Communication Theory.

A typical scheme: source $\longrightarrow$ encoder $\longrightarrow$ channel $\longrightarrow$ decoder $\longrightarrow$ destination

§1. Sources and Coders.

A source emits a 'text' (a sequence of letters): $u_1, u_2, u_3, \dots$ (1), $u_i \in I (= I_m)$; think of $I = \{1, \dots, m\}$. A common approach: consider (1) as a sample of a random text, ie. a sequence of random letters: $U_1, U_2, U_3, \dots$ (2)

Examples: (i) A sequence of IID rvs U_n with values in I. $\mathbb{P}(U_1 = u_1, \dots, U_n = u_n) = \prod \mathbb{P}(U_j = u_j) = \prod p(u_j)$, (3a) where $p(u)$, $u \in I$, is a probability distribution on I. This source is called a Bernoulli source.

(ii) A Markov source: $\mathbb{P}(U_1 = u_1, \dots, U_k = u_k) = p_1(u_1) \prod_{j=1}^{k-1} \mathbb{P}(u_j, u_{j+1})$, where $p_1(u) = \mathbb{P}(U_1 = u)$, and $\mathbb{P}(u, u') = \mathbb{P}(u_{j+1} = u' / u_j = u)$. (3b)

Stationarity: $\mathbb{P}(U_j = u) = \mathbb{P}(U_1 = u) = p_1(u)$, or $p_1 P = p_1$ (p_1 an equilibrium distribution).

Completely reducible: $\mathbb{P}(U_1 = U_2 = \dots = U_k = l) = q(u)$, $\sum_{u \in I} q(u) = 1$. (ie. source emits repeated letters).

A message (string, word) of length n: $u^{(n)} = u_1, \dots, u_n$; the random string, $U^{(n)} = (U_1, \dots, U_n)$.
An encoder (coder) uses a code, ie a map $f : u \in I \mapsto f(u) = x_1 \dots x_s$, $x_i \in J_a$. Think of J as $\{0, \dots, a-1\}$. A typical case is $a = 2$, ie, $J = \{0, 1\}$, ie, binary codes.
The strings (with digits from J) of the form $f(u)$ are called the codewords of f.
If you have $u^{(n)}$, then $f(u^{(n)}) = f(u_1) \dots f(u_n)$ – concatenation.

Definition 1: f is called decipherable if any string with digits from J is the image of ≤ 1 message. A string x is called a prefix in y if $y = xz$. A code f is called prefix-free if no codeword $f(u)$ is a prefix of any other codeword $f(u')$.

Note: A prefix-free code is decipherable. Converse is false – eg, a code f with $I = \{1, 2, 3\}$ and $f(1) = 0$, $f(2) = 01$, $f(3) = 011$ is decipherable, but not prefix-free.

Theorem 1 – Kraft Inequality: Given positive integers $s_1, \dots, s_m$, $\exists$ a decipherable code with codeword-lengths $s_1, \dots, s_m$ iff $\sum_{i=1}^{m} a^{-s_i} \le 1$ (4). If (4) holds, $\exists$ a prefix-free code.

Proof: if: if (4) holds then $\sum_{l=1}^{s} n_l a^{-l} \le 1$, (5), where n_l is the multiplicity of l among $s_1, \dots, s_m$ and $s = \max\{s_i\}$. That is, $n_s a^{-s} \le 1 - \sum_{l=1}^{s-1} n_l a^{-l}$ (6,1),
So, $0 \le n_s \le a^s - \sum_{l=1}^{s-1} n_l a^{s-l}$ $\therefore$ $n_{s-1} a \le a^s - \sum_{l=1}^{s-2} n_l a^{s-l}$. So, $0 \le n_{s-1} \le a^{s-1} - \sum_{l=1}^{s-2} n_l a^{s-l-1}$ (6,2).
... so $n_2 \le a^2 - n_1 a$, (6, s-1), and so $n_1 \le a$ (6, s).
Use (6,1) – (6,s) in the reverse order. (6,s) means that you can form n_1 words of length 1; this leaves $a - n_1$ symbols unused. Use them to form $(a - n_1) a$ words of length 2. (6, s-1) means that you can use n_2 of these words as codewords. This leaves $a^2 - n_1 a - n_2$ words unused. Etc.
At the end you get a prefix-free code that meets the requirements.

only if: if $\exists$ decipherable code then, $\forall r \in \mathbb{Z}_+$, $(a^{-s_1} + \dots + a^{-s_m})^r = \sum_{l=1}^{rs} b_l a^{-l}$, where b_l is the number of ways r codewords may be put together to form a string of length l. As the code is decipherable, these strings are distinct. Ie. $b_l \le a^l$ (the total number of l-strings). So, $(a^{-s_1} + \dots + a^{-s_m})^r \le rs \Rightarrow (a^{-s_1} + \dots + a^{-s_m}) = (rs)^{1/r} \to 1$ as $r \to \infty$. So done.

Question: What are "best" decipherable (or prefix-free) codes?

Remarks: (i) A code obeying (4) is not necessarily decipherable.
(ii) Prefix-free codes suffice.

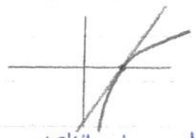
Think of a random source, $\mathbb{P}(U=u) = p(u)$. Want to minimise the expected codeword length, S, under a code f, $\mathbb{E}S = \sum_s s\,\mathbb{P}(S=s) = \sum_{i=1}^{m} s_i\, p(i)$, over decipherable codes.
An optimisation problem: minimise $\mathbb{E}S = \sum s_i\, p(i)$ subject to $\sum a^{-s_i} \le 1$ and $s_i \in \mathbb{Z}_+$.
Change last condition to: $s_1, \dots, s_m \ge 0$.
Use the Lagrange Sufficiency Theorem; the Lagrangian is: $L = \sum s_i p(i) + \lambda\left(1 - \sum a^{-s_i} - z\right)$.
Minimising in s_i yields $\lambda < 0$, $z = 0$, $\frac{\partial L}{\partial s_i} = 0$, whence $-\frac{p(i)}{\lambda \ln a} = a^{-s_i} \Leftrightarrow s_i = -\log_a p(i) - \log_a(-\lambda \ln a)$.
Adjusting the constraint gives $-\lambda \ln a = 0$, so $s_i = -\log_a p(i)$.
This is the solution to the relaxed problem (where we do not require $s_i \in \mathbb{N}$).
The formula above gives a lower bound for the solution to the original problem.
That is, $\min \mathbb{E}S \ge h_a = -\sum p(i) \frac{\log_2 p(i)}{\log_2 a}$.

The quantity $h_2 = -\sum p(i) \log_2 p(i)$ is called the binary entropy of the probability distribution.

In future, we will use: $\log_2 = \log$, $0 \log 0 = 0 = 0 \log \infty$.

Theorem 2.1 (Gibb's Inequality): Let $\{p(i)\}$ and $\{p'(i)\}$ be two probability distributions.
Then, $\forall\, b > 1$ $\sum_{i=1}^{m} p(i) \log_b \frac{p'(i)}{p(i)} \le 0$, ie, $-\sum p(i)\log_b p(i) \le -\sum p(i) \log_b p'(i)$, with equality iff $p = p'$.

Proof: Use $\log_b x \le \frac{x-1}{\log_e b}$ (= iff $x = 1$):



In fact, $\sum_{i\in I'} p(i) \log \frac{p'(i)}{p(i)} \le (\log_e b)^{-1} \sum_{i \in I'} p(i)\left(\frac{p'(i)}{p(i)} - 1\right) = \frac{1}{\log_e b}\left(\sum_{i\in I'} p'(i) - \sum_{i \in I'} p(i)\right) \le 0$,
with = iff $p = p'$. (Must be careful about $p(i) = 0$ – see notes). [$I' = \{i : p(i) > 0\}$]

Theorem 2.2. (Shannon's Noiseless Coding Theorem): If a source emits i with probability $p(i)$, $(i = 1, \dots, m)$, then $\min \mathbb{E}S$ (over the decipherable codes) obeys $\frac{h}{\log a} \le \min \mathbb{E}S \le \frac{h}{\log a} + 1$.

Proof: The LH bound has been proved before.
Take $s_i \in \mathbb{N}$ such that $a^{-s_i} \le p(i) < a^{-s_i + 1}$. Then $\sum_i a^{-s_i} \le \sum p(i) = 1$ (Kraft).
$\therefore \exists$ a decipherable code with codewords lengths $s_1, \dots, s_m$. From RH inequality, we get $s_i < -\frac{\log p(i)}{\log a} + 1$, and $\mathbb{E}S < -\frac{\sum p(i) \log p(i)}{\log a} + 1 = \frac{h}{\log a} + 1$.

Shannon's NC Theorem gives a base for Shannon-Fano encoding rules: Fix $s_1, \dots, s_m \in \mathbb{N}$ as above. Then take a code with the codeword-lengths $s_1, \dots, s_m$, from the shortest word upwards, ensuring that shorter words don't appear as prefixes. The Kraft inequality guarantees that this is possible.

An optimal code was constructed by Huffman. The case $a=2$ only: let the probabilities be $p(i)$ $(i=1,..,m)$. Wlog, assume $p(1) \geq .. \geq p(m)$ Then:
(i) assign symbol 0 to $m-1$ and 1 to m.
(ii) Take a "reduced" alphabet, I_{m-1}, by merging $m-1$ and m. Assign to $(m-1, m)$ the probability $p(m-1)+p(m)$. Rearrange the probabilities.
Then repeat the procedure. Obtain a tree-like stucture.

Example:

i	$p(i)$	$f(i)$	s_i
1	.5	0	1
2	.15	100	3
3	.15	101	3
4	.1	110	3
5	.05	1110	4
6	.025	11110	5
7	.025	11111	5

For tree, see notes.

Lemma 2.3: Any optimal prefix-free code has the codeword-lengths reverse-ordered vs their probabilities.
Proof: obvious, otherwise shuffling would give a better code.

Lemma 2.4: In any optimal prefix-free code, $\exists$, among the codewords of maximum length at least two agreeing in all but the last digit.
Proof: Suppose not. Then either: (i) $\exists$ a unique codeword of maximal length, or (ii) $\exists \geq 2$ codewords of maximal length and they differ before the last digit.
In both cases, you can drop the last digit from the codewords under consideration. The prefix-free condition is retained, but the code becomes shorter. ✳.

Theorem 2.5: The Huffman code is optimal decipherable code.
Proof: Induction on $m=|I|$. For $m=2$, the case is trivial. Suppose the optimality for I_{m-1} $\forall$ probability distributions. Take I_m, assume $\exists$ a code f_m^*, better than f_m, ie. $\mathbb{E}S_m^* \leq \mathbb{E}S_m$. Wlog, $p(1) \geq .. \geq p(m)$. By lemmas 3 and 4, in both codes, the codewords for $m-1$ and m have maximal length and differ only in the last digit.
Reduce both codes to I_{m-1}: "glue" these codewords after dropping the last digit.
The Huffman code f_m becomes f_{m-1}; code f_m^* becomes f_{m-1}^*. In f_m, the contribution to $\mathbb{E}S_m$ from $f_m(m-1)$ and $f_m(m)$ was $s_m(p(m-1)+p(m))$. After reduction, it equals $(s_m - 1)(p(m-1)+p(m))$ $\therefore$ $\mathbb{E}S_m$ is reduced by $p(m-1)+p(m)$.
In f_m^*, the contribution from $f_m^*(m-1)$ and $f_m^*(m)$ was $s_m^*(p(m-1)+p(m))$. After reduction, it equals $(s_m^*-1)(p(m-1)+p(m))$ $\therefore$ $\mathbb{E}S_m^*$ is reduced by $p(m-1)+p(m)$.
As f_m^* was better than f_m, f_{m-1}^* has to be better than f_{m-1}. ✳.

In what follows we set $a=2$. The modern view of encoding is based on the segmentation. We do not encode symbols from $u \in I$, but we divide the source message into 'blocks' or 'segments' and encode these by codewords. It increases the nominal number of letters, as the segments of length n fill the Cartesian product $I^n = I \times \cdots \times I$.

But what matters is the binary entropy of the probability distribution ~~on~~ of our blocks on I^n.

$$h^{(n)} = -\sum_{i_1,\dots,i_n} \mathbb{P}(U_1 = u_1, \dots, U_n = u_n) \log \mathbb{P}(U_1 = u_1, \dots, U_n = u_n).$$

Denote by $S^{(n)}$ the random codeword-length in a code $f_n : I^n \to J$. The minimum expected codeword-length per letter is $e_n = \frac{1}{n} \min_{f_n} \mathbb{E} S^{(n)}$.

By Shannon's NC Theorem, $\frac{h^{(n)}}{n \log a} \le e_n < \frac{h^{(n)}}{n \log a} + \frac{1}{n}$. So $e_n \sim \frac{h^{(n)}}{n \log a} = \frac{h^{(n)}}{n}$, as $\log a = 1$.

Example: For a Bernoulli source, $h^{(n)} = -\sum p(i_1) \cdots p(i_n) \log (p(i_1) \cdots p(i_n))$

$$= -\sum_j \sum_{i_1,\dots,i_n} p(i_1) \cdots p(i_n) \log p(i_j) = -n \sum p(i) \log p(i) = nh,$$

where h is the entropy of the single-letter distribution. Thus $e_n \sim \frac{nh}{n} = h$.

Definition: A source is called reliably encodable at rate $R > 0$ if, $\forall n, \exists$ a set A_n on n-strings such that $\# A_n \le 2^{nR}$ and $\lim_{n \to \infty} \mathbb{P}(U^{(n)} \in A_n) = 1$.

Definition: The information rate of a source is $H = \inf \{ R : R \text{ is reliable} \}$

Theorem 2.7: The information rate of a source with alphabet I_m is $0 \le H \le \log m$, with both bounds being attainable.

Proof: The LH $\le$ holds by definition. Equality holds, eg, for a Markov source repeating the symbols. On the other hand, $|I^n| = m^n$, hence $R = \log m$ is a reliable encoding rate since $2^{nR} = 2^{n \log m} = m^n$. Thus $H \le \log m$.

Equality holds for the equidistributed Bernoulli source. Here, if you take $R < \log m$ then $\mathbb{P}(A_n) = |A_n| \left(\frac{1}{m}\right)^n \le \frac{2^{nR}}{m^n} = 2^{nR - n \log m} \to 0$ as $n \to \infty$. Thus $\forall R < \log m$, rate R is not reliable.

3. Information and Entropy.

Definition: If A is an event, the information gained from observing A is: $i(A) = -\log p(A)$.

If X is a random variable, the entropy of X, $h(X)$, is defined as

$$h(X) = -\sum_{x_i} p(x_i) \log p(x_i) = -\sum p_i \log p_i.$$

The entropy is the expected value of the information gained while observing X. $h(X) = h(p_1, \dots, p_n)$. Given a pair of random variables X, Y, define the joint entropy:

$$h(X,Y) = -\sum_{x_i, y_j} P_{X,Y}(x_i, y_j) \log P_{X,Y}(x_i, y_j).$$

The conditional entropy $h(X|Y)$ of X given Y is: $h(X|Y) = -\sum_{x_i, y_j} P_{X,Y}(x_i, y_j) \log P_{X,Y}(x_i | y_j)$

It is easy too see that $h(X,Y) = h(X|Y) - h(Y)$. [And $h(X|Y) \ne h(Y|X)$.]

If A_1 and A_2 are independent, then $i(A_1 \cap A_2) = i(A_1) + i(A_2)$. For A with $p(A) = \frac{1}{2}$, have $i(A) = 1$. (1 bit of information).

Theorem 3.1: (a) For a random variable X with $\leq m$ values, $0 \leq h(X) \leq \log m$. The LH = occurs iff X = constant with probability 1; the RH = iff $\mathbb{P}(X=i) = \frac{1}{m}$.

(b) $h(X,Y) \leq h(X) + h(Y)$, with = iff X and Y are independent.

Proof: Use the Gibbs inequality. (a) $p(i) = \mathbb{P}(X=i)$, $p'(i) = \frac{1}{m}$. Then, $-\sum p(i) \log p(i) \leq -\sum p(i) \log \frac{1}{m} = \log m$. The LH $\leq$ is trivial.

(b) $p(i) = \mathbb{P}(X=i_1, Y=i_2)$, $i = (i_1, i_2)$, $p'(i) = \mathbb{P}(X=i_1)\mathbb{P}(Y=i_2)$. Then,

$$h(X,Y) = -\sum P_{X,Y}(i_1,i_2) \log P_{X,Y}(i_1,i_2) \leq -\sum P_{X,Y}(i_1,i_2)\left(\log P_X(i_1) + \log P_Y(i_2)\right)$$
$$= -\sum_{i_1,i_2} P_{X,Y}(i_1,i_2) \log P_X(i_1) - \sum_{i_1,i_2} P_{X,Y}(i_1,i_2) \log P_Y(i_2) = h(X) + h(Y).$$

The equality occurs iff $p = p'$, ie, X and Y are independent.

Lemma 3.2: (The pooling inequality): $\forall\, q_1, q_2 \geq 0$, with $q_1 + q_2 > 0$,

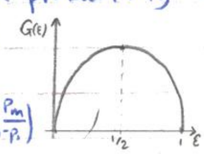
$$-(q_1+q_2)\log(q_1+q_2) \leq -q_1 \log q_1 - q_2 \log q_2 \leq -(q_1+q_2)\log\left(\frac{q_1+q_2}{2}\right);$$ the LH = iff $q_1 q_2 = 0$; the RH = iff $q_1 = q_2$.

Proof: This is equivalent to: $0 \leq h\left(\frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right) \leq \log 2 \quad (=1)$.

Theorem 3.3: If $X = \varphi(Y)$ then $h(X) \leq h(Y)$, the = iff φ is invertible.

Proof: Follows from lemma 3.2.

Theorem 3.4 (The Fano Inequality): Let X take $m > 1$ values, one of them with probability $1-\varepsilon$. Then $h(X) \leq G(\varepsilon) + \varepsilon \log(m-1)$, where $G(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon)$



Proof: Suppose that $p(x_1) = 1-\varepsilon$. Then $h(X) = h(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i$

$$= -p_1 \log p_1 - (1-p_1)\log(1-p_1) + (1-p_1)\log(1-p_1) - \sum_{i=2}^{m} p_i \log p_i = h(p_1, 1-p_1) + (1-p_1)\, h\left(\frac{p_2}{1-p_1}, \ldots, \frac{p_m}{1-p_1}\right)$$

In the RHS, the first term is $G(\varepsilon)$; the second $\leq \varepsilon \log(m-1)$.

Definition: Random variables X, Y, Z. X and Y are conditionally independent given Z if
$$\mathbb{P}(X=x, Y=y \mid Z=z) = P(X=x \mid Z=z)\,\mathbb{P}(Y=y \mid Z=z)$$

Theorem 3.5: (a) $0 \leq h(X|Y) \leq h(X)$; LH $\leq$ = iff $X = \varphi(Y)$, RH = iff X and Y are independent.

(b) $h(X|Y,Z) \leq h(X,Y) \leq h(X|\varphi(Y))$; LH = iff X and Y are conditionally independent given Z, RH = iff X and Y are conditionally independent given $\varphi(Y)$.

Proof: (a) Easy from previous bounds.

(b) For the LH $\leq$, use $h(X|Y,Z) = h(X,Z|Y) - h(Z|Y)$, (1), together with $h(X,Z|Y) \leq h(X|Y) + h(Z|Y)$, (2)
The RH $\leq$ follows from $h(X|Y,\varphi(Y)) = h(X,Y|\varphi(Y)) - h(Y,\varphi(Y))$, together with $h(X|Y,\varphi(Y)) = h(X,Y)$ and, in addition, an inequality which in the form of (2): $h(X,Y|\varphi(Y)) \leq h(X|\varphi(Y)) + h(Y|\varphi(Y))$. Equality cases are identified by inspection.

Theorem 3.6 (Generalised Fano inequality): Let X, Y be a pair of random variables with values: $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$. Assume $\sum_{j=1}^{m} p(X=x_j, Y=y_j) = 1-\varepsilon$. Then $h(X|Y) \leq G(\varepsilon) + \varepsilon \log(m-1)$ [G as above]

Proof: Let $\varepsilon_j = p(X \neq x_j, Y=y_j)$. Then $\sum_j P_Y(y_j)\,\varepsilon_j = \varepsilon$. By using standard definitions, Fano inequality and concavity of $G(\varepsilon)$, get: $h(X|Y) \leq \sum_j P_Y(y_j)\left(G(\varepsilon_j) + \varepsilon_j \log(m-1)\right) = \sum P_Y(y_j)\, G(\varepsilon_j) + \varepsilon \log(m-1) \leq G(\varepsilon) + \varepsilon \log(m-1)$.

Theorem 3.7: If $X^{(n)} = (X_1, .., X_n)$, $Y^{(n)} = (Y_1, .., Y_n)$ are random vectors, then

(a) $h(X^{(n)}) = \sum_{i=1}^{n} h(X_i | X^{(i-1)}) \leq \sum_{i=1}^{n} h(X_i)$ with equality iff $X_1, .., X_n$ are independent.

(b) $h(X^{(n)} | Y^{(n)}) \leq \sum_{i=1}^{n} h(X_i | Y^{(n)}) \leq \sum_{i=1}^{n} h(X_i | Y_i)$, with LH = iff $X_1, .., X_n$ are conditionally independent given $Y^{(n)}$, the RH = iff $\forall\, i = 1, .., n$, X_i and $\{Y_r : r \neq i\}$ are conditionally independent given Y_i.

Proof: Follows from previous results.

Definition: The mutual entropy between X and Y is: $i(X,Y) := \mathbb{E} \log \frac{P_{X,Y}(X,Y)}{P_X(X) P_Y(Y)} = h(X) + h(Y) - h(X,Y)$.

Theorem 3.8: $0 \leq i(X, \varphi(Y)) \leq i(X, Y)$, the LH = iff X and $\varphi(Y)$ are independent; the RH = iff X and Y are conditionally independent given $\varphi(Y)$.

Proof: follows from previous results.

Theorem 3.9: (a) $i(X^{(n)}, Y^{(n)}) \geq h(X^{(n)}) - \sum_{i=1}^{n} h(X_i | Y^{(n)}) \geq h(X^{(n)}) - \sum_{i=1}^{n} h(X_i | Y_i)$

(b) if $X_1, .., X_n$ are independent, $i(X^{(n)}, Y^{(n)}) \geq \sum_{i=1}^{n} i(X_i, Y^{(n)}) \geq \sum_{i=1}^{n} i(X_i, Y_i)$

Proof: Follows from previous results.

4. Shannon's First Coding Theorem

Definition: $D_n(R) := \max_{\substack{A \subset I^n \\ \# A \leq 2^{nR}}} \mathbb{P}(U^{(n)} \in A)$.

Lemma 4.1: $\forall\, \varepsilon > 0$, $\lim_{n\to\infty} D_n(H+\varepsilon) = 1$, and if $H > 0$, $D_n(H-\varepsilon) \not\to 1$

Proof: $R = H + \varepsilon$ is a reliable rate. Thus, $\exists$ a sequence $A_n \subset I^n$ with $\# A_n \leq 2^{nR}$, and $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$. Thus, $D_n(R) \geq \mathbb{P}(U^{(n)} \in A) \to 1$.

If $H > 0$, then $R = H - \varepsilon > 0$ for small ε, but there is no sequence A_n with the above property. Take a set C_n where $\max \mathbb{P}(U^{(n)} \in C_n)$ is attained, then $D_n(R) = \mathbb{P}(U^{(n)} \in C_n) \not\to$ as $n \to \infty$.

Given $u^{(n)} = u_1 .. u_n$, denote $\xi_n(u^{(n)}) = -\frac{1}{n} \log_+ p_n(u^{(n)})$ $\left[\log_+ x = \begin{cases} \log x & \text{if } x > 0 \\ 0 & \text{if } x = 0 \end{cases} \right]$.

If $U^{(n)}$ is a random string, then $\xi_n(U^{(n)}) = -\frac{1}{n} \log_+ p_n(U^{(n)})$ is a random variable.

Lemma 4.2: $\forall\, R, \varepsilon > 0$, $\mathbb{P}(\xi_n \leq R) \leq D_n(R) \leq \mathbb{P}(\xi_n \leq R) + 2^{-n\varepsilon}$.

Proof: Set $B_n = \{u^{(n)} \in I^n : p_n(u) \geq 2^{-nR}\} = \{u^{(n)} \in I^n : -\log p_n(u) \leq nR\} = \{u^{(n)} \in I^n : \xi_n(u) \leq R\}$.

Then $1 \geq \mathbb{P}(U^{(n)} \in B_n) = \sum_{u \in B_n} p(u^{(n)}) \geq 2^{-nR} \cdot \# B_n$. So $\# B_n \leq 2^{nR}$. Hence the LH $\leq$.

On the other hand, $\exists\, C_n \in I^n$ where $D_n(R)$ is attained. For such a C_n,

$D_n(R) = \mathbb{P}(U^{(n)} \in C_n) = \mathbb{P}(U^{(n)} \in C_n, \xi_n \leq R + \varepsilon) + \mathbb{P}(U^{(n)} \in C_n, \xi_n \geq R + \varepsilon) \leq \mathbb{P}(\xi_n \leq R + \varepsilon) + \sum_{\substack{u \in C_n \\ p_n(u) \leq 2^{-n(R+\varepsilon)}}} p_n(u)$.

$\leq \mathbb{P}(\xi_n \leq R + \varepsilon) + 2^{-n(R+\varepsilon)} \cdot \# C_n = \mathbb{P}(\xi_n \leq R + \varepsilon) + 2^{-n(R+\varepsilon)} \cdot 2^{nR}$. So done.

Definition: A sequence of random variables $\{\eta_n\}$ converges in probability to a random variable η (possibly a constant) if, $\forall\, \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\eta_n - \eta| \geq \varepsilon) = 0$.

Write $\eta_n \xRightarrow{\mathbb{P}} \eta$

Theorem 4.5: If $X_1, X_2, ..$ is a sequence of iid rvs with $\mathbb{E}X = a$ then $\frac{1}{n}\sum_{i=1}^{n} X_i \overset{\mathbb{P}}{\Rightarrow} a$.
(Law of Large Numbers)

Theorem 4.3: (Shannon's First Coding Theorem): If the r.v. $\xi_n \overset{\mathbb{P}}{\Rightarrow} \gamma$, a non-random constant, then $\gamma = H$, the information rate of the source.

Proof: Let $\xi_n \overset{\mathbb{P}}{\Rightarrow} \gamma$. Then $\gamma \geq 0$. By the last lemma, $\forall \varepsilon > 0$ $D_n(\gamma+\varepsilon) \geq \mathbb{P}(\xi_n \leq \gamma + \varepsilon)$ $\geq \mathbb{P}(\gamma - \varepsilon \leq \xi_n \leq \gamma + \varepsilon) = \mathbb{P}(|\xi_n - \gamma| \leq \varepsilon) = 1 - \mathbb{P}(|\xi_n - \gamma| > \varepsilon) \to 1$ as $n \to \infty$. Thus, $H \leq \gamma$.
If $\gamma = 0$ then $H = 0$. Assume that $\gamma > 0$. Then, by the last lemma, $D_n(\gamma - \varepsilon) \leq \mathbb{P}(\xi_n \leq \gamma - \frac{\varepsilon}{2}) + 2^{-n\varepsilon/2}$ $\leq \mathbb{P}(|\xi_n - \gamma| \geq \varepsilon/2) + 2^{-n\varepsilon/2} \to 0$ as $n \to \infty$. Thus $H \geq \gamma$. Hence $H = \gamma$.

Remarks: (i) $\xi_n \overset{\mathbb{P}}{\Rightarrow} \gamma$ is equivalent to the asymptotic equipartition property (AEP) $\lim_{n\to\infty} \mathbb{P}(2^{-n(H+\varepsilon)} \leq p_n(U^{(n)}) \leq 2^{-n(H-\varepsilon)}) = 1$. The proof is by inspection - see notes.
In other words, $\forall \varepsilon > 0$, $\exists n_0(\varepsilon)$ such that $\forall n \geq n_0(\varepsilon)$, the whole set I^n is decomposed into two subsets: Π_n, T_n, so that: (a) $\mathbb{P}(U^{(n)} \in \Pi) < \varepsilon$,
(b) $\forall u^{(n)} \in T^n$, $2^{-n(H+\varepsilon)} \leq \mathbb{P}(U^{(n)} = u^{(n)}) \leq 2^{-n(H-\varepsilon)}$.
(ii) The expected value, $\mathbb{E}\xi_n = -\frac{1}{n}\sum_{u^{(n)}} p_n(u^{(n)}) \log p_n(u^{(n)}) = h(U^{(n)})$.

Theorem 4.4: For a Bernoulli source, $H = h\ (= -\sum_{u \in I} p(u) \log p(u))$.

Proof: For a Bernoulli source, $p_n(u^{(n)}) = p(u_1) \cdots p(u_n)$. Thus, $-\frac{1}{n}\log p_n(u^{(n)}) = \frac{1}{n}\sum_{j=1}^{n} -\log p(u_j)$.
For a random string, $U^{(n)}$, $\xi_n := -\frac{1}{n}\log p_n(U^{(n)}) = \frac{1}{n}\sum_{j=1}^{n} -\log p(U_j)$.
Set $\sigma_j := -\log p(U_j)$, then $\sigma_1, \sigma_2, \ldots$ are iid rvs. Then $\xi_n \overset{\mathbb{P}}{\Rightarrow} \gamma$ is a LLNs for this sequence. γ must be $= \mathbb{E}\sigma$. Ie. $\xi_n \overset{\mathbb{P}}{\to} \gamma$ is equivalent to $\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n}\sigma_j - \mathbb{E}\sigma| \geq \varepsilon) \to 0$ $\forall \varepsilon$ $(n\to\infty)$.
Now, the expected value is $\mathbb{E}\sigma = -\sum_{u \in I} p(u) \log p(u) = h$.
You conclude that the statement of the theorem is a LLNs.

The proof of the LLNs is based on:

Lemma 4.6 (Chebyshov's Inequality): $\mathbb{P}(\eta \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}\eta^2$.

Proof: $\mathbb{P}(\eta \geq \varepsilon) = \mathbb{E}\,\mathbb{1}_{(\eta \geq \varepsilon)} \leq \mathbb{E}(\frac{\eta}{\varepsilon})^2 \mathbb{1}_{(\eta \geq \varepsilon)} \leq \frac{1}{\varepsilon^2}\mathbb{E}\eta^2$.

By Chebyshov, $\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n}\sigma_j - h| \geq \varepsilon) \leq \frac{1}{\varepsilon^2 n^2}\mathbb{E}(\sum_{j=1}^{n}(\sigma_j - h))^2 = \frac{1}{\varepsilon^2 n^2} Var(\sum_{j=1}^{n}\sigma_j) = \frac{1}{\varepsilon^2 n^2} \cdot n\, Var\, \sigma_1 \to 0$ as $n \to \infty$.

5. The entropy rate of a Markov source

For a Markov source, assume that $U_1, U_2, \ldots$ is a Markov source, with $\min_{u,u'} P^{(r)}(u,u') = \rho > 0$, ⊛
[In fact, $\rho \in (0,1)$], for some $r \geq 1$. This means that the Markov chain is irreducible and aperiodic. Hence it has a unique invariant (or equilibrium) distribution: $w(1), \ldots, w(m)$, where $w(v) = \sum_{u \in I} w(u) P(u,v)$.
Moreover, the n-step transition probabilities and the probabilities $\mathbb{P}(U_n = v)$ converge to $w(v)$: $\lim_{n\to\infty} P^{(n)}(u,v) = w(v) = \lim_{n\to\infty}\mathbb{P}(U_n = v) = \lim_{n\to\infty}(p_1 P^{n-1})(v)$.

* Theorem 5.1: For a Markov chain, with condition ⊛, $|P^{(n)}(u,v) - w(v)| \leq (1-\rho)^n$, $|\mathbb{P}(U_n = v) - w(v)| \leq (1-\rho)^{n-1}$
*

Theorem 5.2: For a Markov source, under condition ⊛, $H = -\sum_{u,v} \underbrace{w(u)P(u,v)}_{``\mathbb{P}(U_n = u, U_{n+1}=v)} \log P(u,v) = \lim_{n\to\infty} h(U_{n+1}|U_n)$.

For a stationary source, $H = h(U_2|U_1)$ $[= h(U_{n+1}|U_n)\ \forall n]$.

Proof: Again analyse $\xi_n := -\frac{1}{n}\log p_n(U^{(n)})$. By the Markov property, $p_n(u^{(n)}) = p_1(u_1)P(u_1,u_2)\ldots P(u_{n-1},u_n)$, and $-\frac{1}{n}\log p_n(u^{(n)}) = -\frac{1}{n}[\log p_1(u_1) + \log P(u_1,u_2) + \ldots + P(u_{n-1},u_n)]$.

Hence, $\xi_n = \frac{1}{n}(-\log p_1(U_1) - \log P(U_1,U_2) - \ldots - \log P(U_{n-1},U_n))$.

As in the Bernoulli case, set $\sigma_1 = -\log p_1(U_1)$, $\sigma_i = -\log P(U_{i-1},U_i)$ $(i \geq 2)$

Then $\xi_n = \frac{1}{n}\sum_{j=1}^{n}\sigma_j$. This shows that $\xi_n \xrightarrow{\mathbb{P}} \gamma$ is a kind of LLNs.

By Chebyshov, $\mathbb{P}(|\xi_n - H| \geq \varepsilon) < \frac{1}{\varepsilon^2}\mathbb{E}((\xi_n - H)^2) = \cdots = \frac{1}{n^2\varepsilon^2}\mathbb{E}((\sum_{i=1}^{n}(\sigma_i - H))^2)$, and the theorem follows if you can prove that $\mathbb{E}((\Sigma(\sigma_i - H))^2) \leq Cn$, then RHS will be $\leq \frac{C}{n\varepsilon^2} \to 0$ as $n\to\infty$

Now, $\mathbb{E}((\Sigma(\sigma_i - H))^2) = \sum_{i=1}^{n}\mathbb{E}((\sigma_i - H)^2) + 2\sum_{1\leq i\leq j\leq n}\mathbb{E}[(\sigma_i - H)(\sigma_j - H)]$.

The first sum is $\leq c'n$ (some c') and so is okay.

The second is bounded by $2\sum_{i=1}^{n}[\sum_{i\leq j\leq n}|\mathbb{E}[(\sigma_i - H)(\sigma_j - H)]|]$ and the assertion follows, since $\sum_{j=i+1}^{n}|\mathbb{E}(\sigma_i - H)(\sigma_j - H)| \leq \frac{H + \log e}{\rho}$ (2).

To prove (2), compute (3): $\mathbb{E}((\sigma_i - H)(\sigma_j - H)) = \sum_{u,u',v,v'}\mathbb{P}(U_{i-1}=u, U_i = u', U_{j-1}=v, U_j = v')(-\log P(u,u') - H)(-\log P(v,v') - H)$
$= \sum_{u,u',v,v'}(p_1 P^{i-2})(u)P(u,u')P^{j-1-i}(u',v)P(v,v')(-\log P(u,u') - H)(-\log P(v,v') - H)$

Want to compare (3) with $(-\log P(u,u') - H)(-\log P(v,v') - H)$.

(4): $\sum_{u,u',v,v'}(p_1 P^{i-2})(u)P(u,u')w(v)P(v,v')(-\log P(u,u') - H)(-\log P(v,v') - H) = 0$.

[Reminder: an irreducible aperiodic Markov chain has a unique equilibrium distribution, and the chain converges to this invariant distribution.
I.e, $|w(v) - P^{(n)}(u,v)| \leq (1-\rho)^n$, where $\rho = \min_{u,v} P(u,v) > 0$ $[\in (0,1)]$]

So, applying Theorem 5.1, we have $|(3) - (4)| \leq (1-\rho)^{j-i-1}(H + |\log e|)^2$, and thus (2) is bounded by a geometric progression.

§2 - Channels.

The basic scheme: message source → coder → channel → decoder → destination.

A source emits $U_1, U_2, \ldots$ (random text). A segmenting code $f: u^{(n)} \mapsto x^{(N)}$. The code is known both to the sender and the receiver.

Definitions: A channel is subject to 'noise'. The conditional probability $\mathbb{P}_{ch}(\text{receive } y^{(N)} | x^{(N)} \text{ sent})$ describes its ~~the~~ performance. A memoryless channel: $\mathbb{P}_{ch}(y^{(N)}|x^{(N)}) = \prod_{j=1}^{N} p(y_j|x_j)$.
The 2x2 (for binary) matrix $P(y|x)$ is called a channel (probability) matrix.
If $P(1|0) = P(0|1) = p$, it has the form $\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$. The channel is then called symmetric (or a memoryless binary symmetric channel) and p is called the distortion (or error) probability.
A decoding rule, $\hat{f}: y^{(N)} \mapsto v^{(n)} \in I^n$ is a map taking $y^{(N)}$ to a string of length n.

We want to use such a decoding rule $\hat{f}$ that gives a small probability of errors: $\varepsilon = \sum_{u^{(n)}} \mathbb{P}(\hat{f}(y^{(n)}) \neq u^{(n)} \mid u^{(n)}$ is emitted by source$)$. More precisely, we want $\lim_{n\to\infty} \varepsilon = 0$.

Remarks: (i) if the source has the AEP, then the set of 'typical' strings has number $\sim 2^{n(H+o(1))}$. Thus, in the sum for ε, you can restrict to $2^{n(H+o(1))}$ strings and neglect the rest. Ie, the length N of the codeword may be taken $N \sim \lfloor nH \rfloor + 1$.

(ii) if we take $N \sim \lfloor \bar{R}^{-1} nH \rfloor + 1$, a bigger value of the codeword length, we may be able to introduce a redundancy in the code, and 'beat' errors in the channel.

Notation: $u^{(n)}$ - a source message; $f\ (= f^{(N)})$ a code, $f: u^{(n)} \mapsto x^{(N)}$; $x^{(N)}$ a codeword of length N; $\hat{f}\ (= \hat{f}^{(N)})$ a decoding rule, $\hat{f}: \{0,1\}^N \mapsto X_N$, the set of codewords. AEP: # of the $u^{(n)}$'s $\sim 2^{nN}$, $N > \lfloor nH \rfloor$. Try $N \sim \lfloor \bar{R}^{-1} nH \rfloor$, $\bar{R}^{-1} > 1$, ie $\bar{R} \in (0,1)$. Thus, $n \sim \frac{N\bar{R}}{H}$. N will be the main parameter.

Definition 7.1: $\bar{R} \in (0,1)$ is a reliable transmission rate if $\exists$ code f and decoding rule $\hat{f}$ such that, given that the source emits a set U_N of $2^{N(\bar{R}+o(1))}$ equiprobable strings, the error probability $\lim_{N\to\infty} \sum_{u \in U_N} \frac{1}{2^{N(\bar{R}+o(1))}} \sum_{y^{(N)}:\ \hat{f}^{(N)}(y^{(N)}) \neq f^{(N)}(u)} \mathbb{P}_{ch}(y^{(N)} \mid f^{(N)}(u) \text{ sent}) = 0$.

Definition 7.2: The channel capacity $C = \sup[\bar{R} \in (0,1):\ \bar{R}$ is a reliable transmission rate$]$.

In the case of a memoryless binary channel (m.b.c.), $C = \sup_{P_{X_n}} i(X_n, Y_n)$.
Here, $i(X_n, Y_n)$ is the mutual information between a single input/output pair of symbols, taken over all possible input-letter distributions P_{X_n}. If the source is stationary, the index n may be omitted. [Various useful formulae in the handouts - P.29].

Equiprobability in definition 7.1 - gives a worst case.

Theorem 7.5: Suppose a conditional probability $\mathbb{P}_{ch}(y \mid x \text{ sent})$ is fixed. Fix a set U of the source strings and assume that only $u \in U$ are emitted. Consider an arbitrary probability distribution on U and take the error probability minimized over all encoding and decoding rules. Then, this error probability is maximised by the equidistribution over U: $\varepsilon(\mathbb{P}) \leq \varepsilon(\mathbb{P}^{eq})$. $[\varepsilon(\mathbb{P}) = \inf_{f,\hat{f}} \varepsilon(\mathbb{P}, f, \hat{f}),\ \varepsilon(\mathbb{P}^{eq}) = \inf_{f,\hat{f}} \varepsilon(\mathbb{P}, f, \hat{f}).]$

Proof: First fix f and $\hat{f}$. Let $u \in U$ have probability $p(u)$. Set $\beta(u) = \sum_{y:\ \hat{f}(y) \neq f(u)} \mathbb{P}_{ch}(y \mid f(u))$, the conditional error probability. Then, $\varepsilon(\mathbb{P}, f, \hat{f}) = \sum_{u \in U} p(u)\beta(u)$. You can permute the codewords by using a permutation λ of degree $\#U$. (The number of such is $(\#U)!$) Then the overall error probability, $\varepsilon(\lambda) = \sum_{u \in U} p(u)\beta(\lambda u)$. If $\mathbb{P} = \mathbb{P}^{eq}$, $\varepsilon(\lambda) = \varepsilon(\mathbb{P}, f, \hat{f}) = \frac{1}{\#U}\sum \beta(u) =: \bar{\varepsilon}$.

Claim: for any $\mathbb{P}$, $\exists$ a permutation λ such that $\varepsilon(\lambda) \leq \bar{\varepsilon}$. Then, minimising over $f, \hat{f}$ will lead to the assertion of the theorem. Thus, it suffices to prove the claim. Take a random permutation Λ, equidistributed over the set with cardinality $(\#U)!$ Then, $\min_\lambda \varepsilon(\lambda) \leq \mathbb{E}\,\varepsilon(\Lambda) = \mathbb{E}\sum_{u \in U} p(u)\beta(\Lambda u) = \sum_{u \in U} p(u)\mathbb{E}\,\beta(\Lambda u)$
$= \underset{\substack{\uparrow \\ \text{equidistribution of } \Lambda.}}{\sum_{u \in U}} p(u) \frac{1}{(\#U)!}\sum_{\bar{u} \in U} \beta(\bar{u}) = \bar{\varepsilon} = \varepsilon(\mathbb{P}^{eq}, f, \hat{f})$. Thus, $\exists\ \lambda$ with the desired property.

Decoding Rules

There are two possible "good" rules:

(a) an ideal observer rule - used when the receiver knows the source distribution $\mathbb{P}(u)$.

(b) maximal likelihood rule - does not require knowledge of $\mathbb{P}(u)$.

In (a), maximise the posterior distribution; in (b), maximise the prior.

A code, $f: \mathcal{U} \to \mathcal{X}_N$, $\mathcal{U}$ a set of "typical" messages.

A decoding rule: $\hat{f}: \{0,1\}^N \to \mathcal{U}$. If f is 1-1 then $\hat{f}: \{0,1\}^N \to \mathcal{X}_N$ is (can be)

(i) ideal observer - observer decodes a word $y^{(N)} \in \{0,1\}^N$ by $x^{*(N)}$, where $x^{*(N)}$ maximises $\mathbb{P}(x^{(N)} \text{ sent} \mid y^{(N)} \text{ received}) = \frac{\mathbb{P}(x^{(N)})\mathbb{P}_{ch}(y^{(N)} \mid x^{(N)})}{\mathbb{P}_{Y_N}(y^{(N)})}$, where $\mathbb{P}_{Y_N}(y^{(N)}) = \sum_{\tilde{x}^{(N)} \in X} \mathbb{P}(\tilde{x}^{(N)})\mathbb{P}_{ch}(y^{(N)} \mid \tilde{x}^{(N)})$.
The receiver knows $\mathbb{P}_{ch}$. To apply the ideal observer rule, the observer has to know $\mathbb{P}(x^{(N)})$

(ii) Maximum likelihood - decodes $y^{(N)} \in \{0,1\}^N$ by $x_*^{(N)}$, where $x_*^{(N)}$ maximises $\mathbb{P}_{ch}(y^{(N)} \mid x^{(N)})$.

Theorem (a) For any 1-1 encoding rule f, the ideal observer decoder minimises the error probability.

(b) If the source distribution is uniform over $\mathcal{U}$, then the ideal observer and maximal likelihood coincide.

Proof: (a) The ideal observer maximises the quantity $\mathbb{P}(x^{(N)})\mathbb{P}_{ch}(y^{(N)} \mid x^{(N)})$.

$$\varepsilon = \sum_{u \in \mathcal{U}} \mathbb{P}(U=u)\,\mathbb{P}_{ch}(\hat{f}(y) \neq u \mid f(u)) = \sum_{x \in X} \mathbb{P}(x) \sum_{y: \hat{f}(y) \neq u} \mathbb{P}_{ch}(y \mid x) = \sum_{y \in \{0,1\}^N} \sum_{x: \hat{f}(y) \neq x} \mathbb{P}(x)\mathbb{P}_{ch}(y \mid x)$$

$$= \sum_{y \in \{0,1\}^N} \Big[\sum_x \mathbb{P}(x)\mathbb{P}_{ch}(y \mid x) - \sum_{x = \hat{f}(y)} \mathbb{P}(x)\mathbb{P}_{ch}(y \mid x) \Big]$$

$$= \sum_{y \in \{0,1\}^N} \Big[\sum_{x \in X} \mathbb{P}(x)\mathbb{P}_{ch}(y \mid x) \Big] - \sum_{y \in \{0,1\}^N} \mathbb{P}(\hat{f}(y))\,\mathbb{P}_{ch}(y \mid \hat{f}(y))$$

$$= \sum_{x \in X} \mathbb{P}(x) \underbrace{\sum_{y \in \{0,1\}^N} \mathbb{P}(y \mid x)}_{=1} - \sum_y \mathbb{P}(\hat{f}(y))\,\mathbb{P}_{ch}(y \mid \hat{f}(y)) = 1 - \underbrace{\sum_y \mathbb{P}(\hat{f}(y))\,\mathbb{P}_{ch}(y \mid \hat{f}(y))}_{\text{want to maximise this in order to minimise } \varepsilon.}$$

This sum is maximised when $\hat{f}$ is the ideal observer. Thus, $\varepsilon(\hat{f}) \geq \varepsilon(\text{id. obs.})$

In what follows, we use the maximum likelihood decoding rule; the encoding rule will be chosen according to circumstance.

Lemma: Let the source be equidistributed over $\mathcal{U}$ and assume that an encoding rule f is applied. Then $\varepsilon(f) \leq \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{u' \in \mathcal{U}, u' \neq u} \mathbb{P}\big(\mathbb{P}_{ch}(Y \mid f(u')) \geq \mathbb{P}_{ch}(Y \mid f(u)) \mid U = u\big)$.

Proof: Given that $U=u$ and the maximum likelihood decoder is applied, we have the following possibilities: (a), an error when $\mathbb{P}_{ch}(Y \mid f(u')) > \mathbb{P}_{ch}(Y \mid f(u))$ for some $u' \neq u$.

(b), possibly an error when $\mathbb{P}_{ch}(Y \mid f(u')) = \mathbb{P}_{ch}(Y \mid f(u))$ for some $u' \neq u$.

(c), no error when $\mathbb{P}_{ch}(Y \mid f(u')) < \mathbb{P}_{ch}(Y \mid f(u))$ $\forall u' \neq u$.

Thus, $\mathbb{P}(\text{error} \mid U=u) \leq \mathbb{P}(\mathbb{P}_{ch}(Y \mid f(u')) \geq \mathbb{P}_{ch}(Y \mid f(u)) \text{ some } u' \neq u \mid U=u) = \sum_{u' \in \mathcal{U}, u' \neq u} \mathbb{P}(\mathbb{P}_{ch}(Y \mid f(u')) \geq \mathbb{P}_{ch}(Y \mid f(u)) \mid U=u)$

Multiplying by $\frac{1}{|\mathcal{U}|} = \mathbb{P}(U=u)$, and summing u yields result.

Remark: A similar formula holds for a general $\mathbb{P}(U=u)$

Random codes: A *deterministic code* is a map $f: U \to X_N \in \{0,1\}^N$ – given $u \in U$, $f(u)$ is uniquely determined. A *random code* is a map F such that $F(u)$ is a random string for each $u \in U$, from $u \in U$, from $\{0,1\}^N$.
An advantage of random coding is that it is sometimes easy to calculate $E := \mathbb{E}(\varepsilon)$. A disadvantage of random coding is that the chance of case (b) above occurring increases, hence the chance of an error increases. However, as $\exists$ a deterministic code f with $\varepsilon(f) \leq \mathbb{E}(\varepsilon)$, if we manage to prove that $E \to 0$ as $N \to \infty$, then we can guarantee $\exists f$ such that $\varepsilon(f) \to 0$ as $N \to \infty$.

An example of random coding: $F(u^{(1)}), \ldots, F(u^{(s)})$ are iid, and in a codeword $F(u^{(j)})$, symbols are iid, ie. $F(u^{(j)}) = W_1, \ldots, W_N$, $W_i \in \{0,1\}$, iid.

Theorem: (a) $\exists$ a deterministic f such that $\varepsilon(f) \leq \mathbb{E}(\varepsilon(F))$
(b) $\mathbb{P}(\varepsilon(F) \leq \frac{E}{1-\rho}) \leq \rho$ for any $\rho \in (0,1)$.

Proof: (a) Trivial.
(b) by Chebyshev inequality.

Definition: Given random words $x^{(N)}$ (channel input) and $y^{(N)}$ (channel output), define
$C_N = \sup_{P_X^{(N)}} \frac{1}{N} i(X^{(N)}, Y^{(N)})$; sup taken over all possible probability distributions $P_X^{(N)}$

Theorem (Shannon's S.C.T.): converse The channel capacity obeys $C \leq \overline{\lim_{N\to\infty}} C_N$.
Proof: Let's fix a code $f (= f_N): U_N \to X_N \subseteq \{0,1\}^N$, $\# U_N = 2^{N(\bar{R} + o(1))}$. We'll check that $\forall$ decoding rules $\hat{f}$, $\varepsilon(f) \geq 1 - \frac{C_N + o(1)}{\bar{R} + o(1)}$. The result will follow from this bound, because $\underline{\lim}_{N\to\infty} \varepsilon(f) \geq 1 - \frac{1}{\bar{R}} \overline{\lim_{N\to\infty}} C_N$. This is > 0 when $\bar{R} > \overline{\lim_{N\to\infty}} C_N$.
Assume that f is 1-1. (otherwise you would increase $\varepsilon(f)$). As U is assumed to be equidistributed over U_N, the codeword $f(u)$ is equidistributed over X_N. Thus, denoting $r = \# U_N$, you can write, for a given decoding rule $\hat{f}$:

$$NC_N \geq i(X^{(N)}, Y^{(N)}) \geq i(X^{(N)}, \hat{f}(Y^{(N)})) \quad \text{(by Theorem 3.8)}.$$
$$= h(X^{(N)}) - h(X^{(N)} \mid \hat{f}(Y^{(N)})) = r - h(X^{(N)} \mid \hat{f}(Y^{(N)})) \geq \log r - 1 - \varepsilon(f) \log(r-1) \quad \text{(by Theorem 3.6)}$$

In fact, the error probability is $\varepsilon(f) = \sum_{i=1}^{r} \mathbb{P}(X^{(N)} = x_i^{(N)}, \hat{f}(Y^{(N)}) \neq x_i^{(N)})$,
and by the generalised Fano inequality, $h(X^{(N)} \mid \hat{f}(y^{(N)})) \leq g(\varepsilon(f)) + \varepsilon(f) \log(r-1)$.
$$\leq 1 + \varepsilon(f) \log(r-1).$$

Now, from $NC_N \geq \log r - 1 - \varepsilon(f) \log(r-1)$, you conclude that
$$NC_N \geq N(\bar{R} + o(1)) - 1 - \varepsilon(f) \log(2^{N(\bar{R} + o(1))} - 1)$$
So, $$\varepsilon(f) \geq \frac{N(\bar{R} + o(1)) - NC_N - 1}{\log(2^{N(\bar{R} + o(1))} - 1)} = 1 - \frac{C_N + o(1)}{\bar{R} + o(1)}.$$

Theorem (Shannon's SCT), direct: Assume that $\exists$ a constant $c \in (0,1)$ such that $\forall \bar{R} \in (0,c)$ and $\forall N$ $\exists$ a random coding $F(u_1), \ldots, F(u_r)$, $r = 2^{N(\bar{R} + o(1))}$ with iid codewords and such that $\eta_N := \frac{1}{N} \log \frac{P(X^{(N)}, Y^{(N)})}{P_X(X^{(N)}) P_Y(Y^{(N)})} \xrightarrow{\mathbb{P}} c$. Then $C \geq c$.

Proof: Next lecture.

Corollary: $\sup c \le C \le \varlimsup_{N\to\infty} C_N$. Thus, if both quantities coincide, this gives the value of C.

Consider the example of an m.b.c., $\mathbb{P}_{ch}(y^{(N)}|x^{(N)}) = \prod_{i=1}^{N} p(y_i|x_i)$.

Theorem: For this, $i(X^{(N)}, Y^{(N)}) \le \sum_{j=1}^{N} i(X_j, Y_j)$, with equality if $X_1, .., X_N$ are independent.
Proof: The conditional entropy, $h(Y^{(N)}|X^{(N)}) = \sum_{j=1}^{N} h(Y_j|X_j)$ and
$$i(X^{(N)}, Y^{(N)}) = h(Y^{(N)}) - h(Y^{(N)}|X^{(N)}) = h(Y^{(N)}) - \sum_{j=1}^{N} h(Y_j|X_j) \le \sum_{j=1}^{N}\big(h(Y_j) - h(Y_j|X_j)\big) = \sum_{j=1}^{N} i(Y_j, X_j).$$
The "=" iff the Y's are independent. But they are if the X's are independent.

Theorem: For the m.b.c., $C \le \sup_{P_{X_1}} i(X_1, Y_1)$
Proof: $NC_N = \sup i(X^{(N)}, Y^{(N)}) \le \sum_{j=1}^{N} \sup i(X_j, Y_j) = N \sup i(X_1, Y_1)$.
Thus, $C \le \varlimsup_{N\to\infty} C_N \le \sup_{P_{X_1}} i(X_1, Y_1)$.

On the other hand, take a random code F, with codewords $F(u_1), .., F(u_r)$ where $F(u_k) = V_{k1} .. V_{kN}$ with iid digits V_{kj} distributed according to p_{max}, the distribution that maximises $i(X_1, Y_1)$. For this random code,
$$\eta_N = \frac{1}{N}\log\frac{p(X^{(N)}, Y^{(N)})}{p_X(X^{(N)})p_Y(Y^{(N)})} = \frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{p(X_j, Y_j)}{p_{max}(X_j)p_Y(Y_j)}\right) = \frac{1}{N}\sum_{j=1}^{N}\zeta_j \quad \text{where } \zeta_j := \log\frac{p(X_j, Y_j)}{p_{max}(X_j)p_Y(Y_j)}.$$

The rvs $\zeta_1, .., \zeta_N$ are iid, with $\mathbb{E}\zeta_j = i(X_j, Y_j)$
By LLN's, $\eta_N \xrightarrow{\mathbb{P}} i_{max}(X_j, Y_j)$. Thus, for the m.b.c., $C = i_{max}(X_1, Y_1) = \sup_{P_{X_1}} i(X_1, Y_1)$.
For m.b.s.c. $\sim \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$, $C = 1 - h(p, 1-p)$.

Proof of Shannon's SCT direct: The main step is the following lemma:

Lemma 1: Take a random code F, with iid codewords $F(u_1), .., F(u_r)$, $r = 2^{N(\bar{R}+o(1))}$ and with $p_F(v) = \mathbb{P}(F(u) = v)$. Then $\forall t > 0$, under the maximal likelihood decoding rule, $E = \mathbb{E}\,\varepsilon(F) \le \mathbb{P}(\eta_N \le t) + r2^{-Nt}$.

It is easy to deduce the assertion of the SCT from lemma 1. Take $\bar{R} = c - 2\varepsilon$ and $t = c - \varepsilon$. Then by lemma 1, $E \le \mathbb{P}(\eta \le c - \varepsilon) + 2^{N(c - 2\varepsilon + c + \varepsilon + o(1))} = \mathbb{P}(\eta_N \le c - \varepsilon) + 2^{-N\varepsilon}$
$\downarrow$ 0 as $\eta_N \xrightarrow{\mathbb{P}} c$, $\to 0$ as $N \to \infty$.

Thus by theorem 8.4 (i), $\exists$ a sequence of encoding rules f_N such that $\lim_{N\to\infty} f_N = 0$. $\square$

Proof of Lemma 1: Set $\delta(f, u, y) = \begin{cases} 1 & \text{if } f(u') \in S_y(f(u)) \text{ for some } u' \ne u. \\ 0 & \text{otherwise} \end{cases}$
where $S_y(x) = \{x' \in \{0,1\}^N : \mathbb{P}_{ch}(y|x') \ge \mathbb{P}_{ch}(y|x)\}$.
Then, $\forall$ deterministic codes f, $\varepsilon(f) \le \mathbb{E}\,\delta(f, U, V)$ [U - random message, V - random codeword]
and $\forall$ random codes F, $\varepsilon(F) \le \mathbb{E}\,\delta(F, U, V)$ [$\mathbb{P}(A) = \mathbb{E}\,\mathbb{1}_A$.]
For the random code F with iid codewords, $\mathbb{E}\,\delta(F, U, V) = \mathbb{E}\big(1 - \prod_{i=1}^{r-1}(1 - \mathbb{1}_{\{V_i \in S_Y(X)\}})\big)$
because $\delta(f, u, y) = 1 - \prod_{u' \ne y} \mathbb{1}_{\{f(u') \notin S_y(f(u))\}} = 1 - \prod_{u' \ne y}\big(1 - \mathbb{1}_{\{f(u') \in S_y(f(u))\}}\big)$.

Now, $\mathbb{E}\big(1 - \prod_{i=1}^{r-1}(1 - \mathbb{1}_{\{V_i \in S_Y(X)\}})\big) = \sum_x p_X(x)\,\mathbb{P}_{ch}(y|x).\,\mathbb{E}\big(1 - \prod_{i=1}^{r-1}(1 - \mathbb{1}_{\{V_i \in S_y(x)\}}) \,\big|\, X = x, Y = y\big) \ldots$

Lemma 2: For the random code F as indicated in lemma 1, if you define $V_1, \ldots, V_{r-1}$ by: if $U = u_j$, then $V_i = \begin{cases} F(u_i) & \text{for } i<j \text{ (if any)} \\ F(u_{i+1}) & \text{for } i \geq j \text{ (if any)} \end{cases}$, $j = 1, \ldots, r-1$.

Then, U (the message emitted), $X = F(U)$, (a random codeword) and $V_1, \ldots, V_{r-1}$ are independent, and $X, V_1, \ldots, V_{r-1}$ are iid, with distribution $p_F(v) = \mathbb{P}(F(u) = v)$.

Proof: Write $\mathbb{P}(U = u_j, X = x, V_1 = v_1, \ldots, V_{r-1} = v_{r-1}) = \mathbb{P}\left(U = u_j, \begin{pmatrix} F(u_1) \\ \vdots \\ F(u_{j-1}) \\ F(u_j) \\ F(u_{j+1}) \\ \vdots \\ F(u_r) \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_{j-1} \\ x \\ v_j \\ \vdots \\ v_{r-1} \end{pmatrix}\right)$

$= \mathbb{P}_{source}(U = u_j) p_F(x) p_F(v_1) \ldots p_F(v_{r-1})$. Done.

Return to proof of lemma 1:

$$\ldots = \sum_x P_X(x) \mathbb{P}_{ch}(y|x) \left(1 - \prod_{i=1}^{r-1} \mathbb{E}\left(1 - \mathbb{1}_{\{V_i \in S_y(x)\}}\right)\right) = \sum_x P_X(x) \mathbb{P}_{ch}(y|x) \left(1 - (1 - Q_y(x))^{r-1}\right).$$

where $Q_y(x) = \sum_{x' \in S_y(x)} P_X(x')$.

As the result, we have $E \leq 1 - \mathbb{E}(1 - Q_y(X))^{r-1}$.

Denote by $\mathbb{T}$ $(= \mathbb{T}(y))$ the set of pairs (x,y) for which $\frac{1}{N} \log \frac{P_X(x,y)}{P_X(x) P_Y(y)} > t$

Then write the bounds, $1 - (1 - Q_y(x))^{r-1} = \sum_{j=0}^{r-2} (1 - Q_y(x))^j Q_y(x) \leq (r-1) Q_y(x)$ if $(x,y) \in \mathbb{T}$.

and $1 - (1 - Q_y(x))^{r-1} \leq 1$ when $x, y \notin \mathbb{T}$.

This yields $E \leq \mathbb{P}((X,Y) \notin \mathbb{T}) + (r-1) \sum_{(x,y) \in \mathbb{T}} P_X(x) \mathbb{P}_{ch}(y|x) Q_y(x)$

Now, $\mathbb{P}((X,Y) \in \mathbb{T}) \leq \mathbb{P}(\eta_N \leq t)$ and for $x' \in S_y(x)$, $\mathbb{P}_{ch}(y|x') \geq \mathbb{P}_{ch}(y|x) \geq p_Y(y) 2^{Nt}$.

Multiplying by $\frac{P_X(x')}{P_Y(y)}$ gives $\mathbb{P}(X = x' | Y = y) \geq P_X(x') 2^{Nt}$.

Finally you sum over $x' \in S_y(x)$ and get $1 \geq \mathbb{P}(S_y(x) | Y = y) \geq Q_y(x) 2^{Nt}$

$\therefore Q_y(x) \leq 2^{-Nt}$.

This completes the proof of lemma 1, and hence that of Theorem 9.3 (SCT)

Recall: m.b.c. $C = \sup_{P_X} i(\overset{\text{input}}{X}, \overset{\text{output}}{Y})$. Mbs.c. $C = 1 - h(p, 1-p)$.

The formulas for the channel capacity were established for $a = 2$ (ie $J = \{0, 1\}$).

Many features of the theory remain true in a general case, when Y may take values $\{0, \ldots, t\}$. The memoryless property is defined similarly.

A m.c. is called symmetric if the rows of the channel matrix are permutations of each other, and double symmetric if both the rows and columns are permutations of each other.

Theorem: For a m.s.c., $C \leq \log(t+1) - h(Y_1|X_1)$, $[h(Y_1|X_1) = h(p_0, \ldots, p_t)]$

and in the case of a double symmetric channel, $C = \log(t+1) - h(p_0, \ldots, p_t)$

Proof: By repeating the proof given for $a = 2$, obtain that $C = \sup_{P_X} i(X, Y)$, and

$i(X, Y) = h(Y) - h(Y|X) \leq \log(t+1) - h(Y|X)$.

Now, $h(Y|X) = -\sum_{x,y} \mathbb{P}(X = x) P_{ch}(y|x) \log P_{ch}(y|x) = -\sum_x \mathbb{P}(X = x) \sum_y P(y|x) \log P(y|x)$

$= \sum_x \mathbb{P}(X = x) h(p_0, \ldots, p_t) = h(p_0, \ldots, p_t)$

Assuming that the channel is double symmetric, we have
$\mathbb{P}(Y=y) = \sum_x \mathbb{P}(X=x)\, P(y|x)$, taking P_X equidistributed,
$= \frac{1}{2} \sum_x P(y|x)$, which does not depend on y because of the double symmetry. Hence, $\mathbb{P}(Y=y)$ does not depend on y, so $\mathbb{P}(Y=y) = \frac{1}{t+1}$.

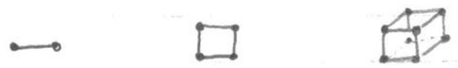
<u>Note</u>: you can think of an arbitrary input or output alphabet; the statements of the main theorems remain true.

In the case of a m.b.s.c., with row-error probability p,
$\mathbb{P}_{ch}(y^{(N)}|x^{(N)}) = p^{d(x^{(N)},y^{(N)})} \cdot (1-p)^{N-d(x^{(N)},y^{(N)})} = (1-p)^N \cdot \left(\frac{p}{1-p}\right)^{d(x^{(N)},y^{(N)})}$.
If $0<p<\frac{1}{2}$, then $\frac{p}{1-p}<1$, and the maximum likelihood decoder wants to minimise $d(x^{(N)},y^{(N)})$. Here, $d(x^{(N)},y^{(N)}) = \#$ distinct digits in $x^{(N)}, y^{(N)}$.
$d(x^{(N)},y^{(N)})$ is a <u>metric</u> in the space $\{0,1\}^N$

<u>Proof</u>: $d \geq 0$ is obvious, as is $d=0 \iff x^{(N)}=y^{(N)}$, as is symmetry.
The triangle inequality: $d(x^{(N)},z^{(N)}) \leq d(x^{(N)},y^{(N)}) + d(y^{(N)},z^{(N)})$. Obvious also.

$\{0,1\}^N$ is the <u>Hamming space</u> of "length" N. $d(x^{(N)},y^{(N)})$ is called the <u>Hamming distance</u>. $\{0,1\}^N \sim$ the collection of the vertices of a unit cube in $\mathbb{R}^N$.

<u>Eg</u>:



N=1 N=2 N=3

The Hamming space is a <u>group</u> wrt the component wise addition mod 2:
$x^{(N)}+y^{(N)} \pmod 2 = x_1+y_1 \bmod 2 \ldots x_N+y_N \bmod 2$.
It is also a <u>vector space</u> [linear] with binary coefficients. Eg; $\lambda x^{(N)} \in \{0,1\}^N$ $(\lambda = 0,1)$, and $\lambda(x^{(N)}+y^{(N)}) = \lambda x^{(N)} + \lambda y^{(N)}$.

<u>Lemma 11.4</u>: The Hamming distance is preserved under group translations.
Ie, $d(x^{(N)}+z^{(N)}, y^{(N)}+z^{(N)}) = d(x^{(N)},y^{(N)})$.

In geometrical terms, for $0<p<\frac{1}{2}$, the maximum likelihood decoder wants to find a codeword $x_*^{(N)}$ that is closest to $y^{(N)}$, the received word. In algebraic terms, we represent $y^{(N)} = x^{(N)} + e^{(N)}$, where $e^{(N)}$ is an error vector. You want to find $e^{(N)}$ such that $x^{(N)}$ is a codeword and $e^{(N)}$ contains a minimal number of 1's.

<u>Recall</u>: a code, f_N, was a map $f_N : \mathcal{U} \to X_N \subset \{0,1\}^N$. If f_N is 1-1, then it may be identified with X_N.

So, from now on, a code is understood as a <u>set</u> $X_N \subset \{0,1\}^N$, known to the receiver.
The Shannon SCT does not produce an example of a deterministic code for which $\varepsilon \to 0$ as $N \to \infty$. It only guarantees its existence.

§3. Coding Theory

Definition: X_N is called an N-code, or a code of length N. $\#X_N = r$ is called the size of the code. $\rho = \frac{\log r}{N}$ is the transmission rate.
A code X_N is called D-error detecting if changing up to D digits in any codeword does not produce another codeword. It is called E-error correcting if changing up to E digits does not produce a word that is within distance $\leq E$ of another codeword. The minimal distance of a code X is $\delta = \min\{ d(x^{(N)}, x^{(N)\prime}) : x^{(N)}, x^{(N)\prime} \in X, x^{(N)} \neq x^{(N)\prime} \}$.

Theorem 1: (a) X is D-error detecting iff $\delta \geq D+1$.
(b) X is E-error correcting iff the balls of radius E about the codewords are pairwise disjoint.

Proof: (a) Obvious.
(b) if the E-balls are disjoint then making up to E changes you are still closer to the original codeword than to any other one. Conversely, if X is E-error correcting then any word obtained by $\leq E$ changes falls in exactly one ball, hence the E-balls are disjoint.

Remark: If X detects D errors and D is even then X corrects $D/2$ errors.
If D is odd then it corrects $\frac{D-1}{2}$ errors.
The volume of an R-ball about $z^{(N)} \in \{0,1\}^N$ is $v_N(R) = \sum_{i=0}^{R} \binom{N}{i}$.

Theorem 2 (The Hamming Bound): Any E-error correcting code obeys $r \leq \frac{2^N}{v_N(E)}$

Proof: The E-balls about the codewords must be disjoint. Altogether they contain $r v_N(E)$ words. These must be within $\{0,1\}^N$, hence $r v_N(E) \leq 2^N$.

Definition: An E-error correcting code X with $\#X = r$ is called perfect if $r = \frac{2^N}{v_N(E)}$. I.e., every word belongs to exactly one E-ball. That is, you are never stuck while decoding.

There are quite few perfect codes. See the notes.

Theorem 3 (The Gilbert-Varshamov bound): $\exists$ a code X of minimal distance δ such that $r \geq \frac{2^N}{v_N(\delta-1)}$.

Proof: Take a code of maximum size with a given δ. Then, all $y^{(N)} \in \{0,1\}^N$ must be within distance $\leq \delta-1$ from the codewords. Thus the $(\delta-1)$-balls cover the whole space, hence $r v_N(\delta-1) \geq 2^N$.

Theorem 4 (The Singleton bound): $\forall$ codes X_N of minimal distance δ, $r \leq 2^{N-\delta+1}$

Proof: Use the "truncation" procedure. That is, delete the last digit from any codeword. Then, you obtain a code of length N-1 and minimum distance $\geq \delta-1$. If $\delta > 1$, the size of the code is preserved. You can continue this procedure $\delta-1$ times. The resulting codes should fit the corresponding spaces. Thus, $r \leq 2^{N-\delta+1}$.

Corollary: If $r^*(N,\delta)$ is the maximal size of a code of length N with minimal distance then $\frac{2^N}{v_N(\delta-1)} \le r^*(N,\delta) \le \min\left[\frac{2^N}{v_N(\lfloor \delta/2 \rfloor)}, 2^{N-\delta+1}\right]$.

The Hamming and Singleton bounds become too rough when $\delta \sim \frac{N}{2}$. [In general, the most interesting domain is where $\delta \sim \alpha N$ (a linear fraction of errors is detected and corrected)]. See notes.

The Plotkin bounds.

Theorem A1: $\forall$ codes X with minimal distance $\delta > N/2$, $r \le 2\left[\frac{\delta}{2\delta - N}\right]$

Theorem A2: If $r^*(N,\delta)$ is as before, then $r^*(N, 2l-1) = r^*(N+1, 2l)$, and $r^*(N-1, l) = \frac{1}{2} r^*(N, l)$

Theorem A3: $\forall N$ and even l with $l > N/2$, $\forall$ codes of minimal distance l, $r^* \le 2\left[\frac{l}{2l-N}\right]$. For a code of maximum size, $r^*(2l, l) \le 4l$.
If l is odd and $l > \frac{N-1}{2}$, then $r^*(N,l) \le 2\left[\frac{l+1}{2l+1-N}\right]$ and $r^*(2l+1, l) \le 4l+4$.

Proofs: A1: For a code of minimal distance δ you have $r(r-1)\delta \le 2\sum_{x,x' \in X} d(x,x') = \sum_{x\in X}\sum_{x'\in X} d(x,x')$.
On the other hand, you can write X in the form of an $(r\times N)$ matrix, by listing the codewords as rows. If column i in this matrix contains s_i zeroes and $r-s_i$ ones, then $\sum\sum d(x,x') \le 2\sum_{i=1}^{N} s_i(r-s_i)$.
If r is even, the rhs is maximised when $s_i = \frac{r_i}{2}$. This yields $r(r-1)\delta \le \frac{1}{2}Nr^2$, so $r \le \frac{2\delta}{2\delta - N}$. As r is even, this gives $r \le 2\left[\frac{\delta}{2\delta-N}\right]$.
If r is odd, then $r(r-1)\delta \le N\frac{(r^2-1)}{2}$.

Lemma 13.1: Let $\lambda \in (0, \frac{1}{2})$. Then $\lim \frac{1}{N}\log v_N([\lambda N]) = h(\lambda, 1-\lambda)$ [$= G(\lambda) = -\lambda \log\lambda - (1-\lambda)\log(1-\lambda)$].
Proof: Write $v_N(R) = \sum_{i=0}^{R}\binom{N}{i}$, $R = [\lambda N]$. The maximal term is the last one.
$\frac{\binom{N}{i+1}}{\binom{N}{i}} = \frac{N-i}{i+1} \geq 1$, as $R \le N/2$. Hence $\binom{N}{R} \le v_N(R) \le (R+1)\binom{N}{R}$.

By Stirling, $N! \sim N^{N+\frac{1}{2}}e^{-N}\sqrt{2\pi}$, and $\log\binom{N}{R} = -R\log\frac{R}{N} - (N-R)\log(1-\frac{R}{N}) + O(\log N)$
and $\frac{1}{N}\log v_N(R) \le \frac{1}{N}\log(R+1) + (1-\frac{R}{N})\log(1-\frac{R}{N}) - \frac{R}{N}\log\frac{R}{N} + O(\log N)$.
Similar lower bound holds, too. Then, using $\frac{R}{N}\to\lambda$ yields the result.

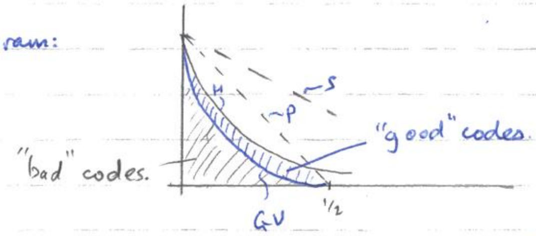
Denote by $r^*(N, [\lambda N])$ the maximal size of a code of length N, and minimal distance $[\lambda N]$, and $\alpha(\lambda) = \lim_{N\to\infty}\frac{1}{N}\log r^*(N, [\lambda N])$.

Theorem 2: (a) $\alpha(\lambda) \le 1 - G(\lambda/2)$ (Hamming).
(b) $\alpha(\lambda) \le 1-\lambda$
(c) $\alpha(\lambda) \geq 1 - G(\lambda)$

Proof: Straightforward

The asymptotical Plotkin bound: $\alpha(\lambda) \leq 1 - 2\lambda$.

The diagram:



Good codes were "produced" recently for $a = b^2 \geq 49$.

Linear Codes.

Definition: A code X is said to be linear if, together with x and y it contains $x+y \pmod 2$.

To identify a linear code, only have to fix a basis, ie a maximal set of linearly independent codewords. The cardinality of such a set is called the rank of the code. A linear code of length N and rank k is called an (N,k)-code.

Lemma 3: Any (N,k)-code contains 2^k codewords.

Proof: A codeword <=> a linear combination of vectors from a basis <=> a sum of vectors from a basis, and there 2^k of these.

An (N,k)-code is identified with a $k \times N$ matrix: $G = \begin{pmatrix} g_{11} & \cdots & g_{1N} \\ & & \vdots \\ g_{k1} & \cdots & g_{kN} \end{pmatrix} \Big\}$ basis, the generating matrix.

An (N,k)-code may also be described in terms of a parity-check matrix.

$X = \{x : xH = 0\}$, H is a parity-check matrix.

Example: The Hamming $(7,4)$-code. The parity-check matrix H is 7×3; its rows are all non-zero binary words of length 3. In the lexicographic order,

$$H^{lex} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

The corresponding generating matrix (one of them) is in the notes. [By permuting the rows/columns of a generating or parity-check matrix produces an equivalent code].

A particular form of G and H is: $G^{can} = (G' \; I_k)$, $H^{can} = \begin{pmatrix} I_{N-k} \\ H' \end{pmatrix}$.

The canonical form is convenient because you may write the result of encoding a binary word v of length r as $vG = (\underbrace{v.G'}_{\text{parity-check bit}}, \underbrace{v}_{\text{information bit}})$.

Definition: The weight of a binary word x is $w(x)$ = # non-zero digits in x.

<u>Theorem 13.5</u>: (i) The minimal distance of a linear code = the minimal weight of a non-zero codeword.

(ii) The minimal distance of a linear code = the minimal number of linearly dependent rows of the parity-check matrix.

<u>Proof</u>: (i) $d(x,y) = d(x+y, 0) = w(x+y)$. As $x+y \in X$, minimal distance $\leq$ minimal weight. $\geq$ similarly.

(ii) Let X have parity-check matrix H. Let the minimal distance of X be δ. Then $\exists$ a codeword $x \in X$ with $w(x) = \delta$. The equality $xH = 0$ means that the sum of δ rows of H gives zero. So we have δ linearly ~~in~~dependent rows of H. Assume there are $\delta-1$ linearly dependent rows of H. Their sum is zero. Thus, $\exists$ a vector x with $w(x) = \delta - 1$ with $xH = 0$. Hence $x \in X$ #.

<u>Theorem 14.1</u>: The Hamming $(7,4)$ has minimum distance 3, i.e. it detects 2 errors and corrects 1.

<u>Proof</u>: No pairs of rows of H^{lex} are linearly dependent. $\exists$ linearly dependent triples of rows: $\forall\, h, h'$, rows of H, you can add their sum. The triplet $h, h', h+h'$ is linearly dependent.

<u>Theorem 14.2</u>: The Hamming $(7,4)$ is a perfect 1-error correcting code.

<u>Proof</u>: $v_7(1) = 1+7 = 8 = 2^3$. The size is 2^4, and $2^4 \cdot 2^3 = 2^7$.

A general construction: take $N = 2^\ell - 1$ and $k = N - \ell = 2^\ell - 1 - \ell$. Take the matrix formed by all non-zero words of length ℓ. $H^{can} = \begin{pmatrix} 1\,0 \cdots 0 \\ 0\,1\,0 \cdots 0 \\ 0 \cdots\cdots 1 \\ H' \end{pmatrix} = \underbrace{\begin{pmatrix} I_\ell \\ H' \end{pmatrix}}_{\ell} \Big\} 2^\ell - 1$.

The code with the parity-check matrix H is called the <u>Hamming $(2^\ell - 1, 2^\ell - 1 - \ell)$-code</u>.

<u>Theorem 14.3</u>: The Hamming $(2^\ell - 1, 2^\ell - 1 - \ell)$-code has minimal distance 3. It detects 2 errors and corrects 1, and is a perfect 1-error correcting code of length $2^\ell - 1$.

<u>Proof</u>: First part - as Theorem 14.1.
The volume $v_{2^\ell - 1}(1) = 1 + 2^\ell - 1 = 2^\ell$. Size $\times$ volume $= 2^{2^\ell - 1 - \ell} \times 2^\ell = 2^{2^\ell - 1} = 2^N$

<u>Syndrome decoding</u>

<u>Theorem 14.4</u>: $\forall$ linear codes X $\exists$ an equivalent (isomorphic) code X' with the generating and parity-check matrices in a canonical form: $G^{can} = (G' \; I_k)$, $H^{can} = \begin{pmatrix} I_{n-k} \\ H' \end{pmatrix}$, with $G' = H'$.

<u>Proof</u>: A standard procedure - see the notes.

<u>Definition</u>: Let X be a linear code. If $u = u_1 \ldots u_N \in \{0,1\}^N$ then the <u>coset</u> $u + X$ is the set of words of the form $u + x$, $x \in X$.

<u>Theorem 14.5:</u> (i) If $u \in v \in X$ then $v \in u+X$, ie each word in a coset determines this coset.

(ii) $u \in u+X$.

(iii) u, v are in the same coset iff $u+v \in X$ [$v = u+x$, $u+v = u+u+x = x \in X$].

(iv) each word u belongs to a single coset [ie, the cosets form a partition of $\{0,1\}^N$].

(v) all cosets have the same number of words in them, equal to $\#X = 2^k$.

Altogether there are 2^{N-k} distinct cosets. X is a coset of any of the codewords.

(vi) the coset of $u+v$ is the set of words of the form $x+y$, $x \in v+X$, $y \in u+X$.

<u>Proof:</u> exercise from linear algebra and set theory.

Now, syndrome decoding: upon receiving the word y you find the coset of y. Then you take a <u>leader</u> of this coset, ie, a word of minimal weight. You decode y by $x = y+u \in X$. A drawback of this procedure is that the leader is not always unique. However, we have:

<u>Theorem 14.6:</u> The word x minimises the distance $d(y, x')$ over $x' \in X$.

<u>Proof:</u> $\forall x' \in X$, $d(y, x') = w(y+x') \geq \min_{v \in y+X} w(v) = w(u) = d(x,y)$.

<u>Theorem 14.7:</u> Cosets $u+X$ are in 1-1 correspondance with vectors of the form yH, ie, y and y' are in the same coset iff $yH = y'H$.

<u>Proof:</u> y, y' are in the same coset iff $y+y' \in X$. Ie, $0 = (y+y')H = yH + y'H$, ie $yH = y'H$.

Vectors of the form yH are called <u>syndromes</u>.

<u>Theorem 14.8:</u> For a Hamming code, $\forall$ syndromes $\exists$ a unique leader u, and u contains ≤ 1 non-zero digit. More precisely, if $yH = s$, a word of length l, then you decode y by y when $s = 0$, and by $y + e_i$ if s ~~is not~~ coincides with row i of H.

<u>Proof:</u> See the notes.

<u>Cyclic Codes.</u>

Polynomials with binary coefficients: $a = a_0 \cdots a_N \in \{0,1\}^{N+1} \leftrightarrow a_0 + a_1 X + \cdots + a_N X^N =: a(X)$, X a formal variable.

Addition and multiplication of the polynomials is as usual. The division - as in the case of "usual" polynomials (Euclid's algorithm).

<u>Examples:</u> $(1+X+X^3+X^4)(X+X^2+X^3) = X+X^7$

$1+X^N = (1+X)(1+X+\cdots+X^{N-1})$

$1+X^{2^l} = (1+X)^{2^l}$

<u>Theorem 1:</u> If $f(x)$ and $h(x)$ are two polynomials with $h \neq 0$ then $\exists$ unique polynomials $g(x)$ and $r(x)$ such that $\deg r(x) < \deg h(x)$ and $f(x) = g(x)h(x) + r(x)$.

<u>Proof:</u> If $\deg h(x) > \deg f(x)$, set $g(x) = 0$ and $r(x) = f(x)$. Otherwise perform the division algorithm.

g is called the quotient and r the remainder of f divided by h.

Definition: $f_1(X)$ is called equivalent $f_2(X)$ mod $h(X)$ if the remainders of $f_1(X)$ and $f_2(X)$ coincide. So $f_i(X) = g_i(X)h(X) + r(X)$. We write $f_1(X) = f_2(X) \bmod h(X)$.

Theorem 2: If $f_1(X) = f_2(X) \bmod h(X)$ and $p_1(X) = p_2(X) \bmod h(X)$ then
$f_1(X) + f_2(X) = p_1(X) + p_2(X) \bmod h(X)$, $f_1(X)f_2(X) = p_1(X)p_2(X) \bmod h(X)$.

Proof: Straightforward – see the notes.

Linear codes: a word of length N $\longleftrightarrow$ a polynomial of degree $N-1$
$a = a_0 \cdots a_{N-1}$ $\quad$ $a_0 + a_1X + \cdots + a_{N-1}X^{N-1}$.

$a(X) \in \mathcal{X}$ iff $a \in \mathcal{X}$. Ie, a linear code is closed under the addition of polynomials and multiplication by a 'scalar' ($= 0$ or 1).

For $a = a_0 \cdots a_{N-1}$, define the cyclic shift $\pi a = a_{N-1}a_0 \cdots a_{N-2}$.

Definition: $\mathcal{X}$ is called cyclic if, with any $a \in \mathcal{X}$, it contains πa.

Lemma 3: $\mathcal{X}$ is cyclic iff, $\forall$ vectors a from a basis, $\pi a \in \mathcal{X}$.

Proof: Each $u \in \mathcal{X}$ is a sum of vectors from the basis. As $\pi(u+v) = \pi(u) + \pi(v)$, the result follows.

Lemma 4: If $a \leftrightarrow a(X)$ then $\pi a \leftrightarrow Xa(X) \bmod (1+X^N)$

Proof: $Xa(X) = a_{N-1} + a_0X + a_1X^2 + \cdots + \underbrace{a_{N-1}X^N + a_{N-1}}_{= a_{N-1}(1+X^N)} = (\pi a)(X) + a_{N-1}(1+X^N)$.

Theorem 5: A cyclic code contains, with $a(X)$ and $b(X)$, the sum $a(X)+b(X)$, and $a(X)v(X) \bmod (1+X^N)$

Proof: The sum $\in \mathcal{X}$ by linearity. Write $v(X) = v_0 + v_1X + \cdots + v_{N-1}X^{N-1}$, and notice that $X^k a(X) \bmod (1+X^N) \in \mathcal{X}$ by lemma 4. Then $v(X)a(X) \bmod (1+X^N) = \sum_{i=0}^{N-1} v_iX^i a(X) \bmod (1+X^N) \in \mathcal{X}$.

Theorem 6: Let $c(X) = \sum_{i=0}^{N-k} c_iX^i$ be a non-zero polynomial of minimum degree from a cyclic code $\mathcal{X}$. Then (i) c is a unique polynomial of minimal degree.
(ii) $\mathcal{X}$ has rank k.
(iii) the codewords $c(X), Xc(X), \ldots, X^{k-1}c(X)$ form a basis in $\mathcal{X}$.
(iv) $a(X) \in \mathcal{X}$ iff $a(X) = v(X)c(X)$ for some $v(X)$ of degree $< k$.

Proof: (i) Let $c'(X) = \sum_{i=0}^{N-k} c_i'X^i$ be an arbitrary polynomial of minimal degree from $\mathcal{X}$. Then $c'_{N-k} = c_{N-k} = 1$. Thus $\deg[c(X)+c'(X)] < N-k$, the minimal degree. But $c(X)+c'(X) \in \mathcal{X}$, so $c(X)+c'(X) = 0$, so $c(X) = c'(X)$.

(ii) from (iii)

(iv) $\forall\, a(X) \in \mathcal{X}$, $\deg a(X) > \deg c(X)$, and by Theorem 1, $a(X) = v(X)c(X) + r(X)$, $\deg v(X) < k$, $\deg r(X) < N-k-\deg c(X)$. The product $v(X)c(X) \in \mathcal{X}$ by Theorem 5. Thus, $r(X) = a(X) + v(X)c(X) \in \mathcal{X}$. But then $r(X) = 0$.

(iv) $\Rightarrow$ (iii) By (iv), each $a(X) \in \mathcal{X}$ has the form $c(X)v(X) = \sum_{i=1}^{r} v_iX^ic(X)$, $r = \deg v(X) \le k-1$.
That is, each a is a linear combination of $c(X), Xc(X), \ldots, X^{k-1}c(X)$.

Corollary 1: All cyclic codes may be obtained from its polynomial of minimum degree by cyclic shifts and linear combinations. $c(X)$ is called the generator of a cyclic code.

Theorem 2: A polynomial ~~of~~ $c(X)$ of degree $\leq N-1$ is the generator of a cyclic code iff it divides $1+X^N$: $1+X^N = h(X)\,c(X)$.

Proof: By the division algorithm, $1+X^N = h(X)c(X) + r(X)$, $\deg r(X) < \deg c(X)$.
Ie, $r(X) = h(X)c(X) + 1 + X^N$, ie $r(X) = h(X)c(X) \mod (1+X^N)$.
By Theorem 5 above, $r(X) \in \mathcal{X}$, the cyclic code generated by $c(X)$. But $c(X)$ must be a unique polynomial of minimum degree in $\mathcal{X}$. $\therefore r(X) = 0$. This does "only if".
The "if" part is done similarly.

Examples: (i) $1+X^N = \underbrace{(1+X)}_{\text{parity-check code}}\underbrace{(1+X+\dots+X^{N-1})}_{\text{symbol-repetition code}}$ - cyclic.

(ii) The Hamming (7, 4) code: after permuting columns, the generating matrix takes the form $G^{cycl} = \begin{pmatrix} 1&1&0&1&0&0&0 \\ 0&1&1&0&1&0&0 \\ 0&0&1&1&0&1&0 \\ 0&0&0&1&1&0&1 \end{pmatrix}$ $1101 \sim 1+X+X^3$ - the generator.

So, the code is equivalent to a cyclic code with generator $1+X+X^3$.

Theorem 1: Any Hamming code is equivalent to a cyclic code.
Proof: omitted (for time!)

Encoding and decoding a cyclic code.

By Theorem 15.6, the basic of a cyclic code with generator $c(X)$ is formed by $c(X), Xc(X), \dots, X^{k-1}c(X)$, where $N-k = \deg c(X)$. The corresponding generating matrix is $G^{cycl} = \begin{pmatrix} c(X) \\ Xc(X) \\ \vdots \\ X^{k-1}c(X) \end{pmatrix}$
Then, given a word $a = a_0 \dots a_{k-1}$ (a source message), you encode it by $a(X)c(X) \in \mathcal{X}$.
To decode a code, you have to calculate the syndrome corresponding to the received word.

Theorem 2: The cosets $y + \mathcal{X}$ are in 1-1 correspondance with the remainders $u(X) = y(X) \mod c(X)$.
Proof: Two words y and y' belong to the same coset iff $y + y' \in \mathcal{X}$. Write $y(X) = g(X)c(X) + u(X)$, $y'(X) = g'(X)c(X) + u'(X)$. Then $y(X) + y'(X) = (g(X)+g'(X))c(X) + u(X) + u'(X)$.
This $\in \mathcal{X}$ iff $u(X) + u'(X) = 0$, ie, $u(X) = u'(X)$.

So, you can list all polynomials of degree $< \deg c = N-k$. These label the cosets. Still, you have to find a leader of a coset, and if it is non-unique, you have to peform an arbitrary choice or demand a retransmission.

BCH codes First, a summary of the theory of Hamming codes.

Theorem 3: The Hamming $(2^l-1, 2^l-1-l)$-code, with the parity-check matrix $H = \overbrace{\begin{pmatrix} \text{all} \neq 0 \text{ words} \\ \text{of length } l \end{pmatrix}}^{l} \Big\} 2^l-1$ is a perfect 1-error correcting code. To decode a word $y = y_1 \dots y_N$, $N = 2^l - 1$, you form the syndrome $s = yH$, $s = s_1 \dots s_l$. If $s = 0$, set $x_* = y$. If $s \neq 0$, then it coincides with a row of H, eg, $s = $ row i. Then you decode y by $x_* = y + e_i$, $e_i = 0 \dots 0 1 0 \dots 0$.

Suppose we want to construct a 2-error correcting code. Try a parity-check matrix of the form $\tilde{H} = (H, \Pi_H)$, where ΠH is obtained from H by permuting the rows (Π is a permutation of order 2^l-1). You obtain a $(2^l-1, 2^l-1-2l)$-code.
Then, a syndrome of a received word y will be a pair $y\tilde{H} = (s, s')$, s' a row from ΠH.
The idea is to choose Π in such a way that $\Pi s = s^{*q}$, where $*$ is a multiplication of words. Say $q = 3$ (a simplest choice). Then your task is: given a syndrome (s, s'), try to identify possible error-digits. You want your procedure to be correct if the numbers of errors is ≤ 2.

Conclusion: if $s = s' = 0$ you decode y by y.

if $s' = s^{*3}$ you decide that a single error occurred, at digit i, where i is the row of $\tilde{H}$ coinciding with (s, s^{*3}). Decode y by $y + e_i$.

if $s' \neq s^{*3}$, then you try to solve a pair of equations: $s_i + s_j = s$, $s_i^{*3} + s_j^{*3} = s'$ (s_i and s_j are rows of H). If you succeed (ie, if the solution is unique), then you decode y by $y + e_i + e_j$. That is, you decide that the errors occurred at places i and j.

Solving the last system is equivalent to solving the cubic equation: $s * z^{*2} - s^{*2} * z - s' = 0$. (see notes, p. 72). $z \in \{0,1\}^l$, $z \neq 0$, is the unknown. It is well-known that solving such an equation requires not only $*$-multiplication, but $*$-division. Ie, $\{0,1\}^l$ should be endowed with the structure of a field.
This is possible: $*$ must be multiplication mod an irreducible polynomial. Then the whole construction works, and you obtain a BCH code correcting ≤ 2 errors.
For the details, see the notes.