Coding & Cryptography

February 12, 2020

0 Communication Channels

This course will be about modelling communication. In general, we have the following idea:

$$\textsc{Source} \longrightarrow \textsc{Encoder} \xrightarrow[\text{Errors, Noise}]{\textsc{Channel}} \textsc{Decoder} \longrightarrow \textsc{Receiver}$$

For example, the channel might be an optical or electrical telegraph, modems, audio CDs, satellite relays. The encoding and decoding might be something like Ascii, so that each character in the email "Call at 2pm" would be encoded into 8 bits using $a = 1100101, \ldots$, giving an 84 bit message to be transmitted via the internet, and decoded by the receiver's email client. Our general aim here will be, given some source and channel (modelled probabilistically), to design an encoder and decoder to send messages economically and reliably.

Examples

- (Noiseless coding) Morse Code. In this code, more common letters are assigned shorter codes, so that we have $A = \cdot - *, E = \cdot *, Q = - - \cdot - *, Z = - - \cdot \cdot *$. This is adapted to the *source*, in the sense that we chose the codes based off the expected distribution of letters that we will have to transmit.

- (Noisy coding) Isbn. In the Isbn encoding, every book is given a 10 digit number $a_1 a_2 \ldots a_{10}$, with $\sum_{i=1}^{10} (11 - i) a_i \equiv 0 \mod 11$. This is adapted to the *channel*, in the sense that the likely errors to occur will be 1 incorrect digit, or accidentally transposing two digits, which this code is resistant to (will return an error rather than an erroneous result).

A ***communication channel*** accepts symbols from some alphabet $\mathscr{A} = \{a_1, a_2, \ldots, a_r\}$ (e.g. $\{0, 1\}, \{a, b, \ldots, z\}$), and outputs symbols from an alphabet $\mathscr{B} = \{b_1, \ldots, b_s\}$. The channel is modelled by the probabilities:

$$\mathbb{P}(y_1, y_2, \ldots, y_n \text{ received} | x_1, x_2, \ldots, x_n \text{ sent}) = \prod_{i=1}^{n} \mathbb{P}(y_i \text{ received} | x_i \text{ sent})$$

A ***discrete memoryless channel (DMC)*** is a channel with $p_{ij} = \mathbb{P}(b_j \text{ received} | a_i \text{ sent})$ the same for each channel usage and independent of any past or future channel usages.

The ***channel matrix*** is $P = (p_{ij})$, an $r \times s$ stochastic matrix.

Examples: The ***binary symmetric channel (BSC)*** with error probability $p \in [0, 1]$ has $\mathscr{A} = \mathscr{B} = \{0, 1\}$. The channel matrix is $\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$. A symbol is transmitted correctly with probability $1 - p$.

The ***binary erasure channel*** has input alphabet $\{0,1\}$, and output alphabet $\{0,1,*\}$, where we miss a bit is probability p, giving channel matrix $\begin{pmatrix} 1-p & 0 & p \\ 0 & 1-p & 0 \end{pmatrix}$. We can model n uses of a channel by the ***n^{th} extension*** with input alphabet $\mathscr{A}^n$, and output alphabet $\mathscr{B}^n$.

A ***code*** c of ***length*** n is a function $c : M \to \mathscr{A}^n$ where M is the set of all possible messages. Implicitly, we also have a decoding rule $\mathscr{B}^n \to M$.
The ***size*** of c is $m = |M|$.
The ***information rate*** is $\rho(c) = \frac{1}{n}\log_2(m)$.
The ***error rate*** is $\widehat{e}(c) = \max_{x \in M}\{\mathbb{P}(\text{error}|x \text{ sent})\}$.

A channel can transmit reliably at a rate R if there exists a sequence of codes $(c_n : n \geq 1)$ with c_n a code of length n, $\lim_{n\to\infty}(\rho(c_n)) = R, \lim_{n\to\infty}(\widehat{e}(c_n)) = 0$. The capacity of a channel is the supremum of all reliable transmission rates.

Theorem 0.1. *A BSC with error probability $p < \frac{1}{2}$ has a non-zero capacity (i.e. good codes exist).*

Proof. See **9.3** □

1 Noiseless Coding

1.1 Prefix-free Codes

For an alphabet $\mathscr{A}, |\mathscr{A}| < \infty$, let $\mathscr{A}^* = \bigcup_{n\geq 0}\mathscr{A}^n$, the set of all finite strings from $\mathscr{A}$. The ***concatenation*** of strings $x = x_1 \dots x_r$ and $y = y_1 \dots y_s$ is $xy = x_1 \dots x_r y_1 \dots y_s$.

Let $\mathscr{A}, \mathscr{B}$, be alphabets. A ***code*** is a function $c : \mathscr{A} \to \mathscr{B}^*$. The strings $c(a)$ for $a \in \mathscr{A}$ are called ***codewords*** (cws). If $x, y \in \mathscr{B}^*$ then x is a ***prefix*** of y if $y = xz$ for some $z \in \mathscr{B}^*$.

For example, we have the Greek fire code, found in the writings of Polybius around 280 BC. $\mathscr{A} = \{\alpha, \beta, \dots, \omega\}, \mathscr{B} = \{1,2,3,4,5\}$, with code $\alpha \mapsto 11, \beta \mapsto 12, \dots, \psi \mapsto 53, \omega \mapsto, 54$, where xy means "x torches held up, and another y torches nearby".

The English language is even a code: we can let $\mathscr{A}$ be words in a given dictionary, and $\mathscr{B} = \{a, b, \dots, z, \sqcup\}$, where the coding function is to spell the word and follow it with a space.

We send a message $x_1 \dots x_n \in \mathscr{A}^*$ as $c(x_1) \dots c(x_n) \in \mathscr{B}^*$. So c extends to a function $c^* : \mathscr{A}^* \to \mathscr{B}^*$.

c is ***decipherable/decidable*** if c^* is injective, so that each string in $\mathscr{B}^*$ could have come from at most one message. Note that it isn't sufficient to just have c injective, although clearly this is necessary:

$\mathscr{A} = \{1,2,3,4\}, \mathscr{B} = \{0,1\}, c : 1 \mapsto 0, 2 \mapsto 1, 3 \mapsto 00, 4 \mapsto 01$. Then $c^*(114) = 0001 = c^*(312)$.

If $|\mathscr{A}| = m, |\mathscr{B}| = a$, then we say c is an ***a-ary code of size m***. 2-ary = ***binary***, 3-ary=***ternary***.

We aim to construct decipherable codes with short word lengths. Assuming c is injective, the following are always decipherable:

- Block codes, where every codeword has the same length (e.g. Greek fire, ASCII)

- Comma codes, where we have an "end of word" character (e.g. English language)
- Prefix-free codes, where no codeword is a prefix of any other distinct words.

Note that both of the first two are special cases of prefix-free codes. Prefix-free codes are often called ***instantaneous*** or ***self-punctuating*** codes. Note that not all decipherable codes are prefix-free: $0 \mapsto 01, 1 \mapsto 011$ is decipherable but not prefix free.

Theorem 1.1 (Kraft's Inequality)**.** *Let $|\mathscr{A}| = m, |\mathscr{B}| = a$. A prefix-free code $c : \mathscr{A} \to \mathscr{B}^*$ with word lengths $\ell_1, \ldots, \ell_m$ exists if and only if:*

$$\sum_{i=1}^{m} a^{-\ell_i} \leq 1 \tag{$*$}$$

Proof. Rewrite $(*)$ as $\sum_{\ell=1}^{s} n_\ell a^{-\ell} \leq 1 (\star)$, where n_ℓ is the number of codewords of length ℓ and $s = \max_{1 \leq i \leq m} \ell_i$.

$\implies$ If $c : \mathscr{A} \to \mathscr{B}^*$ is prefix-free, then $n_1 a^{s-1} + n_2 a^{s-2} + \ldots n_s \leq a^s$, since the LHS is the number of strings of length s in $\mathscr{B}$ with some codeword of c as a prefix, and RHS is the number of strings of length s. Dividing by a^s gives $(\star)$.

$\impliedby$ Given $n_1, \ldots, n_s$ satisfying $(\star)$, we need to construct a prefix-free code c with n_ℓ codewords of length ℓ for all $\ell \leq s$. We use induction on s. The case $s = 1$ is clear: we have $(\star)$ gives $n_1 \leq a$, so we can choose a code.

By the induction hypothesis there is a prefix-free code $\widehat{c}$ with n_ℓ codewords of length ℓ for all $\ell \leq s - 1$. Then $(\star)$ gives:

$$n_1 a^{s-1} + n_2 a^{s-2} + \ldots + n_{s-1} a + n_s \leq a^s$$

where the first $s - 1$ terms on LHS sum to the number of strings of length s with some codeword of $\widehat{c}$ as a prefix, and the RHS is the number of strings of length s. Hence we can add at least n_s new codewords of length s to $\widehat{c}$ and maintain the prefix-free property, giving our code.

□

Theorem 1.2 (McMillan)**.** *Any decipherable code satisfies Kraft's inequality*

Proof (Karush). Let $c : \mathscr{A} \to \mathscr{B}^*$ be a decipherable code with codewords of lengths $\ell_1, \ldots, \ell_m$. Let $s = \max_{1 \leq i \leq m} \ell_i$. Then for $R \in \mathbb{N}$:

$$\left(\sum_{i=1}^{m} a^{-\ell_i} \right)^R = \sum_{l=1}^{Rs} b_\ell a^{-\ell}$$

where $b_\ell = |\{x \in \mathscr{A}^R : c^*(x) \text{ has length } \ell\}| \leq |\mathscr{B}^\ell| = a^\ell$, using the fact that c^* is injective. Then:

$$\left(\sum_{i=1}^{m} a^{-\ell_i} \right)^R \leq \sum_{l=1}^{R} a^\ell a^{-\ell} = Rs$$

$$\sum_{i=1}^{m} a^{-\ell_i} \leq (Rs)^{\frac{1}{R}} \to 1 \text{ as } R \to \infty$$

□

Corollary 1.3. *A decipherable code with prescribed word lengths exists iff a prefix-free code with same word lengths exists.*

Proof.

$\Longrightarrow$ Use **1.2** to generate a prefix-free code by **1.1**

$\Longleftarrow$ Prefix-free codes are decipherable.

□

2 Shannon's Noiseless Coding Theorem

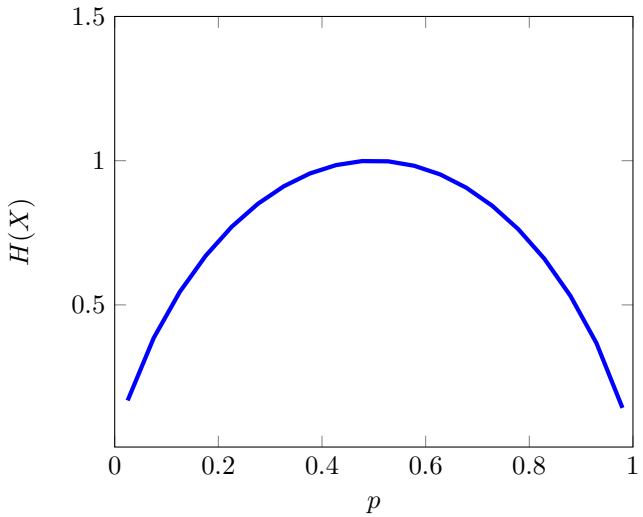
Entropy is a measure of 'randomness' or 'uncertainty'. Suppose we have a random variable X that takes values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. Then the ***entropy*** (roughly speaking) is the expected number of fair coin tosses needed to simulate X.

Examples:

- $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$. We can identify $\{x_1, x_2, x_3, x_4\}$ with $\{HH, HT, TH, TT\}$, and so the entropy of this random variable is 2.
- $(p_1, p_2, p_3, p_4) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$. Here, the entropy is $1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} = \frac{7}{4}$. We might say then, since the entropy is greater, that the first example is "more random" than the second.

More concretely, the ***Shannon (or information) entropy*** of X is $H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$. Note that $H(X) \geq 0$ with equality if and only if $\mathbb{P}(X = x_i) = 1$. This is measured in ***bits***. We take the convention that $0 \log 0 = 0$.

Example: Consider a biased coin, where $\mathbb{P}(X = H) = p, \mathbb{P}(X = T) = 1 - p$. Then $H(X) = -p \log p - (1-p) \log(1-p) = p \log(\frac{1-p}{p} - \log(1-p)$.



Proposition 2.1 (Gibb's Inequality)**.** *Let* $(p_1, \ldots, p_n)$ *and* $(q_1, \ldots, q_n)$ *be probability distributions. Then:*

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i$$

with equality if and only if $p_i = q_i$ *for all* i*.*

Proof. Since $\log x = \frac{\ln x}{\ln 2}$, we may replace log by ln in the proof. Put $I = \{1 \leq r \leq n : p_i \neq 0\}$. Now $\ln x \leq x - 1$ with equality if and only if $x = 1$. So we have $\ln \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1$, and hence:

$$\begin{aligned}
\sum_{i \in I} p_i \ln \frac{q_i}{p_i} &\leq \sum_{i \in I} q_i - \sum_{i \in I} p_i \\
&= \sum_{i \in I} q_i - 1 \leq 0 \\
\therefore -\sum_{i \in I} p_i \ln p_i &\leq -\sum_{i \in I} p_i \ln q_i \\
\therefore -\sum_{i=1}^{n} p_i \log p_i &\leq -\sum_{i=1}^{n} p_i \log q_i
\end{aligned}$$

If equality holds, then $\sum_{i \in I} p_i = 1$ and $\frac{p_i}{q_i} = 1$ for all $i \in I$, so $p_i = q_i$ □

Corollary 2.2. $H(p_1, \ldots, p_n) \leq \log n$ *with equality if and only if* $p_1 = \ldots = p_n = \frac{1}{n}$*.*

Proof. Take $q_1 = \ldots = q_n = \frac{1}{n}$ in **2.1**. □

Let $\mathscr{A} = \{\mu_1, \ldots, \mu_m\}$, and $|\mathscr{B}| = a$, where $m, a \geq 2$. The random variable X takes values $\mu_1, \ldots, \mu_m$ with probabilities $p_1, \ldots, p_m$. We say a code $c : \mathscr{A} \to \mathscr{B}^*$ is ***optimal*** if it is a decipherable code with smallest possible expected word length, $\mathbb{E}S = \sum_i p_i \ell_i$.

Theorem 2.3 (Shannon's Noiseless Coding Theorem)**.** *The expected word length* $\mathbb{E}S$ *of an optimal code satisfies:*

$$\frac{H(X)}{\log a} \leq \mathbb{E}S < \frac{H(X)}{\log a} + 1$$

Proof. For the lower bound, take $c : \mathscr{A} \to \mathscr{B}^*$ decipherable with word lengths $\ell_1, \ldots, \ell_m$. Then set $q_i = \frac{a^{-\ell_i}}{D}$ where $D = \sum_{i=1}^{m} a^{-\ell_i}$. Now we have that $\sum_{i=1}^{m} q_i = 1$. By Gibbs,

$$\begin{aligned}
H(X) &\leq -\sum_{i=1}^{m} p_i \log q_i \\
&= -\sum p_i \left(-\ell_i \log a - \log D\right) \\
&= \left(\sum_{i=1}^{m} p_i \ell_i\right) \log a + \log D
\end{aligned}$$

By McMillan, $D \leq 1$, so $\log D \leq 0$, and so $H(X) \leq (\sum_{i=1}^{m} p_i \ell_i) \log a = \mathbb{E}S \cdot \log a$, and we have equality if and only if $p_i = a^{-\ell_i}$ for some integers $\ell_1, \ldots \ell_m$.

For the upper bound, take $\ell_i = \lceil -\log_a p_i \rceil$. Then $-\log_a p_i \le \ell_i \implies p_i \ge a^{-\ell_i}$.

Now $\sum_{i=1}^m a^{-\ell_i} \le \sum_{i=1}^m p_i = 1$. By Kraft, there is some prefix-free code c with word lengths $\ell_1, \ldots, \ell_m$, and the expected word length of c is $\mathbb{E}S = \sum p_i \ell_i < \sum p_i(-\log_a p_i + 1) = \frac{H(X)}{\log a} + 1$. $\square$

Example: ***Shannon-Fano coding***
We mimic the above proof: given probabilities $p_1, \ldots, p_n$, set $\ell_i = \lceil -\log_a p_i \rceil$. Construct the prefix-free code with word lengths $\ell_1, \ldots, \ell_m$ by choosing in order of increasing length, ensuring that previous codewords are not prefixes. For example, if $a = 2, m = 5$ we have:

i	p_i	$\lceil -\log_2 p_i \rceil$	Codewords
1	0.4	2	00
2	0.2	3	010
3	0.2	3	011
4	0.1	4	1000
5	0.1	4	1001

$\mathbb{E}S = \sum p_i \ell_i = 2.8$, entropy $= 2.12$

3 Huffman Coding Algorithm

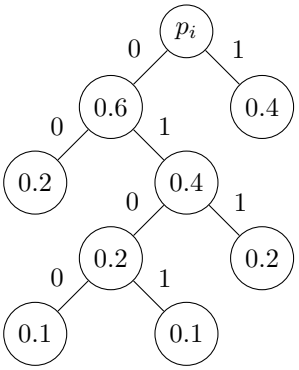
Huffman was a student of Fano, and was thinking about how to construct an optimal code. For simplicity, we will take $a = 2$. Suppose we get messages with orders $p_1 \ge p_2 \ge \ldots \ge p_m$. Huffman gave a recursive definition of codes that we can prove are optimal. If $m = 2$, then take codewords 0 and 1.

If $m > 2$, we first have a Huffman code for messages $\mu_1, \ldots, \mu_{m-2}, \nu$ with probabilities $p_1, \ldots, p_{m-2}, p_{m-1} + p_m$, then append 0 and 1 to give codewords for μ_{m-1} and μ_m.

Note:

- Huffman codes are prefix-free.
- We have some choices to make if some of the p_j are equal, so Huffman codes are not unique.

Example: Reconsider the previous example:



i	p_i	Codewords
1	0.4	1
2	0.2	00
3	0.2	011
4	0.1	0100
5	0.1	0101

This code has expected length 2.2, which is less than Shannon-Fano gave.

Theorem 3.1 (Huffman, 1952)**.** *Huffman codes are optimal.*

Proof. We show this by induction on m. The case of $m = 2$ is trivial. For $m > 2$, let c_m be a Huffman code for source X_m which takes values $\mu_1, \ldots, \mu_m$ with probabilities $p_1 \geq \ldots \geq p_m$. Then c_{m-1} is constructed from a Huffman code c_{m-1} for values $\mu_1, \ldots, \mu_{m-1}, \nu$ with probabilities $p_1, \ldots, p_{m-2}, p_{m-1} + p_m$.

Observe that $\mathbb{E}S_m = \mathbb{E}S_{m-1} + p_{m-1} + p_m$ by construction of c_m from c_{m-1}.

Now let c'_m be an optimal code for X_m. Without loss of generality, we may take c'_m to be prefix-free and the last two codewords of c'_m have maximal length and differ only in the last digit (see **3.2** below). Say $c'_m(\mu_{m-1}) = y0, c'_m(\mu_m) = y1$ for some $y \in \{0,1\}^*$.

Let c'_{m-1} be the prefix free code for X_{m-1} given by $c'_{m-1}(\mu_i) = c'_m(\mu_i), c'_{m-1}(\nu) = y$.

Then the expected word length is $\mathbb{E}S'_m = \mathbb{E}S'_{m-1} + p_{m-1} + p_m \geq \mathbb{E}S_{m-1} + p_{m-1} + p_m = \mathbb{E}S_m$ by the inductive hypothesis, and so c_m is optimal. □

Lemma 3.2. *Suppose messages $\mu_1, \ldots, \mu_m$ are sent with probabilities $p_1, \ldots, p_m$, with an optimal code c with word lengths $\ell_1, \ldots, \ell_m$. Then:*

1. *If $p_i > p_j$ then $\ell_i \leq \ell_j$.*
2. *Among all codewords of maximal length, there are two that differ only in the last digit.*

Proof. Otherwise, modify c by swapping the i^{th} and j^{th} codewords, or deleting the last letter of each codeword of maximal length. The modified code is still prefix-free but has shorter expected word length, contradicting optimality of c. □

4 Joint Entropy

If X, Y are random variables with value in $\mathscr{A}$ and $\mathscr{B}$. Then (X, Y) is also a random variable with entropy $H(X, Y)$, the ***joint entropy*** of X, Y.

$$H(X,Y) = -\sum_{x \in \mathscr{A}} \sum_{y \in \mathscr{B}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x, Y = y)$$

We can of course generalise this to any finite number of random variables. We will use Gibb's (**2.1**) to prove:

Lemma 4.1. *Let X, Y be random variables taking values in $\mathscr{A}, \mathscr{B}$. Then:*

$$H(X,Y) \leq H(X) + H(Y)$$

with equality if and only if X and Y are independent.

Proof. Let $\mathscr{A} = \{x_1, \ldots, x_m\}, \mathscr{B} = \{y_1, \ldots, y_n\}$. Set $p_{ij} = \mathbb{P}(X = x_i, Y = y_i), p_i = \mathbb{P}(X = x_i), q_i = \mathbb{P}(Y = y_i)$. Then Gibb's inequality with $\{p_{ij}\}$ and $\{p_i q_j\}$ gives:

$$\begin{aligned} -\sum_{i,j} p_{ij} \log p_{ij} \leq -\sum_{i,j} p_{ij} \log(p_i q_j) &= -\sum_i \left(\sum_j p_{ij} \right) \log p_i - \sum_j \left(\sum_i p_{ij} \right) \log q_j \\ &= -\sum_i p_i \log p_i - \sum_j q_j \log q_j \end{aligned}$$

i.e. $H(X,Y) \leq H(X) + H(Y)$, with equality if and only if $p_{ij} = p_i q_j$ for all i, j, i.e. when X, Y are independent. □

Example: Let X be a random variable that takes D values with probability $\frac{1}{D}$. Then $H(X) = \log_2(D)$. Suppose $X_1, \ldots, X_N$ are i.i.d. with the same distribution as X. Then $H(X_1, \ldots, X_N) = N \log_2 D$.

5 Error Correcting Codes

5.1 Noisy Channels and Hamming's Code

A ***binary [n,m]-code*** is a subset $C \subseteq \{0,1\}^n$ of ***size*** $m = |C|$, ***length*** n. The elements of C are called ***codewords***. We use an $[n, m]$-code to send one of m messages through a binary symmetric channel, making n uses of the channel. Clearly $1 \leq m \leq 2^n$, so $0 \leq \frac{1}{n} \log m \leq 1$. If $|C| = 1$ then $\rho(C) = 0$, and if $C = \{0,1\}^n$ then $\rho(C) = 1$.

For $x, y \in \{0,1\}^n$, the ***Hamming distance*** $d(x,y) = |\{i : 1 \leq i \leq n, x_i \neq y_i\}|$, i.e. the number of positions where x and y differ.

We have three possible decoding rules:

1. The ***ideal observer*** decoding rule decodes $x \in \{0,1\}^n$ as $c \in C$ maximising $\mathbb{P}(c \text{ sent} | x \text{ received})$.
2. Then ***maximum likelihood*** decoding rule decodes $x \in \{0,1\}^n$ as $c \in C$ maximising $\mathbb{P}(x \text{ received} | c \text{ sent})$.
3. The ***minimum distance*** decoding rule decodes $x \in \{0,1\}^n$ as $c \in C$ minimising $d(x,c)$.

Lemma 5.1.

1. *If all the messages are equally likely, then 1. and 2. agree.*
2. *If $p < \frac{1}{2}$, then 2. and 3. agree.*

Proof.

1. By Bayes' Rule:

$$\mathbb{P}(c \text{ sent} | x \text{ received}) = \frac{\mathbb{P}(c \text{ sent})}{\mathbb{P}(x \text{ received})} \mathbb{P}(x \text{ received} | c \text{ sent})$$

By Hypothesis, $\mathbb{P}(c \text{ sent})$ is independent of $c \in C$, and so for fixed x, maximising $\mathbb{P}(c \text{ sent} | x \text{ received})$ is the same as maximising $\mathbb{P}(x \text{ received} | c \text{ sent})$.

Let $r = d(x, c)$. Then $\mathbb{P}(x \text{ received}|c \text{ sent}) = p^r(1-p)^{n-r} = (1-p)^n \left(\frac{p}{1-p}\right)^r$. Since $p < \frac{1}{2}$, $\frac{p}{1-p} < 1$, and so maximising $\mathbb{P}(x \text{ received}|c \text{ sent})$ is the same as minimising $d(x, c)$. □

For instance, suppose 000 is sent with probability $\frac{9}{10}$, and 111 with probability $\frac{1}{10}$, through a binary symmetric channel with error probability $\frac{1}{4}$. If we receive 110, the ideal receiver computes $\mathbb{P}(000 \text{ sent}|110 \text{ received}) = \frac{3}{4}; \mathbb{P}(110 \text{ sent}|110 \text{ received}) = \frac{1}{4}$, and so decodes it as 000. But the minimum distance (and so maximal likelihood) code is 111. Henceforth, we will decide to use minimal distance decoding.

Note that minimal distance decoding can be expensive in terms of time and storage if $|C$ is large, and we also need to specify a convention in the case of a tie (e.g. make a random choice, request the message again).

A code is ***d-error detecting*** if changing up to d digits in each codeword can never produce another codeword. It is ***e-error correcting*** if, knowing that $x \in \{0,1\}^n$ differs from some codeword in at most e places, we can deduce uniquely what the codeword is.

Examples

1. A ***repetition code*** of length n has codewords $00\ldots0, 11\ldots1$. This is an $[n, 2]$-code. It is $(n-1)$ error detecting and $\lfloor \frac{n-1}{2} \rfloor$-error correcting. But the information rate is only $\frac{1}{n}$.
2. A ***simple parity check code*** or ***paper tape code***: identify $\{0,1\}$ with $\mathbb{F}_2$ (i.e. arithmetic modulo 2), and let $C = \{(x_1, \ldots, x_n) \in \{0,1\}^n : \sum x_i = 0\}$. This is an $[n, 2^{n-1}]$-code. It is 1-error detecting, but cannot correct errors. Its information rate is $\frac{n-1}{n}$.
3. ***Hamming's Original Code*** is a 2-error detecting and 1-error correcting binary [7,16]-code:

$$C = \left\{ c \in \mathbb{F}_2^7 : \begin{array}{r} c_1 + c_3 + c_5 + c_7 = 0 \\ c_2 + c_3 + c_6 + c_7 = 0 \\ c_4 + c_5 + c_6 + c_7 = 0 \end{array} \right\}$$

The bits c_3, c_5, c_6, c_7 are arbitrary and c_1, c_2, c_4 are forced. The information rate is $\frac{4}{7}$.

Given $x \in \mathbb{F}_2^7$, we form the ***syndrome*** $z = (z_1, z_2, z_4) \in \mathbb{F}_2^7$, where $z_1 = x_1 + x_3 + x_5 + x_7$, $z_2 = x_2 + x_3 + x_6 + x_7$, $z_4 = x_4 + x_5 + x_6 + x_7$. If $x \in C$ then $z = (0,0,0)$. If $d(x, c) = 1$ for some $c \in C$ then x_i and c_i differ for $i = z_1 + 2z_2 + 4z_4$. This can be checked easily for $c = 0$ with a case by case check of the seven binary sequences of six 0s and one 1, e.g. $x = 0010000$ gives a syndrome $z = (1,1,0), i = 1 + 2 + 0 = 3$.

Lemma 5.2. *d is a metric on $\mathbb{F}_2^n$.*

Proof. Immediately, $d(x, y) \geq 0$, with equality if and only if $x = y$, and $d(x, y) = d(y, x)$. For the triangle inequality, note that if x and z differ at position i then either x, y differ at i or y, z differ at i. So every difference appearing in $d(x, z)$ appears in $d(x, y) + d(y, z)$, so $d(x, z) \leq d(x, y) + d(y, z)$. □

Note that $d(x, y) = \sum_i d_1(x_i, y_i)$ where d_1 is the discrete metric on $\mathbb{F}_2$. We define the ***minimum distance*** of a code to be $\min_{c_1 \neq c_2} d(c_1, c_2)$.

Lemma 5.3. *Let C have minimal distance d. Then:*

1. *C is $(d-1)$-error detecting, but cannot detect all sets of d errors.*
2. *C is $\lfloor \frac{d-1}{2} \rfloor$-error correcting, but cannot correct all sets of $\lfloor \frac{d-1}{2} \rfloor + 1$ errors.*

Proof.

1. $d(c_1, c_2) \geq d$ for all distinct $c_1, c_2 \in C$. So C is $(d-1)$-error detecting. But $d(c_1, c_2) = d$ for some $c_1, c_2 \in C$. So C cannot detect all sets of errors.
2. Define the closed Hamming ball with center $x \in \mathbb{F}_2^n$, radius $r \geq 0$ as $B(x, r) = \{y \in \mathbb{F}_2^n : d(x, y) \leq r\}$. Now C is e-error correcting if and only if, for all $c_1 \neq c_2 \in C$, we have $B(c_1, e) \cap B(c_2, e) = \emptyset$.

   If $x \in B(c_1, e) \cap B(c_2, e)$, then $d(c_1, c_2) \leq d(c_1, x) + d(x, c_2) \leq 2e$. So if $d \geq 2e + 1$, then C is e-error correcting, with $e = \lfloor \frac{d-1}{2} \rfloor$. For the second part, take $c_1, c_2 \in C$ with $d(c_1, c_2) = d$. Then suppose $x \in \mathbb{F}_2^n$ differs from c_1 in e digits where c_1, c_2 differ too. Then $d(x, c_1) = e, d(x, c_2) = d - e$. If $d < 2e$ then $B(c_1, e) \cap B(c_2, d - e) \neq 0$, and so C cannot correct all sets of e-errors. Then take $e = \lceil \frac{d}{2} \rceil = \lfloor d - 1 \rfloor 2 + 1$.

□

As a point of a notation, an $[n, m]$-code with minimum distance d will be denoted as an $[n, m, d]$-code.

Examples:

1. Repetition of length n is an $[n, 2, n]$-code.
2. Simple parity check code of length n is an $[n, 2^{n-1}, 2]$-code.
3. Hamming's code is 1-error correcting, so $d \geq 3$. 0000000, 1110000 are both codewords, so it is a $[7, 16, 3]$-code, and hence 2-error correcting.

6 Covering Estimates

Denote $V(n, r) = |B(x, r)| = \sum_{i=0}^{r} \binom{n}{i}$, independent of $x \in \mathbb{F}_2^n$, as the ***volume*** of the ball (i.e. the number of points it contains).

Lemma 6.1 (Hamming's Bound)**.** *An e-error correcting code of length n has:*

$$|C| \leq \frac{2^n}{V(n, e)}$$

Proof. Suppose C is e-error correcting. Then $B(c_1, e) \cap B(c_2, e) = \emptyset$ for all $c_1 \neq c_2 \in C$. Then $\sum_{c \in C} |B(c, e)| \leq |\mathbb{F}_2^n| = 2^n$, i.e. $|C|V(n, e) \leq 2^n$. □

A code C of length n that can correct e errors is ***perfect*** if $|C| = \frac{2^n}{V(n,e)}$. Equivalently, a code is perfect if for all $x \in \mathbb{F}_2^n$ there is a unique $c \in C$ such that $d(x, c) \leq e$, or $\mathbb{F}_2^n = \bigcup_{c \in C} B(c, e)$, i.e. any $e + 1$ errors will make you decode incorrectly.

For example, Hamming's $[7, 16, 3]$-code is perfect, as $\frac{2^7}{7+1} = 2^4 = |C|$. Note that if $\frac{2^n}{V(n,e)} \notin \mathbb{Z}$ then there is no perfect e-error correcting code of length n, and even if $2^n / V(n, e)$ is an integer is may be the case that no perfect code exists.

Define $\boldsymbol{A(n,d)} = \max\{m : \exists\ [n, m, d]\text{-code}\}$. For instance, $A(n, 1) = 2^n, A(n, n) = 2, A(n, 2) = 2^{n-1}$.

Lemma 6.2. $A(n, d + 1) \leq A(n, d)$

Proof. Let $m = A(n, d + 1)$, and pick a code C with parameters $[n, m, d + 1]$. Let $c_1, c_2 \in C$ with $d(c_1, c_2) = d + 1$. Let c_1' differ from c_1 in exactly one of the places where c_1, c_2 differ. Then $d(c_1', c_2) = d$. If $c \in C \setminus \{c_1\}$, then $d(c, c_1) \leq d(c, c_1') + d(c_1', c_1) \implies d(c_1, c_1') \geq d$.

Replacing c_1 by c_1' gives an $[n, m, d]$-code i.e. $m \leq A(n, d)$. □

Corollary 6.3. *Equivalently,* $A(n, d) = \max\{m : \exists\ [n, m, d'] - \textit{code for some } d' \geq d\}$.

Theorem 6.4.

$$\frac{2^n}{V(n, d-1)} \leq A(n, d) \leq \frac{2^n}{V(n, \lfloor \frac{d-1}{2} \rfloor)}$$

The lower bound is called the Gilbert-Shannon-Varshanov (GSV) bound, whilst the upper bound follows from Hamming's bound.

Proof of GSV. Let $m = A(n, d)$, and let C be a $[n, m, d]$-code. Then there cannot exist $x \in \mathbb{F}_2^n$ with $d(x, c) \geq d$ for all $c \in C$, otherwise we could replace C by $C \cup \{x\}$, contradicting maximality of $d(x, c)$. Hence $\mathbb{F}_2^n = \bigcup_{c \in C} B(c, d - 1)$. Hence $2^n \leq mV(n, d - 1)$. □

For example, take $n = 10, d = 3$. Then $V(n, 2) = 56, V(n, 1) = 11$, and so these bounds give $\frac{2^10}{56} \leq A(10, 3) \leq \frac{2^10}{11}$, i.e. $19 \leq A(10, 3) \leq 93$, but in fact we know computationally that it is between 72 and 79.

6.1 Asymptotics

We study $\frac{\log A(n, \lfloor n\delta \rfloor)}{n}$ as $n \to \infty$ to see how large the information rate can be for a given error rate.

Proposition 6.5. *Let* $0 < \delta < \frac{1}{2}$. *Then:*

1. $\log V(n, \lfloor n\delta \rfloor) \leq nH(\delta)$.
2. $\frac{1}{n} \log A(n, \lfloor n\delta \rfloor) \geq 1 - H(\delta)$.

Proof. Assuming *1.* we see by the GSV bound, $A(n, \lfloor n\delta \rfloor) \geq \frac{2}{V(n, \lfloor n\delta \rfloor - 1} \geq \frac{2^n}{V(n, \lfloor n\delta \rfloor)}$, so $\frac{\log A(n, \lfloor n\delta \rfloor)}{n} \geq 1 - \frac{\log V(n, \lfloor n\delta \rfloor)}{n} \geq 1 - H(\delta)$, and so *1.* $\implies$ *2.*

For *1.* observe $H(\delta)$ is increasing for $\delta \leq \frac{1}{2}$, so WLOG we can assume $n\delta \in \mathbb{Z}$. Then:

$$\begin{aligned}
1 &= (\delta + (1-\delta))^n \\
&= \sum_{i=0}^{n} \binom{n}{i} \delta^i (1-\delta)^{n-i} \\
&\geq \sum_{i=0}^{n\delta} \binom{n}{i} \delta^i (1-\delta)^{n-i} \\
&= (1-\delta)^n \sum_{i=0}^{n\delta} \binom{n}{i} \left(\frac{\delta}{1-\delta}\right)^i \\
&\geq (1-\delta)^n \sum \binom{n}{i} \left(\frac{\delta}{1-\delta}\right)^{n\delta} \\
&= \delta^{n\delta}(1-\delta)^{n(1-\delta)} V(n, n\delta) \\
0 &\geq n\delta \log \delta + n(1-\delta)\log(1-\delta) + \log V(n, n\delta) \\
0 &\geq -nH(\delta) + \log V(n, n\delta)
\end{aligned}$$

□

This constant $H(\delta)$ is the best possible, in the sense that:

Proposition 6.6.

$$\lim_{n\to\infty} \frac{\log V(n, \lfloor n\delta \rfloor)}{n} = H(\delta)$$

Proof. WLOG we may assume that $0 < \delta < \frac{1}{2}$. Let $0 \leq r \leq \frac{n}{2}$. Recall that $V(n,r) = \sum_{i=0}^{r} \binom{n}{i}$. Then:

$$\binom{n}{r} \leq V(n,r) \leq (r+1)\binom{n}{r} \qquad ((*))$$

From Stirling's approximation, $\log \binom{n}{r} = -r \log \frac{r}{n} - (n-r) \log \frac{n-r}{n} + \mathcal{O}(\log n) = nH(\frac{r}{n}) + \mathcal{O}(\log n)$.

Then from $(*)$, we have:

$$H(\frac{r}{n}) + \mathcal{O}\left(\frac{\log n}{r}\right) \leq \frac{\log V(n,r)}{n} \leq H\left(\frac{r}{n}\right) + \mathcal{O}\left(\frac{\log n}{n}\right)$$

$$\therefore \lim_{n\to\infty} \frac{\log V(n, \lfloor n\delta \rfloor)}{n} = H(\delta)$$

□

7 New Codes from Old

Suppose C is an $[n, m, d]$-code. Then the ***parity check digit extension*** C^+ of C is the code of length $n+1$ given by:

$$C^+ = \{(c_1, \ldots, c_n, \sum_{i=1}^{n} c_i) : (c_1, \ldots, c_n) \in C\}$$

where the summation is done modulo 2. It is an $[n+1, m, d']$-code where $d' = d$ or $d+1$.

We can also delete the i^{th} digit from each codeword for $1 \leq i \leq n$, giving a ***truncated*** or ***punctured*** codeword (depending on if $i = n$ or $i < n$ respectively), called C^-, with parameters $[n-1, m, d']$ where $d - 1 \leq d' \leq d$.

Finally, given some $1 \leq i \leq n, \alpha \in \mathbb{F}_2$, we can create the ***shortened*** or ***punctured*** code C' of C is $\{(c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n) : (c_1, \ldots, c_{i-1}, \alpha, c_{i+1} \ldots, c_n) \in C\}$. This has parameters $[n-1, m', d']$ with $d' \geq d$ and $m' \geq \frac{m}{2}$ for c a suitable choice of α.

8 AEP and Shannon's First Coding Theorem

A ***source*** is a sequence of random variables $X_1, X_2, \ldots$ taking values in some alphabet $\mathscr{A}$. A source is ***Bernoulli*** or ***memoryless*** if $X_1, \ldots$ are independently identically distributed (IID). A source $X_1, \ldots$ is ***reliably encodable at rate r*** if there are subsets $A_n \subseteq \mathscr{A}^n$ such that:

1. $\lim_{n\to\infty} \frac{\log |A_n|}{n} = r$
2. $\lim_{n\to\infty} \mathbb{P}[(X_1, \ldots, X_n) \in A_n] = 1$

The ***information rate*** H of a source is the infimum of all reliable encoding ratesso that $0 \leq H \leq \log |\mathscr{A}|$. Shannon's first coding theorem computes the information rate of certain sources, including Bernoulli sources.

8.1 Reminder from 1A Probability

A ***probability space*** is given by a triple $(\Omega, \mathscr{F}, \mathbb{P})$ where $\mathscr{F} \subset \mathcal{P}(\Omega)$ is a set of ***events*** and $\mathbb{P}$ is a ***probability measure***, and a ***random variable*** X is a function defined on Ω with some range. It has a ***probability mass function*** $p : x \mapsto \mathbb{P}(X = x)$

We say that a sequence of random variables $X_1, X_2, \ldots$ ***converges in probability*** to $\lambda \in \mathbb{R}$ means that

$$\forall\, \epsilon > 0, \lim_{n\to\infty} \mathbb{P}(|X_n - \lambda| \geq \epsilon) = 0$$

We write $X_n \xrightarrow{\mathbb{P}} \lambda$ as $n \to \infty$.

Theorem 8.1 (Weak Law Of Large Numbers, WLLN)**.** *Let $X_1, X_2, \ldots$ be* IID *discrete real-valued random variables with finite expected value μ. Then:*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu \text{ as } n \to \infty$$

Proof. See Carne, theorem 10.3. □

Lemma 8.2. *The information rate of a Bernoulli source $X_1, X_2, \ldots$ is at most the expected word length of an optimal code $c : \mathscr{A} \to \{0,1\}^*$ for X_i.*

Proof. Let $\ell_1, \ell_2, \ldots$ be the lengths of codewords when we encode $X_1, X_2, \ldots$ using c. Then given $\epsilon > 0$, let $A_n = \{x \in \mathscr{A}^n : c^*(x) \text{ has length } < n(\mathbb{E}[\ell_i] + \epsilon)\}$. Then:

$$\begin{aligned}\mathbb{P}\left[(X_1, \ldots, X_n) \in A_n\right] &= \mathbb{P}\left[\sum \ell_i < n(\mathbb{E}[\ell_i] + \epsilon)\right] \\ &\geq \mathbb{P}\left(\left|\frac{1}{n}\sum \ell_i - \mathbb{E}[\ell_i]\right| < \epsilon\right) \\ &\to 1 \text{ as } n \to \infty\end{aligned}$$

Now c is decipherable so c^* is injective, and hence $|A_n| \leq 2^{n(\mathbb{E}[\ell_i]+\epsilon)}$. Making A_n larger if required, we may take $|A_n| = \lfloor 2^{n(\mathbb{E}[\ell_i]+\epsilon)} \rfloor$. Hence $\frac{\log |A_n|}{n} \to \mathbb{E}[\ell_i] + \epsilon$. So $X_1, X_2, \ldots$ is reliably encodable at a rate $r = \mathbb{E}[\ell_i] + \epsilon$ for any $\epsilon > 0$, and hence the information rate is at most $\mathbb{E}[\ell_i]$. □

Corollary 8.3. *A Bernoulli source has information rate less than $H(X_1) + 1$.*

Proof. Use **8.2** and the Noiseless Coding theorem **2.3**. □

Now suppose we encode $X_1, X_2, \ldots$ in blocks:

$$\underbrace{X_1, \ldots, X_N}_{Y_1}, \underbrace{X_{N+1}, \ldots, X_{2N}}_{Y_2}, \ldots$$

such that $Y_1, Y_2, \ldots$ take values in $\mathscr{A}^N$. We can check that if $X_1, X_2, \ldots$ has information rate H, then $Y_1, Y_2, \ldots$ has information rate NH.

Proposition 8.4. *The information rateH of a Bernoulli source $X_1, X_2, \ldots$ is at most $H(X_1)$.*

Proof. Apply **8.3** to $Y_1, Y_2, \ldots$ to get:

$$NH < H(Y_1) + 1 = H(X_1, \ldots, X_N) + 1 = \sum_{i=1}^{N} H(X_i) + 1 = NH(X_i) + 1$$

i.e. $H < H(X_1) + \frac{1}{N}$ for all $N \geq 1$, and so $H \leq H(X_1)$. □

8.2 Typical Sequences

This content is nonexaminable, but is required to prove the examinable result that $H = H(X_1)$.

As a motivational example, toss a biased coin with head probability p, and let X_i be the outcome of the i^{th} flip. If we toss a large number, say N, times, we expect that we will get about pN heads and $(1-p)N$ tails. The probability of any particular sequence of pN heads and $(1-p)N$ tails is $p^{pN}(1-p)^{(1-p)N} = 2^{-NH(X)}$.

We say that a source $X_1, X_2, \ldots$ satisfies the ***Asymptotic Equipartition Property (AEP)*** for some constant $H \geq 0$ if:

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) \to H \text{ as } n \to \infty$$

Lemma 8.5. *The AEP for a source $X_1, x_2, \ldots$ is equivalent to the following:*
$\forall\, \epsilon > 0 \exists\, n_0(\epsilon)$ s.t. $\forall\, n \geq n_0(\epsilon) \exists\, T_n \subseteq \mathscr{A}^n$ s.t.

- $P[(X_1, \ldots, X_n) \in T_n] > 1 - \epsilon$
- $\forall\ (x_1, \ldots, x_n) \in T_n, 2^{-n(H+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H-\epsilon)}$

The T_n are called **typical sets***, and the $(x_1, \ldots, x_n) \in T_n$ are* **typical sequences**.

Proof. If $(x_1, \ldots, x_n) \in \mathscr{A}^n$ then we have the following equivalence:

$$2^{-n(H+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H-\epsilon)} \iff |-\frac{1}{n}\log p(x_1, \ldots, x_n) - H| \leq \epsilon \qquad (\dagger)$$

Both AEP and the claimed equivalent results say that $P((X_1, \ldots, X_N)$ satisfies $\dagger) \to 1$ as $n \to \infty$. □

Theorem 8.6 (Shannon's First Coding Theorem)**.** *If a source $X_1, X_2, \ldots$ satisfies the AEP with constant H then it has information rate H.*

Proof. Let $\epsilon > 0$ and let $T_n \subseteq \mathscr{A}^n$ be typical sets. Then for all $(x_1, \ldots, x_n) \in T_n$:

$$p(x_1, \ldots, x_n) \geq 2^{-n(H+\epsilon)} \implies 1 \geq |T_n| 2^{-n(H+\epsilon)} \implies \frac{\log |T_n|}{n} \leq (H + \epsilon)$$

Taking $A_n = T_n$ in the definition of reliable encoding, we see that the source is reliably encodeable at rate $H + \epsilon$. As $\epsilon > 0$, the information rate is $\leq H$.

Conversely, if $H = 0$ we're done, otherwise pick $0 < \epsilon < \frac{H}{2}$. Suppose for a contradiction that the source is reliably encodeable at rate $H - 2\epsilon$, say, with sets $A_n \subseteq \mathscr{A}^n$. Let $T_n \subseteq \mathscr{A}^n$ be typical sets. Then for all $(x_1, \ldots, x_n) \in T_n, p(x_1, \ldots, x_n) \leq 2^{-n(H-\epsilon)}$

Hence $\mathbb{P}(A_n \cap T_n) \leq 2^{-n(H-\epsilon)}$, and so $\frac{\log \mathbb{P}(A_n \cap T_n)}{n} \leq (H - \epsilon) + \frac{\log |A_n|}{n} \xrightarrow{n\to\infty} -\epsilon$. So $\mathbb{P}(A_n \cap T_n) \to 0$ as $n \to \infty$. But $\mathbb{P}(T_n) \leq \mathbb{P}(T_n \cap A_n) + \mathbb{P}(\mathscr{A}^n \setminus A_n) \to 0 + 0 = 0$ as $n \to \infty$, contradicting typicality of T_n. Hence the source cannot be reliably encoded at rate $H - 2\epsilon$, and so the information rate must be $\geq H$, and hence $= H$. □

9 Capacity and Shannon's Second Coding Theorem

Given a random variable X with mass function p_X, we can construct a new random variable $p(X) = p_X \circ X$, taking values in $[0, 1]$. Then $H(X) = \mathbb{E}(-\log p(X))$. For example, if X, Y are independent, then $p(X, Y) = p(X)p(Y)$, and so $-\log p(X, Y) = -\log p(X) - \log p(Y) \implies H(X, Y) = H(X) + H(Y)$.

Corollary 9.1. *A Bernoulli source $X_1, X_2, \ldots$ has information rate $H(X_1) = H$.*

Proof.* $p(X_1, \ldots, X_n) = p(X_1) \ldots p(X_n)$, and hence we have:

$$-\frac{\log p(X_1, \ldots X_n)}{n} = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) \xrightarrow{\mathbb{P}} H(X_i)$$

by the weak law of large numbers, and using the fact that the X_i are i.i.d. Check as an exercise that the AE holds with constant $H(X_1)$ using the definition of convergence in probability. Hence by **8.6** we are done. □

Note that **8.4** gave us an information rate $\leq H(X_1)$, without the use of the AEP. The AEP can also be used for noiseless coding - we encode the typical sequences with a block code and the atypical sequences arbitrarily, since they rarely occur. Many sources of interest, not just Bernoulli sources, satisfy the AEP. Under suitable conditions, the sequence $\frac{1}{n}H(X_1,\ldots,X_n)$ is decreasing and the AEP is satisfied with constant $H\lim_{n\to\infty}\frac{H(X_1,\ldots,X_n}{n}$.

Consider a communication channel with input of alphabet $\mathscr{A}$, and output B. A code of length n is a subset $C\subseteq\mathscr{A}^n$. The ***error rate***

$$\widehat{e}(C)=\max\mathbb{P}(\text{error }|c\text{ sent})$$

The ***information rate*** is $\rho(C)=\frac{\log|C|}{n}$.

The channel can ***transmit reliably*** at rate R if there are codes $C_1,C_2,\ldots$ with C_n of length n, and:

- $\lim_{n\to\infty}\rho(C_n)=R$.
- $\lim_{n\to\infty}\widehat{e}(C_n)=0$.

The ***operational capacity*** is the supremum of all reliable transmission rates.

Assume a source has information rate r bits per symbol, and emits symbols at s symbols per second, whilst the channel has capacity R bits per transmission and can can transmit symbols at S transmissions per second. Usually $S=s=1$. If $rs\leq RS$ then we can encode and transmit reliably, and if $rs>RS$ we cannot.

Proposition 9.2. *A binary symmetric channel with error probability $p<\frac{1}{4}$ has non-zero capacity.*

Proof. We use the *GSV* bound. Pick δ with $2p<\delta<\frac{1}{2}$. We claim reliable transmission rate of $R=1-H(\delta)>0$.

Let C_n be a code of length n with minimum distance $\lfloor n\delta\rfloor$ of maximal size. Them $|C_n|=A(n,\lfloor n\delta\rfloor)\geq 2^{n(1-H(\delta)}=2^{nR}$.

Replacing C_n by a subcode we can assume $|C_n|=\lfloor 2^{nR}\rfloor$ with minimum distance still $\geq\lfloor n\delta\rfloor$

Now, with minimum distance decoding, $\widehat{e}(C_n)\leq\mathbb{P}$(in n uses the BSC makes more than $\frac{n\delta-1}{2}$ errors).

Pick $\epsilon>0$ with $p+\epsilon<\frac{\delta}{2}$. For n sufficiently large we have that $\frac{n\delta-1}{2}=n(\frac{\delta}{2}-\frac{1}{2n})>n(p+\epsilon)$.

Hence $\widehat{e}(C)\leq\mathbb{P}$(BSC makes more than $n(p+\epsilon)$ errors) $\to 0$, as we will see in the next lemma. □

Lemma 9.3. *Let $\epsilon>0$. A BSC with error probability p is used to transmit n digits. Then:*

$$\lim_{n\to\infty}\mathbb{P}(\textit{BSC makes at least } n(p+\epsilon)\textit{ errors})=0$$

Proof. If U_i is the Bernoulli random variable that digit i is mistransmitted. Then U_i are i.i.d. with probability p. So $\mathbb{E}[U_i]=p$. Then the probability we are interested in is $\mathbb{P}(\sum U_i\geq n(p+\epsilon))\leq\mathbb{P}(|\frac{1}{n}\sum U_i-p|\geq\epsilon)\to 0$ by the WLLN. □

9.1 Fano's Inequality

Let X, Y be random variables taking values in the alphabets $\mathscr{A}, \mathscr{B}$. Then:

- $H(X|Y=y) = -\sum_{x\in\mathscr{A}} \mathbb{P}(X=x|Y=y)\log\mathbb{P}(X=x|Y=y)$
- $H(X|Y) = \sum_{y\in\mathscr{B}} \mathbb{P}(Y=y)H(X|Y=y)$

Clearly $H(X|Y) \geq 0$.

Lemma 9.4.

$$H(X,Y) = H(X|Y) + H(Y)$$

Proof.

$$\begin{aligned}
H(X|Y) &= -\sum_{x\in\mathscr{A}}\sum_{y\in\mathscr{B}} \mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y)\log\mathbb{P}(X-x|Y=y) \\
&= -\sum_{x\in\mathscr{A}} \mathbb{P}(X=x,Y=y)\log\frac{\mathbb{P}(X=x,Y=y}{\mathbb{P}(Y=y)} \\
&= -\sum_{(x,y)\in\mathscr{A}\times\mathscr{B}} \underbrace{\mathbb{P}(X=x,Y=y)\log\mathbb{P}(X=x,Y=y)}_{H(X,Y)} + \underbrace{\mathbb{P}(X=x,Y=y)\log\mathbb{P}(Y=y)}_{-H(Y)} \\
&= H(X,Y) - H(Y)
\end{aligned}$$

□

Corollary 9.5. *$H(X|Y) \leq H(X)$ with equality if and only if X, Y are independent.*

Proof. Combine **4.1** with **9.4**. □

We can replace X, Y by random variables $X_1, X_2, \ldots, X_r$ and $Y_1, Y_2, \ldots, Y_s$, and similarly define $H(X_1, \ldots, X_r|Y_1, \ldots, Y_s)$[1].

Lemma 9.6.

$$H(X|Y) \leq H(X|Y,Z) + H(Z)$$

Proof.

$$\begin{aligned}
H(X,Y,Z) &= H(Z|X,Y) + H(X,Y) = H(Z|X,Y) + H(X|Y) + H(Y) \\
H(X,Y,Z) &= H(X|Y,Z) + H(Y,Z) = H(X|Y,Z) + H(Z|Y) + H(Y) \\
\therefore H(X|Y) &= -H(Z|X,Y) + H(X|Y,Z) + H(Z|Y) \\
&\leq H(X|Y,Z) + H(Z)
\end{aligned}$$

□

Proposition 9.7 (Fano's Inequality)**.** *Let X, Y be random variables taking values in $\mathscr{A}$, where $|\mathscr{A}| = m$. Let $p = \mathbb{P}(X \neq Y)$. Then:*

$$H(X|Y) \leq H(p) + p\log(m-1)$$

[1] $H(X,Y|Z)$ means "the entropy of (X and Y) given Z", NOT "the entropy of X and (Y given Z)"

Proof. Let $Z = \begin{cases} 0 & X = Y \\ 1 & X \neq Y \end{cases}$. Then $\mathbb{P}(Z=0) = 1-p, \mathbb{P}(Z=1) = p$. So $H(Z) = H(p)$. By **9.6**,

$$H(X|Y) \leq H(p) + H(X|Y,Z) \tag{$*$}$$

Then we have two cases:

$Z = 0$: We must have $X = y$, so $H(X|Y=y, Z=0) = 0$.

$Z = 1$: Just $m-1$ remaining possibilities for X, so $H(X|Y=y, Z=1) \leq \log(m-1)$.

Hence we have:

$$\begin{aligned} H(X|Y,Z) &= \sum_{y,z} \mathbb{P}(Y=y, Z=z) H(X|Y=y, Z=z) \\ &\leq \sum_{y} \mathbb{P}(Y=y, Z=1) \log(m-1) \\ &= \mathbb{P}(Z=1)\log(m-1) = p\log(m-1) \end{aligned}$$

Then by $(*)$, $H(X|Y) \leq H(p) + p\log(m-1)$. □

We will apply this result later when X takes values in $\mathscr{A}$ and Y is the result of passing codewords through a channel and then decoding, where p will be the probability of error.

If X, Y are random variables, then we define the ***mutual information*** of X and Y to be:

$$I(X;Y) \coloneqq H(X) - H(X|Y)$$

By **4.1** and **9.4**, $I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$, with equality if and only if X and Y are independent. We also see from this form that $I(X;Y) = I(Y;X)$.

Given a DMC with input alphabet $\mathscr{A}$ of size m, and output alphabet $\mathscr{B}$. Let X be a random variable taking values in $\mathscr{A}$ used as an input to the channel. Let Y be the random variable output, depending on both X and the channel matrix.

We define the ***information capacity*** to be the $\max_X I(X;Y)$, where the maximum is taken over all probability distributions $(p_1, \ldots, p_m)$ for X. Since the space of random variables for X is a closed and bounded subset of $\mathbb{R}^m$, by Heine-Borel it is compact and, since I is continuous, the maximum is attained. Note that the information capacity depends only on the channel matrix.

Theorem 9.8 (Shannon's Second Coding Theorem)**.** *For a DMC, the operational capacity is equal to the information capacity.*

We will prove $\leq$ in general, and $\geq$ for a binary symmetric channel only.

For example, with a BSC with error probability p, input X, output Y, then $\mathbb{P}(X=0) = \alpha, \mathbb{P}(X=1) = 1-\alpha$, and so $\mathbb{P}(Y=0) = \alpha(1-p) + (1-\alpha)p; \mathbb{P}(Y=1) = (1-\alpha)(1-p) + \alpha p$.

Then $C = \max_\alpha I(X;Y) = \max_\alpha [H(\alpha(1-p) + (1-\alpha)p) - H(p)] = 1 - H(p$, where the max attained at $\alpha = \frac{1}{2}$. So the information capacity $C = 1 + p\log p + (1-p)\log(1-p)$. We can plot this on a graph:

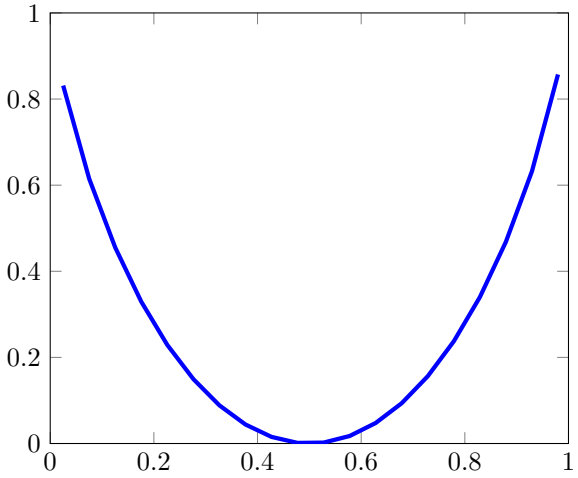


Figure 2: Capacity as a function of p for a binary symmetric channel

At $p = 0, 1$ the channel transmits perfectly, whilst at $p = \frac{1}{2}$ no information can be transmitted. We can choose to calculate $H(Y) - H(Y|X)$ or $H(X) - H(X|Y)$ to find the information - often one is easier than the other, for example with the binary erasure channel with erasure probability p.

$\mathbb{P}(X = 0) = \alpha, \mathbb{P}(X = 1) = 1 - \alpha, \mathbb{P}(Y = 0) = \alpha(1 - p), \mathbb{P}(Y = *) = p, \mathbb{P}(Y = 1) = (1 - \alpha)(1 - p)$

Then $H(X|Y = 0) = 0; H(X|Y = *) = H(\alpha); H(X|Y = 1) = 0$, and so $H(X|Y) = pH(\alpha)$. So $C = \max_\alpha(H(\alpha) - pH(\alpha)) = 1 - p$, where the maximum value is attained for $\alpha = \frac{1}{2}$.

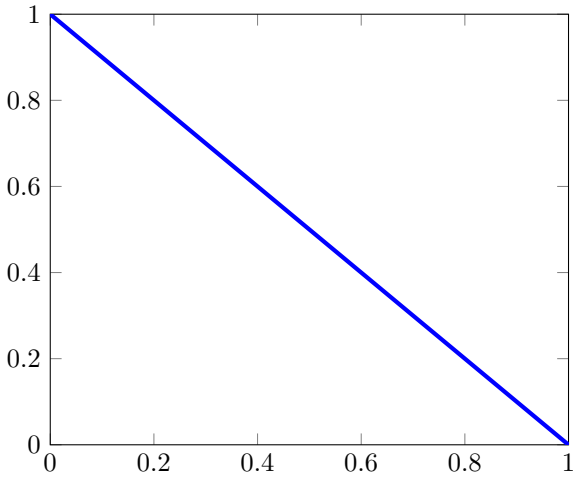


Figure 3: Capacity as a function of p for a binary erasure channel

We can now model using a channel n times, called the ***n^{th} extension***, i.e. replace input and output alphabets $\mathscr{A}, \mathscr{B}$ by $\mathscr{A}^n, \mathscr{B}^n$.

Lemma 9.9. *The n^{th} extension of a discrete memoryless channel with information capacity C has information capacity nC.*

Proof. The input $X_1, \ldots, X_n$ determines the output $Y_1, \ldots, Y_n$, and since the channel is memoryless, $H(Y_1, \ldots, Y_n | X_1, \ldots, X_n) = \sum_i H(Y_i | X_1, \ldots, X_n) = \sum_i H(Y_i | X_i)$.

$$\begin{aligned} I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) &= H(Y_1, \ldots, Y_n) - H(Y_1, \ldots, Y_n | X_1, \ldots, X_n) \\ &= H(Y_1, \ldots, Y_n) - \sum_{i=1}^{n} H(Y_i | X_i) \\ &\leq \sum_{i=1}^{n} [H(Y_i) - H(Y_i | X_i)] \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \leq nC \end{aligned}$$

To finish, we must find $X_1, \ldots X_n$ giving equality - take X_i to be i.i.d. with $I(X_i, Y_i) = C$, giving equality at the second $\leq$. Then the Y_i are i.i.d. so we have equality at the first $\leq$. So the maximum possible value is nC. □

Proposition 9.10. *For a DMC the operational capacity is at most the information capacity.*

Proof. Let C be the information capacity. Suppose that reliable transmission is possible at some $R > C$, i.e. there is a sequence $C_1, C_2, \ldots$, with C_n of length n such that $\lim_{n\to\infty} \rho(C_n) = R$, and $\lim_{n\to\infty} \widehat{e}(C_n) = 0$.

We define the ***average error rate*** $e(C_n) = \frac{1}{|C_n|} \sum_{c \in C_n} \mathbb{P}(\text{error} | c \text{ sent})$, so that $e(C_n) \leq \widehat{e}(C_n)$. Hence $e(C_n) \to 0$ as $n \to \infty$. Then let X be a random variable equidistributed in C_n. Transmit X and decode to obtain Y. So $e(C_n) = \mathbb{P}(X \neq Y)$. Then $H(X) = \log |C_n| = \log \lfloor 2^{nR} \rfloor \geq nR - 1$ for sufficiently large n. By Fano's inequality,

$$H(X|Y) \leq 1 + e(C_n) \log(|C_n| - 1) \leq 1 + e(C_n) n \rho(C_n)$$

By the previous proposition,

$$\begin{aligned} nC \geq I(X, Y) &= H(X) - H(X|Y) \\ &\geq \log |C_n| - (1 + e(C_n) n \rho(C_n)) \\ &= n\rho(C_n) - e(C_n) n \rho(C_n) - 1 \\ \therefore e(C_n) n \rho(C_n) &\geq n(\rho(C_n) - C) - 1 \\ e(C_n) &\geq \frac{\rho(C_n) - C}{\rho(C_n)} - \frac{1}{n\rho(C_n)} \to \frac{R - C}{R} \end{aligned}$$

But $R > C$, so $e(C_n) \not\to 0$ as $n \to \infty$. ↯ □

Proposition 9.11. *Consider the Binary Symmetric Channel with error probability p. Let $R < 1 - H(p)$. Then there exists a sequence of codes $C_1, C_2, \ldots$, with C_n of length n and size $\lfloor 2^{nR} \rfloor$, such that:*

$$\lim_{n\to\infty} \rho(C_n) = R$$
$$\lim_{n\to\infty} e(C_n) = 0$$

Note this is $e(C_n)$, not $\widehat{e}(C_n)$.

Proof. We use the method of random codes.

Without loss of generality, assume $p < \frac{1}{2}$. Let $\epsilon > 0$ be such that $p+\epsilon < \frac{1}{2}$ and $R < 1-H(p+\epsilon)$. This is always possible since H is continuous. Let $m = \lfloor 2^{nR} \rfloor$ and pick $C \in \mathscr{C} = \{[n,m]-\text{binary codes}\}$ and random. Note $|\mathscr{C}| = \binom{2^n}{m}$. Let $\mathscr{X}$ be a random variable equidistributed throughout $\mathscr{C}$, say $\mathscr{X} = \{X_1, \ldots, X_m\}$ where the X_i are random variables taking values in $\mathbb{F}_2^n$, such that:

$$\mathbb{P}(X_i = x_i | \mathscr{X} = C) = \begin{cases} \frac{1}{m} & x_i \in C \\ 0 & \text{otherwise} \end{cases}$$

Notice that $\mathbb{P}(X_2 = x_2 | X_1 = x_1) = \begin{cases} \frac{1}{2^n - 1} & x_1 \neq x_2 \\ 0 & \text{otherwise} \end{cases}$

Send $X = X_1$ through the BSC, receive Y, and decode to obtain Z. Under the minimum distance decoding, $\mathbb{P}(X \neq Z) = \frac{1}{|\mathscr{C}|} \sum_{C \in \mathscr{C}} e(C)$.

It then suffices to show that $\mathbb{P}(X \neq Z) \to 0$ as $n \to \infty$. Let $r = \lfloor n(p+2) \rfloor$.

$$\begin{aligned} \mathbb{P}(X \neq Z) &\leq \mathbb{P}(B(Y,r) \cap \mathscr{X} \neq \{X\}) \\ &= \underbrace{\mathbb{P}(X \notin B(Y,r))}_{(i)} + \underbrace{\mathbb{P}(B(Y,r) \cap \mathscr{X} \supsetneq \{X\})}_{(ii)} \end{aligned}$$

(i): $\mathbb{P}(X \notin B(Y,r)) = \mathbb{P}(\text{BSC makes more than } r \text{ errors}) \to 0$ as $n \to \infty$ by **9.3**.

(ii):

$$\begin{aligned} \mathbb{P}(B(Y,r) \cap \mathscr{X} \supsetneq \{X\}) &\leq \sum_{i=2}^{m} \mathbb{P}(X_i \in B(Y,r) \text{ and } X_1 \in B(Y,r)) \\ &\leq \sum_{i=2}^{m} \mathbb{P}(X_i \in B(Y,r) | X_1 \in B(Y,r)) \\ &= (m-1) \frac{V(n,r) - 1}{2^n - 1} \\ &\leq m \frac{V(n,r)}{2^n} \\ &\leq 2^{nR} 2^{nH(p+\epsilon)} 2^{-n} = 2^{n(R-(1-H(p+\epsilon)))} \to 0 \end{aligned}$$

since $R < 1 - H(p+\epsilon)$

□

Proposition 9.12. *We can replace e by $\widehat{e}$ in the previous proposition.*

Proposition 9.13. *Pick R' such that $R < R' < 1 - H(p)$. Then the previous proposition constructs $C_1', C_2', \ldots$, with C_n' of length n, size $\lfloor 2^{nR'} \rfloor$, and $e(C_n') \to 0$ as $n \to \infty$.*

Order then codewords of C_n' by $\mathbb{P}(\text{error}|\text{c sent})$, and delete the worse half them to give C_n. Then we have $|C_n| = \lfloor \frac{|C_n'|-1}{2} \rfloor$, $\widehat{e}(C_n) \leq 2e(C_n')$. Then $\rho(C_n) \to R$ and $\widehat{e}(C_n) \to 0$ as $n \to \infty$.

Note that:

1. **9.12** says we can transmit reliably at any rate $R < 1 - H(p)$ so the capacity is at least $1 - H(p)$. But, by **9.10**, the capacity at most $1 - H(p)$, and hence the BSC with error probability has capacity $1 - H(p)$.
2. The proof shows that good codes exist, but it does not tell us how to find them.

10 Interlude: An Application to Gambling

Let $0 < p < 1$, $n > 0$, and $0 \leq w < 1$. A coin is tossed n times in succession. $P(H) = p$ and, if I pay k ahead of a particular throw, then I get back kn if the throw is head, but nothing if the throw is a tail.

What is my strategy?

- If $pn < 1$, then don't bet - your expected winnings is < 0.
- If $pn > 1$, then we want to bet, but how much? A larger bet means more winnings, but also more risk. What proportion w of our total wealth should we bet?

Note that w is always the same, only the size of the fortune changes. Let the fortune by X_j after the j^{th} throw. I bet wX_j, retaining $(1-w)X_j$. My fortune X_{j+1} after the $(j+1)^{\text{th}}$ throw is:

$$X_{j+1} = \begin{cases} X_j(Xn + 1 - w) & j^{\text{th}} \text{ throw is H} \\ X_j(1-w) & j^{\text{th}} \text{ throw is T} \end{cases}$$

Put $Y_{j+1} = \frac{X_{j+1}}{X_j} = \begin{cases} wn + (1-2) & H \\ 1 - w & T \end{cases}$.

Then we try to maximise the log of our fortune: let $\mathbb{E}\log Y_i = \bar{\mu}4, \text{Var}(\log Y_i) = \bar{\sigma}^2$. If $a > 0$ then:

$$\mathbb{P}\left(\left|\frac{\log Y_1 + \ldots + \log Y_n}{n} - \bar{\mu}\right| \geq a\right) \leq \frac{\bar{\sigma}^2}{na^2}$$

by Chebyshev's inequality. But $\sum \log Y_i = \log X_i$, so we have:

$$\mathbb{P}\left(\left|\frac{\log X_n}{n} - \bar{\mu}\right| \geq a\right) \leq \frac{\bar{\sigma}^2}{na^2}$$

i.e., for all $\epsilon > 0$, $\delta > 0$, there is some N such that

$$\mathbb{P}\left(\left|n^{-1}\log(X_n) - \bar{\mu}\right| \geq \delta\right) \leq \epsilon \ \ \forall\, \mathbb{N}$$

Lemma 10.1. *Consider one single toss of a coin with $\mathbb{P}(H) = p < 1$. Suppose that a bet on heads has payout ratio of n. Suppose that we have a bankroll of 1 unit and we bet w on H, retaining $1 - w$ for $0 \leq 1 \leq 1$. If Y is the expected value of our fortune after the throw then*

$$\mathbb{E}(\log Y) = p\log(1 + (n-1)w) + (1-p)\log(1-w)$$

The value of $\mathbb{E}(\log Y)$ is maximised by taking $w = 0$ if $np \leq 1$ and setting $w = \frac{np-1}{n-1}$ if $np > 1$.

A better who maximises the log of his or her fortune is called a ***Kelly better***, and is following ***Kelly's rule***, since Kelly in 1956 showed how to do this using information theory.