

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

$$p + \delta < 1/2 \tag{3.3}$$

and

$$D_2(p + \delta\|1/2) > D_2(p\|1/2) - \varepsilon/2. \tag{3.4}$$

Note that if both

1. $H(\mathcal{E}(X)+R, \mathcal{E}(X)) < (p+\delta)n$ ("channel noise is low"), and
2. $\forall X' \neq X : H(\mathcal{E}(X)+R, \mathcal{E}(X')) > (p+\delta)n$ ("code design is good for X, R")

then Bob will decode correctly.

The contrapositive is that if Bob decodes X 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: $\exists X' \neq X : H(\mathcal{E}(X)+R, \mathcal{E}(X')) \leq (p+\delta)n$

Lemma 50 *$\exists c > 0$ s.t. $E_{\mathcal{E}}(M_X) < 2^{1-cn}$*

Proof: Specifically we show this for $c = \min\{D_2(p+\delta\|p), \varepsilon/2\}$. In what follows when we write a bound on $\Pr_W(\ldots)$ we mean that "conditional on anything else, the randomness in W is enough to ensure the bound".

$$\begin{aligned}
E_{\mathcal{E}}(M_X) &\leq \Pr_R(\text{Bad}_1) + \sum_{X' \neq X} \Pr_{\mathcal{E}(X')}(\text{Bad}_2) \\
&\leq \Pr(H(\vec{0}, R) \geq (p+\delta)n) + 2^k \Pr(H(\vec{0}, \mathcal{E}(X')) \leq (p+\delta)n) \\
&\leq 2^{-nD_2(p+\delta\|p)} + 2^{k-nD_2(p+\delta\|1/2)} \\
&= 2^{-nD_2(p+\delta\|p)} + 2^{n(D_2(p\|1/2)-\varepsilon-D_2(p+\delta\|1/2))} \quad \text{substituting value of } k \\
&\leq 2^{-nD_2(p+\delta\|p)} + 2^{-\varepsilon n/2} \quad \text{using inequality (3.4)} \\
&\leq 2^{1-cn} \quad \text{using value of } c
\end{aligned}$$

□

All of the above analysis treated an arbitrary but fixed message X. We showed that, picking the code at random, the expected value of $M_X = \Pr_R(\text{Error on } X|\mathcal{E})$ is small.

Let Z be the rv which is the *fraction* of X's for which $M_X \leq 2E(M_X)$. By the Markov inequality, $\exists \mathcal{E}$ s.t. $Z \geq 1/2$. Let $\mathcal{E}^*$ be a specific such code.

$\mathcal{E}^*$ works well for most messages X, but this isn't quite what we want—we want M_X to be small for *all* messages X.

There is a simple solution. Choose a code $\mathcal{E}^*$ as above for $k+1$ bits, then map the k-bit messages to the good half of the messages. Note that removal of some codewords from $\mathcal{E}^*$ can only decrease any M_X. (Assuming we still use closest-codeword decoding.)

So now the bound $\Pr_R(\text{Error on } X) \leq 2E(M_X) \leq 2^{2-cn}$ applies to *all* X.

The asymptotic rate is unaffected by this trick; the error exponent is also 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 for all $X \in \{0,1\}^k$

$$\Pr_R(\text{Error on } X) \leq 2^{2-cn}$$

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

3.3.2 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, measurable $S \subseteq \mathbb{R}$, $\Pr(X \in S) = \mu(S)$. For $X_1, \ldots, X_n$ sampled independently from μ set $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Theorem 51 *Suppose that μ possesses both first and second moments:*

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

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

Then for all $a < b$,

$$\lim_n \Pr\left(\frac{a\sigma}{\sqrt{n}} < \overline{X} - \theta < \frac{b\sigma}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} \, dt. \tag{3.5}$$

The form of convergence to the normal distribution in 3.5 is called *convergence in distribution* or *convergence in law.* For a proof of the CLT see [17] Sec. 27 or for a more accessible proof for the case that the X_i are bounded, see [3] Sec. 3.8.