Eleventh Lecture
November 27, 2023

Channel coding cont.

We consider only *discrete memoryless channels (DMC)*, so the input and output alphabets are finite and the behaviour of the channel is described by the same matrix W at every channel use. (In particular, the probabilities described by this matrix do not depend on what happened in the past, e.g., what input letters were sent previously and what output letters they resulted in.)
Channel matrix is a stochastic matrix, where rows belong to input letters, columns belong to output letters. $W_{i,j} = W(v_j|u_i)$, which is the probability of receiving v_j when u_i was sent.

The input and output alphabets are denoted by $\mathcal{U}, \mathcal{V}$, respectively.

Binary symmetric channel ($BSC(p)$): $\mathcal{U} = \mathcal{V} = \{0, 1\}$, $W(1|1) = W(0|0) = 1-p$, $W(1|0) = W(0|1) = p$.

Goal: Communicating reliably and efficiently.
Reliably means: with small probability of error.
Efficiently means: with as few channel use as possible.

Def. *Code*: is a set $\mathcal{C} = \{\boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(M)}\} \subset \mathcal{U}^n$, the elements of $\mathcal{C}$ are the codewords. The relevance of $\mathcal{C}$ will be its size $M := |\mathcal{C}|$. The codeword length is n.
The *encoder* is an invertible function $f : \mathcal{Y}^k \to \mathcal{C}$, i.e. it maps from the set of possible messages to the set of codewords.
The *decoder* consists of two parts. First a function $g : \mathcal{V}^n \to \mathcal{C}$ observing the output of the channel decides on which codeword have been sent. Then with the inverse function $f^{-1} : \mathcal{C} \to \mathcal{Y}^k$ the message is decoded.

The conditional probability of error if the mth message $\boldsymbol{y}^{(m)}$, that is the codeword $\boldsymbol{c}^{(m)}$, was sent is

$$P_{e,m} = \sum_{v:g(v)\neq c^{(m)}} Prob(\boldsymbol{v} \text{ was received}|\boldsymbol{c}^{(m)} \text{ was sent}) = \sum_{v:g(v)\neq c^{(m)}} \prod_{i=1}^{n} W(v_i|c_i^{(m)}),$$

where v_i and $c_i^{(m)}$ denote the ith character in the sequences $\boldsymbol{v}$ and $\boldsymbol{c}^{(m)}$, respectively.

The error probability

$$P_e = \sum_{m=1}^{M} Prob(\boldsymbol{y}^{(m)} \text{ was sent}) \cdot P_{e,m}$$

We want small error independently of the probability distribution on the messages. So we define the average error probability that is the average of the $P_{e,m}$ values on the M messages:

$$\bar{P_e} = \frac{1}{M} \sum_{m=1}^{M} P_{e,m}$$

We might as well be interested in the maximal probability of error, which is also independent of the distribution on the messages:

$$P_{e,max} = \max_{1\leq m\leq M} P_{e,m}$$

Clearly, $P_e \leq P_{e,max}$.

The efficiency of the code is measured by its rate:

$$R = \frac{\log_2 M}{n}$$

Shannon's Channel Coding Theorem, one of the most fundamental results in information theory, says that discrete memoryless channels have a characteristic value, their *capacity*, with the property that one can communicate reliably with any rate below it, and one cannot, above it. Here "reliably" means "with arbitrary small probability of error".

First we define the capacity:

Def. The *capacity* C_W of a discrete memoryless channel given by its matrix W is

$$C_W := \max I(U, V),$$

where the maximization is over all joint distributions of the pair of random variables (U, V) that satisfy that the conditional probability of V given U is what is prescribed by W.

The above expression can be rewritten as

$$C_W = \max \left\{ \sum_{u \in \mathcal{U}, v \in \mathcal{V}} p(u, v) \log \frac{p(u, v)}{p(u)p(v)} \right\}$$

$$= \max \left\{ \sum_{u \in \mathcal{U}, v \in \mathcal{V}} p(u)p(v|u) \log \frac{p(v|u)}{\sum_{u' \in \mathcal{U}} p(u')p(v|u')} \right\}$$

The advantage of the last expression is that it shows very clearly that when maximizing $I(U, V)$ what we can vary is the distribution of U, that is the input distribution. (All other values in the last expression are conditional probabilities given by the channel matrix W.)

Capacity of the binary symmetric channel
$I(U, V) = H(V) - H(V|U)$ and it follows from the channel characteristics that $H(V|U = 0) = H(V|U = 1) = h(p)$, so $H(V|U) = h(p)$ irrespective of the distribution of U. So $I(U, V) = H(V) - h(p) \leq \log 2 - h(p) = 1 - h(p)$. Observing that if we let U have uniform distribution, then V will also have uniform distribution (that results in $H(V) = 1$), we conclude that this upper bound can be achieved. Thus the *capacity of the binary symmetric channel* is $1 - h(p)$.

Home-work:
1. Find the capacity of the binary erasure channel.
2. (Mod 11 channel): We have a channel with input and output alphabet $\mathcal{U} = \mathcal{V} = \{0, 1, \ldots, 10\}$. When input i is sent the output is one of $i + 1, i + 2$, and $i + 3$ (where addition is meant modulo 11), each with probability 1/3. Determine the capacity of this channel.

Now we state the Channel Coding Theorem and its converse:

Theorem 1 *(Converse to the Channel Coding Theorem) Given a channel with capacity C, for any code with rate R and codeword length n*

$$P_{e,max} \geq \bar{P}_e \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

Theorem 2 *(Channel Coding Theorem) For any $\varepsilon > 0$ and $r < C$ there exists an $N(r, \varepsilon)$ such that if $n \geq N(r, \varepsilon)$ then there exists a code $\mathcal{C} = \{\boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(M)}\} \subset \mathcal{U}^n$ with $P_{e,max} < \varepsilon$ and $R > r$*

In short one can say that all rates below capacity are achievable with an arbitrarily small error probability, and this is not true for any rate above capacity. We must have $R \leq C$ in order to achieve arbitrarily small error probability.

In order to prove the Converse to the Channel Coding Theorem, we need a theorem that gives lower bound to the error probability when we wish to estimate the value of an (inobservable) random variable from the value of another random variable (that can be observed).

Theorem 3 *(Fano's inequality) Assume that we would like to estimate Y, which can take n different values, and the random variable that we can observe is Z. Let denote our guess by $\hat{Y} = g(Z)$. Then for the error probability $P_e = Prob(\hat{Y} \neq Y)$*

$$H(Y|Z) \leq h(P_e) + P_e \log(n-1)$$

Remark: Since $h(P_e) \leq 1$, we can get the following lower bound for the error probability:

$$P_e \geq \frac{H(Y|Z) - 1}{\log(n-1)}$$

Proof. Let E be the random variable defined by

$$E \in \{0,1\}, E = 1 \Leftrightarrow \hat{Y} \neq Y,$$

i.e., the indicator variable of the error. Clearly, E is determined by the pair (Y, Z), so $H(E|Y,Z) = 0$

For $H(Y|E,Z)$ we can get from definition that

$$H(Y|E,Z) = Prob(E=0) \cdot H(Y|E=0,Z) + Prob(E=1) \cdot H(Y|E=1,Z)$$

Obviously $H(Y|E=0,Z) = 0$, since if there is no error, then $Y = g(Z)$.
If an error occurs then $Y \neq g(Z)$, thus Y can take one of the remaining $n-1$ possible values, so $H(Y|E=1,Z) \leq \log(n-1)$. And $Prob(E=1) = P_e$
Then using the Chain rule to expand $H(E,Y|Z)$ in two different ways, we can write

$$H(Y|Z) = H(Y|Z) + H(E|Y,Z) = H(E,Y|Z) = H(E|Z) + H(Y|E,Z)$$
$$\leq H(E) + H(Y|E,Z) = h(P_e) + H(Y|E,Z) \leq h(P_e) + P_e \log(n-1)$$

□

Proof of the converse of the channel coding theorem.

We will use the following lemma.

Lemma 1 *Let V^n be the output of a discrete memoryless channel with capacity C resulting from the input U^n. Then*

$$I(U^n, V^n) \leq nC.$$

Proof.

$$I(U^n, V^n) = H(V^n) - H(V^n|U^n) = H(V^n) - \sum_{i=1}^{n} H(V_i|V_1, ..., V_{i-1}, U^n)$$

$$= H(V^n) - \sum_{i=1}^{n} H(V_i|U_i) \leq \sum_{i=1}^{n} H(V_i) - \sum_{i=1}^{n} H(V_i|U_i) = \sum_{i=1}^{n} I(U_i, V_i) \leq nC.$$

Here the second equality follows from the Chain rule, and the third equality used the discrete memoryless property of the channel, which implies that V_i depends only on U_i among $V_1, ..., V_{i-1}, U_1, \ldots U_n$ and thus the used equality of conditional entropies. (The other relations should be clear: the first and fourth equality follows from the definition of mutual information, the first "$\leq$" is a consequence of the standard property of the entropy of joint distributions, while the final inequality follows from the definition of channel capacity.) □

If $\boldsymbol{Y}$ denotes the message that was sent and $\hat{\boldsymbol{Y}} = f^{-1}(g(V^n))$ is the decoded message (the estimate), then since f is invertible

$$I(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = I(\boldsymbol{Y}, f^{-1}(g(V^n))) = I(f(\boldsymbol{Y}), f^{-1}(g(V^n))) \leq I(f(\boldsymbol{Y}), V^n) \leq nC$$

follows from the properties of mutual information and the lemma above.

Let $\boldsymbol{Y}$ be uniformly distributed, i.e. the probability of message $\boldsymbol{y}_m$ is $\frac{1}{M}$ for all m. Then $P_e = \bar{P}_e$ and $H(\boldsymbol{Y}) = \log M$. Thus from Fano's inequality

$$h(\bar{P}_e) + \bar{P}_e \log(M-1) \geq H(\boldsymbol{Y}|\hat{\boldsymbol{Y}}) = H(\boldsymbol{Y}) - I(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \log M - I(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) \geq \log M - nC$$

Obviously $\log(M-1) < \log M$, so using that $h(P_e) \leq 1$ and that by definition $\log M = nR$

$$1 + \bar{P}_e \log M \geq \log M - nC \Rightarrow 1 + \bar{P}_e nR \geq nR - nC \Rightarrow \bar{P}_e \geq 1 - \frac{C}{R} - \frac{1}{nR}$$ □