Noisy Channel Coding Theorem (Shannon, 1948)

Here we present two versions of the Noisy Channel Coding Theorem, the first for the specific case of the binary symmetric channel and the second an extension to the general discrete memoryless channel.

Noisy Channel Coding Theorem for the BSC.

Theorem. Given a Binary symmetric Channel with capacity \( C \) and any real number \( R \) such that \( 0 < R < C \), then if \( (M_n : 1 \leq N < \infty) \) is any sequence of integers satisfying

\[
1 \leq M_n \leq 2^{RN},
\]

and \( \epsilon > 0 \) is any positive real number, there exists a sequence of codes \( (C_N : 1 \leq N < \infty) \) and an integer \( N_0(\epsilon) \), with \( C_N \) having \( M_N \) codewords of length \( N \), and with maximum error probability

\[
\max_{1 \leq m \leq M_N} \{P_{e,m}\} \leq \epsilon,
\]

for all \( N \geq N_0(\epsilon) \).

Proof: The main ideas behind the proof are:

(a) Select codewords for each of the \( M_N \) messages at random;
(b) Decode the messages using (an essentially) maximum likelihood decoding rule;
(c) Upper bound the error the code and decoding rule rule and show that, as \( N \rightarrow \infty \), if \( M_N \leq 2^{RN} \), \( P_{e,m} \rightarrow 0 \), \( 1 \leq m \leq M_N \).

Assume the BSC has crossover probability \( p \). Select the \( M_N \) binary codewords at random, independently of each other; for each codeword select each of the \( N \) binary digits at random, independently of each other, with \( \Pr(0) = \Pr(1) = 1/2 \). Decode by the following method: Fix \( r > 0 \) and let the \( r \)-sphere \( S_r(y) \) about \( y \) be defined as

\[
S_r(y) = \{ z : z \in V_N, d(y, z) \leq r \},
\]

where \( V_N \) is the set of all binary vectors of length \( N \) and \( d(y, z) \) is the Hamming distance between \( y \) and \( z \). If \( y \) is the received vector, decode \( y \) as \( x_j \) if \( x_j \) is the unique codeword in \( S_r(y) \). Otherwise, intentionally make an error by decoding \( y \) as some arbitrary codeword (i.e., \( x_1 \)).

Let \( \sim \), denote the received vector when a particular \( x \) is transmitted. Define

\[
E = \text{the event that an error is made}.
\]
That is, \( E \) denotes the event that \( y \) is decoded as some codeword other than \( x \). Then \( E \) occurs if either

(a) \( d(x, y) > r \), or

(b) \( d(x, y) \leq r \) and \( d(x', y) \leq r \) for some other codeword \( x' \neq x \).

Call these two events event A and event B respectively. It follows that

\[
\Pr(E) = \Pr(A \cup B) \leq \Pr(A) + \Pr(B),
\]

so it is sufficient to upper-bound \( \Pr(A) \) and \( \Pr(B) \) in order to upper-bound \( \Pr(E) \).

Consider the event B. B occurs if both

(i) Not more than \( r \) errors occur in transmission, and

(ii) Some codeword other than \( x \) is within distance \( r \) of \( y \).

Call these two events \( B_1 \) and \( B_2 \), respectively. Note that

\[
\Pr(B) \leq \Pr(B_2). \tag{1}
\]

Consider the event \( B_2 \). Since the codewords are chosen randomly, the probability that some \( x_i \) is within distance \( r \) of \( y \) is

\[
\sum_{k=0}^{r} \binom{N}{k} \frac{r}{2^N}. \tag{2}
\]

Hence, the probability that at least one of the \( M - 1 \) codewords not equal to \( x \) belongs to \( S_r(y) \) can be upper bounded by

\[
\Pr(B_2) \leq \frac{(M - 1)}{2^N} \sum_{k=0}^{r} \binom{N}{k}. \tag{3}
\]

For \( \epsilon > 0 \), we take \( \tilde{r} = \lceil Np + N\epsilon \rceil \). Then using Eqs. (1), (2), and (3), and the Tail Inequality, which states that for \( 0 < \lambda < 1/2 \),

\[
\sum_{k=0}^{\lfloor \lambda n \rfloor} \binom{n}{k} \leq 2^{n\mathcal{H}_2(\lambda)},
\]

we get

\[
\Pr(B) \leq \frac{M - 1}{2^N} 2^{N\mathcal{H}_2(p + \epsilon)} < M N 2^{-N(1 - \mathcal{H}_2(p + \epsilon))}.
\]

Noting that the capacity \( C(p + \epsilon) \) of a BSC with crossover probability \( p + \epsilon \) is given by \( 1 - \mathcal{H}_2(p + \epsilon) \), we can rewrite this as

\[
\Pr(B) \leq M 2^{-NC(p + \epsilon)}. \tag{4}
\]

Next consider the event A. Here we have the situation in which \( d(x, y) > r \)—that more than \( r \) errors have occurred in transmission of the \( N \) binary digits. Let \( U \) be defined
as the random variable giving the number of errors that have occurred in transmitting $x$ across the channel. Then $U$ is binomially distributed ($U \sim b(N; p)$), and we have

$$\Pr(A) = \Pr(U > r) = \Pr(u > Np + N\epsilon)$$

$$\leq \Pr(|U - Np| > N\epsilon)$$

$$\leq \frac{\text{var}(U)}{N^2\epsilon^2} \quad \text{(Chebyshev’s Inequality)} \cdot$$

Thus

$$\Pr(A) \leq \frac{p(1-p)}{N\epsilon^2}. \quad (5)$$

From Eqs. (4) and (5), it follows that the probability of average probability of error, $\Pr(E)$, is upper bounded by

$$\Pr(E) \leq \frac{p(1-p)}{N\epsilon^2} + M_N 2^{-N(C(p + \epsilon)} \quad (6)$$

Since $\epsilon > 0$, $\Pr(E)$ can be made arbitrarily small if $N$ is taken to be sufficiently large, provided that $M_N$ as a function of $N$ grows at a rate less than $2^{C(p)}$. This is equivalent to saying that

$$M_N \leq 2^{NR} \leq 2^{NC(p)}$$

or $R < C$.

We have bounded the average error probability rather than the maximum error probability as specified in the theorem statement. In order to complete the proof, we need to show the existence of codes $C_N$ with $M_N$ codewords, where $M_N \leq 2^{NR}$, and with maximum error probability less than or equal to $\epsilon$. To do this, take $\epsilon' = \epsilon/2$ and $M' = 2M_N$. Since $M_N \leq 2^{NR}$ and $R < C$, there exists a real number $R'$ with $R < R' < C$, and $N_\epsilon'$ such that, for $N \geq N_\epsilon'$,

$$M' \leq 2^{NR'},$$

and there exists a sequence of codes $C'_{N'}$ such that $C'_{N'}$ has $M'_{N'}$ codewords and average error probability less than or equal to $\epsilon'$ for $N \geq N_\epsilon'$.

If $x_1, \ldots, x_{M'}$ are the codewords of $C'_{N'}$, then

$$\sum_{i=1}^{M'} \Pr(E|x_i) \leq \epsilon'M'_{N'},$$

So at least half of these codewords $x_i$ must satisfy

$$\Pr(E|x_i) \leq 2\epsilon' = \epsilon.$$

Let $C_N$ be any $M_N$ of these codewords. Then we have our required code with maximum error probability less than or equal to $\epsilon$. 

- 3 -
Noisy Channel Coding Theorem for the DMC.

We now look at generalizing the Noisy Channel Coding Theorem for the general Discrete Memoryless Channel (DMC). Consider a DMC with input alphabet $\{x_1, \ldots, x_K\}$, output alphabet $\{y_1, \ldots, y_J\}$, channel transition matrix $[p(y_j|x_k)]$, and capacity $C$. Let $P^*_x$ be the input distribution that achieves capacity, and without loss of generality, assume $p^*_x(x_k) \neq 0$, $k = 1, \ldots, K$. Let the output distribution corresponding to $P^*_x$ be $P^*_y$. A statement of the noisy channel coding theorem for this DMC is as follows:

Theorem. If an information source generates information at a rate $R < C$, then there exist codes such that the information may be transferred across the discrete memoryless channel with arbitrarily small error probability.

Proof: In order to generalize the result for the BSC to the case of the general DMC, we must generalize the idea of distance between sequences. The Maximum Likelihood (ML) decoding rule specifies that if $y_i$ is received, it is decoded as an $x_k$ that maximized the forward probability

$$p(y_i|x_k), \quad m = 1, \ldots, M.$$ 

Here we assume that the channel is used $N$ times, that there are $M = K^N$ length-$N$ input vectors $x_n$ and that there are $L = J^N$ length-$N$ output vectors $y_l$. We would like to define a distance metric that is compatible with the ML decoding rule—a distance metric that is monotonically decreasing in $p(y_l|x_m)$. Such a distance metric would have the property that an ML decoding rule would be a minimum distance decoding rule. One such distance metric is

$$D(x_m, y_l) = \log \frac{p^*_y(y_l)}{p(y_l|x_m)}.$$ 

(n.b., $p(y_l|x'_m) > p(y_l|x_m) \iff D(x'_m, y_l) < D(x_m, y_l)$.)

Assuming a particular codeword $x_{i0}$ has been sent across the channel, the average distance between $x_{i0}$ and the received $y_l$ is

$$= \sum_{i=1}^L p(y_l|x_{i0}) D(x_{i0}, y_l)$$

$$= -\sum_{i=1}^L p(y_l|x_{i0}) \log \frac{p(y_l|x_{i0})}{p^*_y(y_l)}.$$ 

Since the channel is memoryless, if $y_l = (y_{l1}, \ldots, y_{lN})$ and $x_{i0} = (x_{i1}, \ldots, x_{iN})$, we can write

$$p(y_l|x_{i0}) = \prod_{n=1}^N p(y_{ln}|x_{in}).$$

Since the codewords are obtained by independent random codeword selection according to the distribution $p^*_x(x_k)$,

$$p^*_y(y_l) = \prod_{n=1}^N p^*_y(y_{ln}).$$

$$-4-$$
Thus $D(x_{i_0}, y_1)$ can be written as

$$D(x_{i_0}, y_1) = - \sum_{n=1}^{N} \sum_{j=1}^{J} p(y_j | x_{in}) \log \frac{p(y_j | x_{in})}{p^*_y(y_j)}.$$  

The inner summation here is just $I(x_{in}; Y)$, and for codewords selected according to $P^*_x$ it is equal to $C$ for all $n = 1, \ldots, N$.

The decoding rule we will use is as follows: Let $r = -N(C - \epsilon)$. Assuming the sequence $y_1$ is received, define the $r$-sphere about $y_1$ as

$$S_r(y_1) = \{ x : D(x, y_1) \leq r \}.$$  

If one and only one codeword $x \in S_r(y_1)$, choose it as the decoded codeword. Otherwise, intentionally make an error by decoding $y_1$ as some arbitrary codeword.

Given two arbitrary positive numbers $\epsilon$ and $6$, the weak law of large numbers asserts that, for sufficiently large $N$, the probability of receiving a sequence $y_1$ such that

$$D(x_{i_0}, y_1) > r = -N(C - \epsilon)$$

can be made less than 6. So if $x_i$ is the transmitted codeword, $P_\epsilon$ can be upper bounded by

$$P_\epsilon \leq 6 + \sum_{m=1}^{M} \Pr(x_m \in S_r(y_1)).$$

Using the random coding argument, we can find the average probability of error, $\overline{P_\epsilon}$, for randomly chosen codewords chosen according to the distribution $P^*_x$.

The average value of $\Pr(x_m \in S_r(y_1))$ is

$$\overline{\Pr(x_m \in S_r(y_1))} = \sum_{y} p^*_y(y) \sum_{x : D(x, y) \leq r} p^*_x(x)$$

$$= \sum_{(x, y) : D(x, y) \leq r} p^*_x(x) p^*_y(y).$$

Now all $(x, y) : D(x, y) \leq r$ satisfy

$$\log \frac{p^*_y(y)}{p(y|x)} < r = -N(C - \epsilon),$$

and thus

$$p^*_y(y) \leq p(y|x) \cdot 2^{-N(C-\epsilon)}.$$

Consequently,

$$\left[ \sum_{(x, y) : D(x, y) \leq r} p^*_x(x) p^*_y(y) \right] \leq 2^{-N(C-\epsilon)} \left[ \sum_{(x, y) : D(x, y) \leq r} p^*_{xy}(x, y) \right] \leq 2^{-N(C-\epsilon)}.$$
Combining all of the above results, we have

\[ P_e \leq \delta + M \cdot 2^{-N(C-\epsilon)} \leq \delta + 2^{-N(C-\epsilon-R)}, \]

where \( M \leq 2^R \) is the number of messages encoded. Since \( \epsilon \) can be made arbitrarily small, the average probability of error \( P_e \) can be made to approach zero as \( N \) becomes large if \( R < C \). And using a procedure similar to that in the case of the BSC, it can be shown that a small \( P_e \) implies that at least one appropriate code exists.

Q. E. D.