1 The Theorem

- For every discrete memoryless channel, the channel capacity $C = \max_{\mathcal{P}_X} H(X;Y)$ has the following property. For any $\epsilon > 0$ and $R < C$, for large enough N, there exists a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \epsilon$.
- If a probability of bit error p_b is acceptable, rates up to $R(p_b)$ are achievable, where

$$R(p_b) = \frac{C}{1 - H_2(p_b)}.$$

- For given p_b, rates greater than $R(p_b)$ are not achievable.

JOINTLY TYPICAL SEQUENCES

We formalize the intuitive preview of the last lecture.[1]

We will define codewords $\mathbf{x}^{(m)}$ as coming from an ensemble X^N, thus defining a joint ensemble $(XY)^N$, corresponding to random generation of a codeword and a corresponding output. A received signal $\mathbf{y}$ will be decoded as m if $\mathbf{x}^{(m)}, \mathbf{y}$ are *jointly typical*.

The proof will then centre on determining the probabilities (a) that the true input codeword is not jointly typical with the output sequence; and (b) that a false input codeword will be jointly typical with the output. We will show that for large N both probabilities $\to 0$, as long as there are fewer than 2^{NC} codewords.

Joint typicality. A pair of sequences $\mathbf{x}, \mathbf{y}$ of length N are jointly typical with respect to the distribution $P(x, y)$ if

$$\left|\frac{1}{N}\log\frac{1}{P(\mathbf{x})} - H(X)\right| < \beta, \left|\frac{1}{N}\log\frac{1}{P(\mathbf{y})} - H(Y)\right| < \beta,$$

$$\text{and } \left|\frac{1}{N}\log\frac{1}{P(\mathbf{x},\mathbf{y})} - H(X,Y)\right| < \beta.$$

The jointly typical set $J_{N\beta}$ is the set of all jointly typical sequence pairs of length N.

Joint typicality theorem. Let $\mathbf{x}, \mathbf{y}$ be drawn from the ensemble $(XY)^N$ with $P(\mathbf{x}, \mathbf{y}) = \prod_{n=1}^{N} P(x_n, y_n)$. Then

1. the probability that $\mathbf{x}, \mathbf{y}$ are jointly typical tends to 1 as $N \to \infty$.
2. the number of jointly typical sequences $|J_{N\beta}|$ is $\leq 2^{N(H(X,Y)+\beta)}$.
3. if $\mathbf{x}' \sim X^N$ and $\mathbf{y}' \sim Y^N$, *i.e.*, $\mathbf{x}'$ and $\mathbf{y}'$ are independent samples with the same marginal distribution as $P(\mathbf{x}, \mathbf{y})$, then

$$P((\mathbf{x}', \mathbf{y}') \in J_{N\beta}) \leq 2^{-N(H(X;Y)-3\beta)}.$$

[1] This exposition is based on that of Cover and Thomas.

Proof. The proof of parts 1 and 2 by the law of large numbers follows that of the source coding theorem in lecture 2. For the third part,

$$\begin{aligned} P((\mathbf{x}', \mathbf{y}') \in J_{N\beta}) &= \textstyle\sum_{(\mathbf{x},\mathbf{y}) \in J_{N\beta}} P(\mathbf{x})P(\mathbf{y}) \\ &\leq |J_{N\beta}|\, 2^{-N(H(X)-\beta)} 2^{-N(H(Y)-\beta)} \\ &\leq 2^{N(H(X,Y)+\beta)-N(H(X)+H(Y)-2\beta)} \\ &= 2^{-N(H(X;Y)-3\beta)} \end{aligned} \qquad (1)$$

2 Proof of the channel coding theorem

SKETCH

We wish to show that there exists a code and a decoder having small probability of error. Evaluating the probability of error of any system is not easy. Shannon's innovation was, instead of constructing a good system and evaluating its error probability, to calculate the average probability of block error of *all* codes, and prove that this average is small. There must then exist individual codes that have small probability of block error. Finally to prove that the *maximal* probability of error is small too, we modify one of these good codes by throwing away the worst 50% of its codewords.

RANDOM CODING

Consider the following rate R' encoding-decoding system.

1. We fix $P(x)$ and generate the $2^{NR'}$ codewords of a (N, NR') code $\mathcal{C}$ at random according to

$$P(\mathbf{x}) = \prod_{n=1}^{N} P(x_n).$$

2. The code is known to both sender and receiver.
3. A message m is chosen at random from $\{1, 2, \ldots 2^{NR'}\}$, and $\mathbf{x}^{(m)}$ is transmitted. The received signal is $\mathbf{y}$, with

$$P(\mathbf{y}|\mathbf{x}^{(m)}) = \prod_{n=1}^{N} P(y_n|x_n^{(m)}).$$

4. The signal is decoded by *typical set decoding*, which decodes $\mathbf{y}$ as $\hat{m}$ if $(\mathbf{x}^{(\hat{m})}, \mathbf{y})$ are jointly typical *and* there is no other m' such that $(\mathbf{x}^{(m')}, \mathbf{y})$ are jointly typical; otherwise a failure (m=0) is declared. (This is not the optimal decoding algorithm, but it will be good enough, and easier to analyse.)
5. A decoding error occurs if $\hat{m} \neq m$.

PROBABILITY OF ERROR

There are two different sources of error when we use typical set decoding. Either (a) the output $\mathbf{y}$ is not jointly typical with the transmitted codeword $\mathbf{x}^{(m)}$, or (b) there is some other codeword that is jointly typical with $\mathbf{y}$.

By the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the value of m; we can assume without loss of generality that $m = 1$.

(a) The probability that $(\mathbf{x}^{(1)}, \mathbf{y})$ are not jointly typical is $\leq \delta$ by the joint typicality theorem part 1, where $\delta \to 0$ as $N \to \infty$.

(b) The probability that $(\mathbf{x}^{(m)}, \mathbf{y})$ are jointly typical, for a given $m \neq 1$ is $\leq 2^{-N(H(X;Y)-3\beta)}$, by part 3.

Thus the average probability of error $\langle p_{\rm B} \rangle$ satisfies:

$$\begin{aligned} \langle p_{\rm B} \rangle &\leq \delta + \sum_{m=2}^{2^{NR'}} 2^{-N(H(X;Y)-3\beta)} \\ &\leq \delta + 2^{-N(H(X;Y)-R'-3\beta)} \end{aligned}$$

This is $< 2\delta$ for sufficiently large N *if*

$$R' < H(X;Y) - 3\beta.$$

We are almost there. We make three modifications:

1. We choose $P(x)$ in the proof to be the optimal input distribution of the channel. Then the condition $R' < H(X;Y) - 3\beta$ becomes $R' < C - 3\beta$.

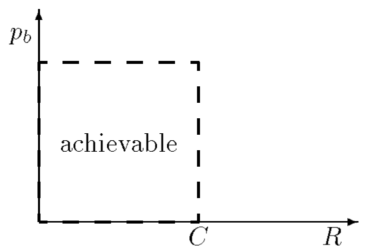
2. Since the average probability of error over all codes is $< 2\delta$, there must exist a code with $p_{\rm B} < 2\delta$.

3. To show that not only the average but also the maximal probability of error can be made small, we modify this code by throwing away the worst half of the codewords—the ones most likely to produce errors. Those that remain must have conditional probability of error less than 4δ. This new code has $2^{NR'-1}$ codewords, *i.e.*, we have reduced the rate negligibly from R' to $R' - \frac{1}{N}$, and achieved $p_{\rm BM} < 4\delta$.

In conclusion, we have 'constructed' a code of rate $R' - \frac{1}{N}$, where $R' < C - 3\beta$, with maximal probability of error $< 4\delta$. We obtain the theorem as stated by setting $R' = (R+C)/2$, $\delta = \epsilon/4$, $\beta < (C - R')/3$, and N sufficiently large for the remaining conditions to hold. The theorem's first part is thus proved.

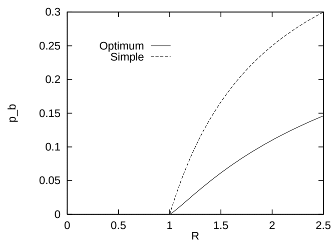
COMMUNICATION ABOVE CAPACITY

We have now proved the achievability of a portion of the R, p_b plane, for any discrete memoryless channel:



We have shown that we can turn any noisy channel into an essentially noiseless binary channel with rate up to C bits per cycle. We now extend the right-hand boundary of the region of achievability at non-zero error rates.

Consider a noiseless binary channel, and assume that we force communication at a rate greater than its capacity of 1 bit. For example, if we require the sender to attempt to communicate at R=2 bits per cycle then they must effectively throw away half of the information. What is the best way to do this if the aim is to achieve the smallest possible probability of bit error? One simple strategy is to just communicate a fraction $1/R$ of the source bits, and ignore the rest. The receiver guesses the missing fraction $1 - 1/R$ at random, and the average probability of bit error is $p_b = \frac{1}{2}(1 - 1/R)$. We can do better than this (in terms of minimizing p_b) by spreading out the risk of corruption evenly among all the bits. In fact we can achieve $p_b = H_2^{-1}(1 - 1/R)$. The two graphs are shown here:



So, how can this optimum be achieved?

We reuse a tool that we just developed, namely the (N, K) code for a noisy binary symmetric channel, and we turn it on its head, using the *decoder* to define a lossy compressor. Assume the original code has a rate $R' = K/N$, and that it is capable of correcting errors in a binary symmetric channel with probability of error q. Assymptotically, rate R' codes exist that have $R' = 1 - H_2(q)$.

We take the signal that we wish to send, and chop it into blocks of length N (NB, not K). We pass each block through the *decoder*, and obtain a shorter signal of length K bits, which we communicate over the noiseless channel. To decode the transmission, we pass the K bit message to the *encoder* of the original code. The reconstituted message will now differ from the original message in some of its bits; the probability of bit error will be $p_b = q$. The rate of this lossy compressor is $R = N/K = 1/R' = 1/(1 - H_2(p_b))$.

Now attaching this lossy compressor to our capacity C error-free communicator, we have proved the achievability of communication up to the curve (p_b, R) defined by:

$$R = \frac{C}{1 - H_2(p_b)}.$$

THE NON-ACHIEVABLE REGION

The source, encoder, noisy channel and decoder define a Markov chain:

$$P(m, \mathbf{x}, \mathbf{y}, \hat{m}) = P(m)P(\mathbf{x}|m)P(\mathbf{y}|\mathbf{x})P(\hat{m}|\mathbf{y}).$$

The data processing inequality must apply to this chain: $H(m; \hat{m}) \leq H(\mathbf{x}; \mathbf{y})$. Furthermore, by the definition of channel capacity, $H(\mathbf{x}; \mathbf{y}) \leq NC$, so $H(m; \hat{m}) \leq NC$.

Assume that a system achieves a rate R and a bit error probability p_b; then the mutual information $H(m; \hat{m})$ is $\geq NR(1 - H(p_b))$. But $H(m; \hat{m}) > NC$ is not achievable, so $R > \frac{C}{1-H_2(p_b)}$ is not achievable.