Lecture 8*

Agenda for the lecture

- Heuristic derivation of Shannon’s channel coding theorem and introduction of mutual information

8.1 Maximum rate transmission over a channel

In the previous lecture, we had setup the stage for the channel coding problem. Our goal was to choose $M$ sequences $\{x(1),...,x(M)\}$, each consisting of $n$ symbols from the input alphabet of the channel, such that the receiver observing the output $Y_1,...,Y_n$ can determine reliably which message was sent. The design of receiver can be addressed using $M$-ary hypothesis testing. But how should we choose $\{x(1),...,x(M)\}$ so that $M$ is as large as possible?

To answer this question, we first choose a reasonable test for the receiver, and then choose the inputs which will keep the error for this receiver under control. We assume that each message is sent with equal probability, i.e., we are in the Bayesian framework with a uniform prior. Of course, we know that the optimal decision rule (with minimum prob. of average error) is the ML rule. However, we will choose a (perhaps) suboptimal

*©Himanshu Tyagi. Feel free to use with acknowledgement. 
decision rule, but one which will give a better insight into the problem. We alluded to this
*hard-thresholding* based decision rule in the previous lecture. Now, we define it formally.

Denote by $P_m$ the probability distribution $W^n(\cdot|x(m))$ on $Y^n$, $1 \leq m \leq M$, i.e.,

$$P_m(y) = \prod_{i=1}^{n} W(y_i|x_{im}),$$

and by $P_0$ the average distribution

$$P_0(y) = \frac{1}{M} \sum_{m=1}^{M} P_m(y).$$

Our decision rule has $M$ detectors, with detector $m$ resolving the binary hypothesis testing
problem $P_m$ vs $P_0$, namely if the received signal corresponds to $m$ or not. Each detector
outputs a bit $B_m$, with $B_m = 1$ representing the belief that message $m$ was sent. Our
decision rule is simply to declare a message $m$ if the vector $(B_1, ..., B_M)$ has a 1 only
at the $m$th location; if no such location is found, we declare an error$^1$.

To analyse the error for this decision rule, denote by $A_m$ the set of $Y^n$ for which the
$m$th detector declares 1. An error can occur if and only if

1. Message $m$ was sent and the $m$th detector does not declare a 1; or
2. A detector $m' \neq m$ declares a 1.

Denoting the first error event as $E_1$ and the second as $E_2$, the probability of error $P_e$ is
bounded above as

$$P_e \leq P(E_1) + P(E_2).$$

$^1$Formally, this is cheating because we never said we are allowed to declare errors in $M$-ary hypothesis
testing. But this more flexible decision rule is practically more relevant since we can ask for a re-transmission
if an error is detected.
The first term is equal to

$$\mathbb{P}(\mathcal{E}_1) = \frac{1}{M} \sum_{m=1}^{M} P_m(A_m^c).$$

(1)

Note that each term $P_m(A_m^c)$ is equal to the probability of false alarm for the $m$th detector.

As a conservative strategy, we might require each detector to use a rule with probability of FA less than $\epsilon$, where $\epsilon$ is our required probability of error. However, the actual requirement by (1) is milder and requires only to the average of probabilities of FA to be less than $\epsilon$.

For the second error term, denoting by $\mathcal{E}_{mm'}$ the error event where detector $m'$ declares a 1 when $m$ is sent, by union bound we get

$$\mathbb{P}(\mathcal{E}_2) = \frac{1}{M} \sum_{m=1}^{M} P_m(\mathcal{E}_2) \leq \frac{1}{M} \sum_{m=1}^{M} \sum_{m' \neq m} P_m(\mathcal{E}_{mm'}).$$

But the event $\mathcal{E}_{mm'}$ happens if and only if $Y^n$ is in $A_m'$, the acceptance region for detector $m'$. Thus,

$$\mathbb{P}(\mathcal{E}_2) \leq \frac{1}{M} \sum_{m=1}^{M} \sum_{m' \neq m} P_m(A_m')$$

$$\leq \frac{1}{M} \sum_{m=1}^{M} \sum_{m'} P_m(A_{m'})$$

$$= \sum_{m'} \frac{1}{M} \sum_{m=1}^{M} P_m(A_{m'})$$

$$= \sum_{m'} P_0(A_{m'}),$$

where the previous two equalities follow upon switching the two summations and using the definition of $P_0$. Once again, we can recognize each term in the summation as the probability of MD for detector $m'$, and can seek detectors which minimize these probabilities individually. But the actual requirement is simply to minimize the average of these
probabilities, since the inequality above can be re-expressed as

\[ \Pr(\mathcal{E}_2) \leq M \cdot \left( \frac{1}{M} \sum_{m' = 1}^{M} \Pr_0(A_{m'}) \right). \]  

(2)

We now make an interesting observation about our requirements (1) and (2) – it seems that our combined decision rule is itself a binary hypothesis test! Specifically, it is a BHT for distributions \( P \) and \( Q \) on \( \{1, \ldots, M\} \times \mathcal{Y}^n \) given by

\[ P(x, y) = \frac{1}{M} P_m(y), \quad Q(m, y) = \frac{1}{M} P_0(y). \]

The test is given by

\[ A = \{(m, y) : y \in A_m, 1 \leq m \leq M\}. \]

The bound in (1) corresponds to the probability of FA \( \Pr(A^c) \) since

\[ \Pr(A) = \sum_{m = 1}^{M} \sum_{y \in A_m} \Pr(m, y) \]
\[ = \frac{1}{M} \sum_{m = 1}^{M} \sum_{y \in A_m} P_m(y) \]
\[ = \frac{1}{M} \sum_{m = 1}^{M} P_m(A_m). \]

Furthermore, the bound in (2) corresponds to the probability of MD \( \Pr(A) \). Thus, we can design the required test by using a BHT for \( P \) vs \( Q \) with prob. of FA less than \( \epsilon/2 \) and probability of MD as small as possible. The overall bound we get for error is

\[ \Pr_e \leq \frac{\epsilon}{2} + M \beta_{\frac{1}{2}}(P, Q). \]
This will be smaller than \( \epsilon \), as long as

\[
M \leq \frac{\epsilon}{2\beta_2(P, Q)}.
\]

This test is heuristically appealing. Note that \( P \) is the joint distribution that the channel induces between the input and the output, and thereby between the message and the output, and \( Q \) is an independent distribution between the message and the output. The corresponding BHT is popularly known as the independence testing problem in statistics and machine learning. Therefore, detector \( m \) is simply testing if the received output is correlated with \( m \) or is it simply noise generated independently of the message.

The final bound above is the essence of information theory (or Shannon theory, as it is sometimes called). It quantizes the number of messages we can send over the channel with required reliability. Our goal will be to choose \( \{x(1), \ldots, x(M)\} \) so that this is as large as possible. But how large can it be? Using the bound above, together with the Little-Big lemmas for BHT, we can get an estimate for it. But that estimate itself will depend on \( P_1, \ldots, P_M \), and therefore also on the chosen \( \{x(1), \ldots, x(M)\} \). Can we obtain an estimate which depends only on the channel \((X, W, Y)\)? Later in the course, we shall rigorously derive this estimate. For now, we have to be satisfied with only a heuristic argument.

A wishful thinking bound: One difficulty with \( \beta_2(P, Q) \) is that the distributions \( P \) and \( Q \) are not independent across \( n \). Note that the distributions could have been independent if the distribution on \( X^n \) was independent, since the channel is memoryless. However, the distribution on \( X^n \) comes from that on messages and is uniform. Suppose that we could have obtained an independent distribution, i.e., suppose the uniform distribution on \( \{x(1), \ldots, x(M)\} \) can be replaced with an i.i.d. distribution \( P_{X^n} \). Then, since the channel is memoryless,

\[
P(x, y) = \prod_{i=1}^n P_X(x_i) W(y_i|x_i) = \prod_{i=1}^n P_{XY}(x_i, y_i),
\]

that is, the distribution \( P \) is i.i.d. \( P_{XY} \). Similarly, the distribution \( Q \) is i.i.d. \( P_X \times P_Y \).
with $X$ and $Y$ independent. Therefore, using Stein’s Lemma, for $n$ sufficiently large

$$
\beta_\frac{1}{2}(P, Q) \leq 2^{-nD(P_{XY} \| P_X \times P_Y)}.
$$

Therefore, we can send $M$ messages over the channel if

$$
M2^{-nD(P_{XY} \| P_X \times P_Y)} \leq \frac{\epsilon}{2},
$$

which is the same as

$$
\frac{1}{n} \log M \leq D(P_{XY} \| P_X \times P_Y) - \log \frac{2}{\epsilon}.
$$

We shall see later that, in fact, this replacement with independent $P_X^n$ can be done if the set $\{x(1), ..., x(M)\}$ is carefully chosen! Furthermore, we can have any distribution $P_X$ in this calculation. The expression

$$
D(P_{XY} \| P_X \times P_Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}
$$

is called the mutual information between $X$ and $Y$. It will be denoted by $I(X \land Y)$ or $I(P_X; W)$ depending on the context. It is one of the fundamental quantities of information. Optimizing over the distribution $P_X$, we have shown that rates upto $\max_{P_X} I(P_Z; W)$ can be attained using sufficiently large $n$. Shannon showed this is the best we can do – the maximum rate, the capacity $C(W)$ of the channel $(\mathcal{X}, W, \mathcal{Y})$ is equal to $\max_{P_X} I(P_Z; W)$. This result is called Shannon’s channel coding theorem. We shall provide several proofs of this result, formalizing the heuristic arguments presented here.