Information Theory I Lecture Notes

Victor Kawasaki-Borruat, Kostas Vergo

HS2022, ETH Zürich

Contents

1 Introductory Formulae

This section will cover the basic notations, formulae and 'characters' of the course.

1.1 Notation

We will only consider finite alphabets of symbols.

- $\mathcal{X}$ denotes a finite set of symbols
- $x \in \mathcal{X}$ is an element of $\mathcal{X}$
- X denotes a *chance variable* (is not a RV, as it does not necessarily take values in $\mathbb{R}$, and thus do not have an expected value)

1.2 Entropy and Friends

Definition 1.1 (Entropy)**.** *The entropy of a chance variable X is commonly referred to as the uncertainty of X.(It's not a function of X)*

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log(\frac{1}{P_X(x)}) \tag{1.1}$$

Also it is expressible as a function of X's probability mass function P_X as $H(P_X)$.

$$H(X) = H(P_X) = E[\log(\frac{1}{P_X})] \tag{1.2}$$

Example 1. *The setup*

$$\mathcal{X} = \{H, T\}, P_X(X) = p, P_X(T) = (1-p) \tag{1.3}$$

Yields the following entropy

$$H(X) = p\log(\frac{1}{p}) + (1-p)\log(\frac{1}{1-p}) \tag{1.4}$$

Remark 1. *Since the PMF of a chance variable can be viewed as a vector, a permutation of the values does not affect the entropy*

Definition 1.2 (Self Information)**.** *The self-information of a chance variable X is defined as*

$$I_X(a) = \log(\frac{1}{P_X(a)}) \tag{1.5}$$

Thus

$$H(X) = E[I_X(x)] \tag{1.6}$$

Proposition 1.3 (Entropy Bounds)**.** *A few properties of entropy:*

1. *$H(X) \geq 0$ w.eq. iff X is deterministic*
2. *$H(X) \leq \log(|\mathcal{X}|)$ w.eq. iff X is uniformly distributed*

Proof. 1. $P_X(x)\log(P_X(x)) \geq 0$ and is zero if either $P_X(x) = 0$ or $\log(P_X(x)) = 1$ thus $P_X(x) = 1 \rightarrow log(P_X(x)) = 0$

2. Let $P_X(x)$ and $P_{uni}(x) = \frac{1}{|\mathcal{X}|}$.

$$D(P_X(x)||P_{uni}(x)) = \sum_{x\in\mathcal{X}} P(x)\log(\frac{P_X(x)}{P_{uni}(x)}) = \sum_{x\in\mathcal{X}} P(x)\log(\frac{P_X(x)}{\frac{1}{|\mathcal{X}|}}) = \sum_{x\in\mathcal{X}} P(x)\log(|\mathcal{X}|P_X(x))$$

So, by splitting the *log* we have

$$D(P_X(x)||P_{uni}(x)) = \sum_{x\in\mathcal{X}} P(x)\log(|\mathcal{X}|) + \sum_{x\in\mathcal{X}} P(x)\log(P_X(x)) = \log(|\mathcal{X}) - \sum_{x\in\mathcal{X}} P(x)\log(\frac{1}{P_X(x)})$$

We will prove later that $D(P_X(x)||P_{uni}(x)) \geq 0 \Rightarrow H(X) \leq \log(|\mathcal{X}|)$

□

Definition 1.4 (Relative Entropy)**.** *Relative entropy of two PMFs P and Q is defined as*

$$D(P||Q) = \sum_{x\in\mathcal{X}} P(x)\log(\frac{P(x)}{Q(x)}) = E_P[\log(\frac{P}{Q})] \tag{1.7}$$

or is infinity if there exists $y \in \mathcal{X}$ s.t. $Q(x) = 0$ and $P(x) > 0$.

Remark 2. *It's obvious that $D(P||Q) \neq D(Q||P)$. So it's not symmetric, thus it can not be used as a metric.*

Definition 1.5 (Concavity)**.** *A function f is said concave if it satisfies $\forall\lambda \in [0, 1]$*

$$f(\lambda x_0 + (1-\lambda)x_1) \geq \lambda f(x_0) + (1-\lambda)f(x_1) \tag{1.8}$$

Proposition 1.6 (Jensen's Inequality)**.** *Let X be. a RV s.t. $P[X = x_0] = \lambda$ and $P[X = x_1] = 1-\lambda$, then*

$$f(E[X]) \geq E[f(X)] \tag{1.9}$$

for any concave function f

Proof. Using a second-order Taylor expansion, we get

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(\eta), \eta \in [x, x_0] \tag{1.10}$$

Since f is concave, $f''(x) \leq 0$, thus

$$E[f(x)] \leq E[f(x_0)] + E[(x - x_0)f'(x_0)] \tag{1.11}$$

By setting $E[X] = x_0$, we get $E[x - x_0] = 0$ thus

$$E[f(x)] \leq f(x_0) = f(E[x]) \tag{1.12}$$

□

Theorem 1.7.

$$D(P||Q) \geq 0 \tag{1.13}$$

with equality iff $P = Q$

Proof.

$$-D(P||Q) = \sum_x P(x) \log(\frac{P(x)}{Q(x)}) = E_P[\log(\frac{P(x)}{Q(x)})] \tag{1.14}$$

by applying Jensen's inequality due to the logarithm's concavity,

$$-D(P||Q) \leq \log(E_P[\frac{P(x)}{Q(x)}]) = \log(\sum_x P(x)\frac{Q(x)}{P(x)}) = \log(\sum_x Q(x)) = 0 \tag{1.15}$$

□

Proposition 1.8 (Log-Sum Inequality)**.** *Let $a_i \geq 0$, $b_i \geq 0$, then*

$$\sum_i a_i \log(\frac{a_i}{b_i}) \geq (\sum_i a_i) \log(\frac{\sum_i a_i}{\sum_i b_i}) \tag{1.16}$$

Proof. Let $a = \sum_i a_i$ and $b = \sum_i b_i$, then $\frac{a_i}{a}$ and $\frac{b_i}{b}$ are PMFs. Hence

$$D(\frac{a_i}{a}||\frac{b_i}{b}) = \sum_i \frac{a_i}{a} \log(\frac{\frac{a_i}{a}}{\frac{b_i}{b}}) \geq 0$$

implying

$$\sum_i \frac{a_i}{a} \log(\frac{a_i}{b_i}) \geq a \log(\frac{a}{b})$$

□

Definition 1.9 (Joint Entropy)**.** *Of two chance variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ is the entropy of the chance variable $Z = (X, Y) \in (\mathcal{X} \times \mathcal{Y})$*

$$H(X,Y) = H(P_{XY}) = \sum_x \sum_y P_{XY}(x,y) \log(\frac{1}{P_{XY}}) \tag{1.17}$$

Definition 1.10 (Conditional Entropy)**.** *Of chance variables X and Y is given by:*

$$H(X|Y=y) = \sum_x P_{X|Y=y}(x) \log(\frac{1}{P_{X|Y=y(x)}}) \tag{1.18}$$

if Y is known, otherwise

$$H(X|Y) = \sum_y P_Y(y) H(X|Y=y) = E_Y[H(X|Y=y)] \tag{1.19}$$

Remark 3. *Quick reminder that the conditional $P_{X|Y=y}(x)$ can be calculated based on the relation*

$$P_{X|Y=y_0}(x) = \frac{P_{XY}(x,y_0)}{P_Y(y_0)}$$

Theorem 1.11 (Chain Rule)**.**

$$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X) \tag{1.20}$$

i.e. the uncertainty of X and Y is the uncertainty of Y plus the uncertainty of X once Y has been observed and vice-versa.

Remark 4. *The foregoing relation can easily be generalized for $X^n = (X_1, X_2, ..X_n)$ variables.*

$$\begin{aligned} H(X^n) &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + ... + H(X_n|X_{n-1}, X_{n-2}...X_1) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}...X_1) =_{notation} \sum_{i=1}^{n} H(X_i|X^{i-1}) \end{aligned}$$

Proof.

$$E[\log(\frac{1}{P_{XY}(x,y)})] = E[\log(\frac{1}{P_Y P_{X|Y}})] = E[\log(\frac{1}{P_Y})] + E[\log(\frac{1}{P_{X|Y}})]$$

□

Remark 5 (Independent RVs' Entropy)**.** *When X and Y are independent, then*

$$H(X,Y) = H(X) + H(Y) \tag{1.21}$$

Definition 1.12 (Mutual Information)**.** *Informally, mutual information $I(X;Y)$ of two chance variables denotes the amount of information one gives about the other.*

1. *$H(X) - H(X|Y)$ (this implies that $I(X;Y) \leq log(|\mathcal{X}|)$)*
2. *$H(Y) - H(Y|X)$*
3. *$H(X) + H(Y) - H(X,Y)$ (this implies symmetry, i.e. $I(X;Y) = I(Y;X)$)*

4. $D(P_{XY}||P_XP_Y)$ *(thus* $I(X;Y)=0$ *iff* $X \perp\!\!\!\perp Y$*)*

Proof. 1. by definition

2.

$$I(X;Y) = H(X) - H(X|Y) = H(X) - [H(X,Y) - H(Y)] = H(X) + H(Y) - H(X,Y)$$

3.

$$\begin{aligned} D(P_{XY}||P_XP_Y) &= \sum_{X,Y} P_{XY} \log(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}) \\ &= \sum_{X,Y} P_{XY} \log(P_{X,Y}(x,y)) + \sum_{X,Y} P_{XY} \log(\frac{1}{P_X(x)}) + \sum_{X,Y} P_{XY} \log(\frac{1}{P_Y(y)}) \\ &= \sum_{X,Y} P_{XY} \log(P_{X,Y}(x,y)) + \sum_{X} P_X \log(\frac{1}{P_X(x)}) + \sum_{Y} P_Y \log(\frac{1}{P_Y(y)}) \\ &= -E[\log(\frac{1}{P_{XY}})] + E[\log(\frac{1}{P_X})] + E[\log(\frac{1}{P_Y})] \end{aligned}$$

□

A very important result from mutual information, is that **conditioning reduces entropy**, i.e.

$$H(X) \geq H(X|Y) \tag{1.22}$$

with equality iff $X \perp\!\!\!\perp Y$. Knowing Y can only remove uncertainty about X, not add any.

Remark 6. *This does NOT apply to* $H(X|Y=y)$*. It could very well be greater than* $H(X)$*. All of* Y *must be observed to affirm that conditioning reduces entropy.*

To better illustrate it, consider a distribution $P[X=0]=\frac{1}{2}, P[X=1]=\frac{1}{2}$ *conditional on* $Y=y$*, but the average conditional distribution is* $P[X=0]=\frac{1}{4}, P[X=1]=\frac{3}{4}$*. The distribution of* $X|Y=y$ *has greater entropy than* $X|Y$*.*

Definition 1.13 (Mixture of PMFs)**.** *Let* P, Q *be PMFs and* $\lambda \in [0,1]$*. Then* $W = \lambda P + (1-\lambda)Q$ *is a convex combination aka mixture of* P *and* Q*. It is a well-defined PMF as well.*

Proposition 1.14.

$$H(W) = H(\lambda P + (1-\lambda)Q) \geq \lambda H(P) + (1-\lambda)H(Q) \tag{1.23}$$

Proof. Let

$$E = \begin{cases} 1 & w.p.\lambda \\ 0 & w.p.1-\lambda \end{cases}$$

and $P(X|E=1)=P$, $P(X|E=0)=Q$. Then

$$P_X(x) = P(E=0)Q(x) + P(E=1)P(x) = W(x)$$

and

$$H(X) = H(W) \geq H(X|E)$$

by the conditioning of entropy. □

Definition 1.15 (Conditional Mutual Information)**.** *If Z is known and has value $Z = \zeta$ then*

$$I(X;Y|Z=\zeta) = H(Y|Z=\zeta) - H(Y|X,Z=\zeta) \tag{1.24}$$

Otherwise

$$\begin{aligned} I(X;Y|Z) &= \sum_z P_Z(z) I(X;Y|Z=z) \\ &= \sum_z P_Z(z)(H(X|Z=z) - H(X|Y,Z=z)) \\ &= H(X|Z) - H(X|Y,Z) \end{aligned} \tag{1.25}$$

$I(X;Y|Z) \geq 0$, with equality iff $X \perp\!\!\!\perp Y$ conditionally on Z, i.e. iff X, Y, Z form a **Markov Chain**.

Definition 1.16 (Chain Rule for Conditional Mutual Information)**.**

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) \tag{1.26}$$

Proof. We consider X, Y as one random variable and write:

$$I(X,Y;Z) = H(X,Y) - H(X,Y|Z) \tag{1.27}$$

and apply the chain rule

$$= H(X) - H(X|Z) + H(Y|X) + H(Y|X,Z) = I(X;Z) + I(Y;Z|X) \tag{1.28}$$

□

Remark 7. *All in all you have to remember the next 3 equalities.*

- $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$
- $I(X_1, X_2, ...X_n; Y) = \sum_i I(X_i; Y|X_1, ..., X_{i-1})$
- $I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$

Proposition 1.17 (Fano's Inequality)**.** *Let $\hat{U}, \hat{U}$ be L-ary random variables taking value in the same alphabet, and let the error probability $P[U \neq \hat{U}] = P_e$. Then*

$$H_b(P_e) + P_e \log(L-1) \geq H(U|\hat{U}) \tag{1.29}$$

where H_b is the binary entropy function.

Proof:
Let Z be a random variable such that

$$Z = \begin{cases} 1, \text{ if } U \neq \hat{U} \\ 0, \text{ if } U = \hat{U} \end{cases}$$

Thus,

$$\begin{aligned} H(U,Z|\hat{U}) &= H(U|\hat{U}) + H(Z|U,\hat{U}) = H(U|\hat{U}) = H(Z|\hat{U}) + H(U|\hat{U},Z) \\ &\leq H(Z) + P_Z(1)H(U|\hat{U}, Z=1) + P_Z(0)H(U|\hat{U}, Z=0) \\ &\leq H_b(P_e) + P_e \log(L-1) \end{aligned}$$

by using the upper bound on entropy and the fact that the alphabet is reduced to size $L-1$ since we presume $\hat{U}$ is known.

2 Source Coding / Data Compression

This section will cover efficient coding of a single random variable. **source coding** refers to the mapping of symbols (from an information source $\mathcal{X}$) to a set of alphabet symbols $\mathcal{D}$.

2.1 Basic Definitions

Definition 2.1. *A **source code** $\mathcal{C}$ for a random variable X is a mapping from $\mathcal{X}$ to $\mathcal{D}^*$.*

Definition 2.2. *The **expected length** of a source code $C(x)$ for a random variable X of PMF P_X is defined by*

$$L = \sum_i p_X(x_i) l(x_i) \tag{2.1}$$

where $l(\cdot)$ denotes the length of a codeword.

Definition 2.3. *A **singular code** is a code that has the same codeword for two different sources.*

Definition 2.4. *The **extension** of a code $C(x)$ is a mapping from finite-length strings in $\mathcal{X}$ to finite-length strings in $\mathcal{D}^*$.*

$$C^*(x_1 x_2 ... x_n) = C(x_1)C(x_2)...C(x_n) \tag{2.2}$$

Definition 2.5. *A **uniquely decodable** code C is a code which it's extension C^* non-singular i.e*

$$C^*(x_1, x_2, ..x_n) = C(x_1)C(x_2)...C(x_n)$$

Definition 2.6. *A **prefix-free** code is a code whose codewords are only leaves of the D-ary tree representing it. It is not singular, not ambiguous and also gives the shortest coding. More formally, C is a prefix free code if no codeword is a prefix of other codeword.*

Proposition 2.7. *Every prefix free code is uniquely decodable. Whereas the reverse doesn't hold*

Lemma 2.8 (Leaf Counting and Depth)**.** *The number of leaves (n) and their depths $\{l_i\}_{i=1}^n$ in a D-ary tree satisfies*

$$n = N(D-1) + 1 \tag{2.3}$$

where $N$is the number of nodes (root included), and

$$\sum_{i=1}^{n} D^{-l_1} = 1 \tag{2.4}$$

Proof:
Induction. Pretty straightforward.

Proposition 2.9 (Kraft's Inequality)**.** *There exists a D-ary prefix-free code wth r codewords of lengths $l_1, ..., l_r$ **if and only if***

$$\sum_{i=1}^{r} D^{-l_i} \leq 1 \tag{2.5}$$

If the equality holds, then there are no unused leaves remaining in the tree.

This proposition offers a practical property on any prefix-free code. Consider a D-ary tree of any D-ary prefix-free code for an r-ary random message U. Let $w_1, w_2, ..., w_n$ be the leafs to this tree, ordered such that w_i is the leaf corresponding to the codeword message u_i, $i = 1, 2.., r$ and let $w_{r+1}, w_{r+2}...., w_n$ are unused. Define two PMF's such that

$$P_w(w_i) = \begin{cases} P_u(u_i) = p_i & i = 1, 2, 3..., r \\ 0 & i = r+1, ..., n \end{cases}$$

$$P_{\tilde{w}}(w_i) = D^{-l_i}, i = 1, 2, ..., n$$

Proposition 2.10. *The expected length of a prefix free code of messages $\{u\}$ is always upper bounded as follows:*

$$\frac{H(u)}{\log(D)} \leq E[L] \tag{2.6}$$

Proof. We can compute

$$\begin{aligned} D(P_w||P_{\tilde{w}}) &= \sum_{i=1}^{r} P_w(w_i) \log(\frac{P_w(w_i)}{P_{\tilde{w}}(w_i)}) \\ &= \sum_i P_w(w_i) \log(P_w(w_i)) - \sum_i P_w(w_i) \log(P_{\tilde{w}}(w_i)) \\ &= \sum_i p_i \log(p_i) + \sum_i p_i l_i \log(D_i) \\ &= -H(u) + \log(D) E[L] \geq 0 \\ &= E[L] \geq \frac{H(u)}{\log(D)} \end{aligned}$$

□

2.2 Kraft's Inequality and Minimum Coding Length

P.S: we only talk about binary codes but the same hold for D-ary codes.

Proposition 2.11 (Kraft's Inequality)**.**

1) The lengths $l_1, l_2, ..., l_n$ of every uniquely decodable binary code are integers satisfying the following:

$$\sum_{i=1}^{r} 2^{-l_i} \leq 1 \tag{2.7}$$

2) Given $l_1, l_2, ..., l_n$ integers that satisfy

$$\sum_{i=1}^{r} 2^{-l_i} \leq 1 \tag{2.8}$$

then there is a prefix free code of these lengths.

Remark 8. *This inequality states that everything that a uniquely decodable code can achieve always there is an equivalent prefix free code which can do the same.*

Proof. We will prove the claim in two parts:

- If $\sum 2^{l_i} \leq 1$, then there is a prefix free code with the same l_i lengths.
  Assume w.l.o.g that $l_1 \leq l_2 \leq l_3 ... \leq l_m$. By construction we can create a tree till $l_1, l_2 ..., l_k$. We are sure that we can find leaves for $l_{k+1}, l_{k+2} ..., l_m$
  In depth l_k the tree has 2^{l_k} leaves. In depth l_k due to the child at length l_1 there are $2^{l_k - l_1}$ leaves which can't be used. In similar fashion due to the child at depth l_2 there are $2^{l_k - l_2}$ leaves which can be used. At level l_{k-1} there are $2^{l_k - l_{k-1}}$ children which can be used. Ruled out all the phantom leaves:

$$\sum_{i=1}^{k-1} 2^{l_k - l_i} < 2^{l_k} \Rightarrow \sum_{i=1}^{k-1} 2^{-l_i} < 1, \textit{So if } \sum 2^{l_i} \leq 1 \textit{the tree can expand}$$

- Let the code $C : \mathcal{X} \to \{0,1\}^\tau$, and $l(X)$ is the length of the $C(X)$ for each symbol x.

$$\begin{aligned}
\left(\sum_{x \in \mathcal{X}} 2^{-l(x)}\right)^m &= \left(\sum_{x \in \mathcal{X}} 2^{-l(x)}\right)\left(\sum_{x \in \mathcal{X}} 2^{-l(x)}\right) ... \left(\sum_{x \in \mathcal{X}} 2^{-l(x)}\right) \\
&= \left(\sum_{x_1 \in \mathcal{X}} \left(\sum_{x_2 \in \mathcal{X}} ... \left(\sum_{x_m \in \mathcal{X}} 2^{-l(x_1)} 2^{-l(x_2)} 2^{-l(x_m)}\right)\right)\right) \\
&= \sum_{\overline{x} \in \mathcal{X}^m} 2^{-\sum l(x_i)} = \sum_{k=1}^{ml_{max}} a(k) 2^{-k} \leq \sum_{k=1}^{ml_{max}} 2^k 2^{-k} = ml_{max} \\
&\Rightarrow \sum_{x \in \mathcal{X}} 2^{-l(x)} \leq (ml_{max})^{\frac{1}{m}}, \ \textit{also} \lim_{m \to \infty} (ml_{max})^{\frac{1}{m}} \to 1 \\
&\textit{Hence,} \sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1
\end{aligned}$$

□

Definition 2.12. *An **optimal code** is a code C that has minimal expected length.*

Proposition 2.13. *Let $l_1, l_2, l_3, ..., l_n$ be the lengths of the symbols $x_1, x_2, ..., x_n$. It can be shown that*

$$H(X) \leq L^* \leq H(X) + 1 \tag{2.9}$$

where

$$\underset{u.d \ codes}{\arg\min} \sum_{i=1,..n} p_i l_i = \underset{\substack{l_1, l_2, l_3 ... \in \mathbf{N} \\ \sum 2^{-l_i} \leq 1}}{\arg\min} \sum p_i l_i = L^*$$

Proof. Relax the problem to the following

$$\underset{\substack{l_1, l_2, l_3 ... \in \mathbf{R} \\ \sum 2^{-l_i} \leq 1}}{\arg\min} \sum p_i l_i = L_R^*$$

Using Lagrange multipliers we can prove that the L_R^* is achieved for $l_i^* = \log(\frac{1}{p_i})$ So, $L_R^* = H(X) \leq L^*$, the equality holds if $\log(\frac{1}{p_i})$ are integers.
To upper bound it we can take $\lceil \log(\frac{1}{p_i}) \rceil$. Hence we got

$$\sum p_i l_i = \sum p_i \lceil log(\frac{1}{p_1}) \rceil < \sum p_i (log(\frac{1}{p_1}) + 1) = H(P) + 1$$

□

Proposition 2.14. *The plus 1 in the foregoing equation sometimes is not convenient. We can avoid this term using k-to-variable codes.*
$C : X \to \{0,1\}^t$ *one-to-variable code.*
$C : X^k \to \{0,1\}^t$ *k-to-variable code.*

$$H(X_1, X_2, ..X_k) \leq L^* \leq H(X_1, X_2, ..X_k) + 1 \overset{i.i.d}{\Rightarrow} kH(X) \leq L^* \leq kH(X) + 1$$
$$\Rightarrow H(X) \leq \frac{L^*}{k} \leq H(X) + \frac{1}{k}$$

Example:
We have symbols $X_1, X_2...$ with true probability mass function equal to $\{P_i\}$ but we design our code with wrong probability mass function equal to $\{Q_i\}$ and thus we choose $l_i = \lceil log(\frac{1}{q_i}) \rceil$. So the error that we will pay is $D(P||Q)$

$$E[L] = \sum p_i log(\frac{1}{q_i}) = \sum p_i log(\frac{p_i}{q_i p_i}) = \sum p_i \left(log(\frac{p_i}{q_i}) + log(\frac{1}{p_i}) \right) = D(P||Q) + H(P)$$

2.3 Huffman Procedure

The Huffman procedure is a way of constructing optimal codes (Huffman Codes).

2.3.1 Construction of Huffman Codes

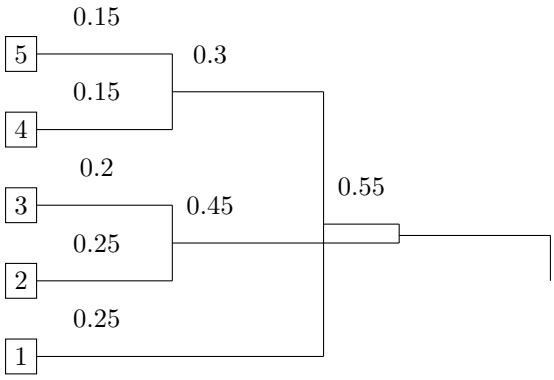
Without loss of generality, we will assume that our codewords are ordered such that

$$p_1 \geq p_2 \geq ... \geq p_n$$

For binary codes, we start with the two lowest-probable codewords, and 'join' them by adding probabilities. We now consider their joined node as a codeword and repeat the procedure by finding the next lowest-probable codeword, joining etc...

Example:

Let $\mathcal{X} = \{(1|0.25), (2|0.25), (3|0.2), (4|0.15), (5|0.15)\}$ be an ensemble, we will construct the Huffman code corresponding to this source:



For ternary codes, we can add 'extra' nodes with probability zero to fit the model adequately.

2.4 Typicality

This section will discuss the three types of typicality, from most intuitive and mathematically complex to least intuitive but mathematically simple.

2.4.1 Empirical Type

Let

$$N(a|X) = \sum_{i=1}^{n} I\{x_i = a\} \tag{2.10}$$

count the number of outcomes of X matching a.

From $N(a|X)$, we can construct an **empirical type**, which corresponds to a PMF:

$$P_X(a) = \frac{1}{n} N(a|X) \tag{2.11}$$

We can thus say that X **is of type** P **if and only if** $\frac{1}{n}N(aX) = P(a)$

2.4.2 Strong and Weak Typicality

Strong typicality is a relaxation of the constraint on empirical typicality.

Example:

Let A be a sequence of a fair coin, of outcomes $\in \{H, T\}$, which $P = (\frac{1}{2}, \frac{1}{2})$.

Then all sequences of empirical type P are all sequences in which H and T occur the same number of times, which is rather limited.

Definition 2.15 (Strongly Typical Set)**.**

$$T_\epsilon^{(n)}(P) = \{x \in \mathcal{X} | \frac{1}{n} N(a|X) - P(a) \leq \epsilon P(a)\} \tag{2.12}$$

This allows for sequences where H occurs almost 50% of the time.

Definition 2.16 (Weakly Typical Set)**.**

$$\mathcal{A}_\epsilon^{(n)}(P) = \{\xi \in \mathcal{X}^n | 2^{-n(H(P)+\epsilon)} \leq \Pi_{i=1}^n P(\xi_i) \leq 2^{-n(H(P)-\epsilon)}\} \tag{2.13}$$

The elements of $\mathcal{A}_\epsilon^{(n)}(P)$ are said to be ***weakly typical w.r.t.*** *P.*

Remark 9. *Notice that*

$$T_\epsilon^{(n)}(P) \subseteq \mathcal{A}_\epsilon^{(n)}(P) \tag{2.14}$$

Also note that if $X_1, X_2, ..., X_n \sim_{i.i.d.} P$, then the sequence $(x_1, x_2, ..., x_n)$ can be in $\mathcal{A}_\epsilon^{(n)}(P)$.

2.5 Asymptotic Equipartition Property (AEP)

This is the tool we will mostly use. Recall the **weak law of large numbers**, which states that for $X_1, ..., X_n \sim_{i.i.d.} P$ then the empirical average converges toward the expectation, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mathbb{E}[X_1] \quad \text{as} \quad n \to \infty \tag{2.15}$$

The AEP is very analogous, but with the entropy $H(P)$ instead of the expected value.

Theorem 2.17 (Asymptotic Equipartition Property)**.** *Let $X_1, X_2, ...X_n \sim_{i.i.d.} P$. Then*

$$\frac{1}{n}\log(\frac{1}{p(x_1, x_2, ...x_n)}) = H(P), \quad n \to \infty \tag{2.16}$$

in other words,

$$p(x_1, x_2, ...x_n) = 2^{-n(H(P)\pm\epsilon)}, \quad n \to \infty \tag{2.17}$$

Proof. By considering the random variable $Y_i = -\log(p(X_i = x_i))$, and the fact that X_i are i.i.d., then we can write out

$$\frac{1}{n}\log(\frac{1}{p(x_1, x_2, ...x_n)}) = -\frac{1}{n}\sum_{i=1}^{n} P(x_i) = -\frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2.18}$$

which by weak law of large numbers yields

$$-\frac{1}{n}\sum_{i=1}^{n} Y_i \to -\mathbb{E}[Y_i] = \mathbb{E}][-\log(P(x_i))] = H(P) \tag{2.19}$$

□

Lemma 2.18 (Convergence of AEP)**.** *If $X_1, X_2, ..., X_n \sim P$ are i.i.d., then*

$$P[x \in \mathcal{A}_\epsilon^{(n)}(P)] \to_{n\to\infty} 1 \tag{2.20}$$

i.e. all sequences of i.i.d chance variables will all find themselves in this weakly typical set for n large enough.

Proof.

$$P[2^{-n(H(P)+\epsilon)} \leq \Pi_{i=1}^{n} P(x_i) \leq 2^{-n(H(P)-\epsilon)}]$$

$$= P[-(H(P)+\epsilon) \leq \frac{1}{n}\sum_{i=1}^{n} log(P_i) \leq -(H(P)-\epsilon)]$$

$$= P[H(P)+\epsilon \geq \frac{1}{n}\sum_{i=1}^{n} log(\frac{1}{P_i}) \geq H(P)-\epsilon]$$

$$= P[-\epsilon \leq \frac{1}{n}\sum_{i=1}^{n} log(\frac{1}{P_i}) - H(P) \leq \epsilon] \to 1$$

□

Lemma 2.19 (Bounds on size of weakly typical set)**.** *We have two bounds for the cardinality of* $\mathcal{A}_\epsilon^{(n)}(P)$*:*

- $|\mathcal{A}_\epsilon^{(n)}(P)| \leq 2^{n(H(P)+\epsilon)}$
- $2^{n(H(P)-\epsilon)} \leq |\mathcal{A}_\epsilon^{(n)}(P)|$

Proof. The upper bound:

$$1 \geq P^{\times n}(\mathcal{A}_\epsilon^{(n)}(P)) = \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} P^{\times n}(\xi) \geq \sum_{\mathcal{A}_\epsilon^{(n)}(P)} 2^{-n(H(P)+\epsilon)} = |\mathcal{A}_\epsilon^{(n)}(P)| 2^{-n(H(P)+\epsilon)}$$

which implies

$$2^{n(H(P)+\epsilon)} \geq |\mathcal{A}_\epsilon^{(n)}(P)|$$

The lower bound:

$$1 - \epsilon \leq P^{\times n}(\mathcal{A}_\epsilon^{(n)}(P)) = \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} P^{\times n}(\xi) \leq \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} 2^{-n(H(P)-\epsilon)} = |\mathcal{A}_\epsilon^{(n)}(P)| 2^{-n(H(P)-\epsilon)}$$

□

2.5.1 Consequences of AEP

$\mathcal{A}_\epsilon^{(n)}(P)$ is rather small within the set of sequences of i.i.d. $\{X_i\}_{i=1}^n$, but contains **most of the probability**. Moreover, since there are less than $2^{n(H(P)+\epsilon)}$ in $\mathcal{A}_\epsilon^{(n)}$, indexing them requires no more than $\lceil n(H(P)+\epsilon)+1 \rceil$ bits. **This is the real data compression implication!** Each code in $\mathcal{A}_\epsilon^{(n)}$ have a short description of bits, which is 'optimal'. Other codes not in $\mathcal{A}_\epsilon^{(n)}$ are 'brute force' indexed, with a flag bit prepended to them to signal they are not in $\mathcal{A}_\epsilon^{(n)}$. *Add a photo or something to show diagrammatically this important implication!*

3 Channels and Capacity

This section will cover the concepts of data transmission through a (potentially noisy / lossy) environment.

Remark 10. *We will only consider* ***discrete-time channels****, as well as* ***finite alphabets****! Moreover, no cost-constraints are taken into account.*

Definition 3.1 (Channel)**.** *A channel is the medium that will transmit information from source to target. In our context, it is the part of the communication system that we cannot change / tamper with. It is out of our control.*

Definition 3.2 (Discrete Memoryless Channel)**.** *A DMC is a channel specified by:*

- *an input alphabet* $\mathcal{X}$
- *an output alphabet* $\mathcal{Y}$
- *a conditional probability distribution* $P_{Y_n|X_n}(y_n|x_n, x_{n-1}, ..., x_1) = P_{Y_n|X_n}(y_n|x_n)$

such that the conditional distribution yields a ***channel law matrix*** $W_{Y|X}$

Remark 11. *For a DMC, the channel transition function can be expressed as such:*[1]

$$W(y^n|x^n) = \Pi_{i=1}^n W(y_i|x_i)$$

where the superscript denotes a sequence of n chance variables.

Definition 3.3 (Channel Law Matrix)**.** *The channel law matrix* $W_{Y|X}$ *satisfies the following:*

- $W(y|x) \geq 0, \forall y, x$
- $\sum_{y\in\mathcal{Y}} W(y|x) = 1$

It is built by considering the X's in the rows, and the Y's in the columns. Thus, all rows sum to 1.

Thus for any **source distribution** $Q(\cdot)$, we have:

- $Q \in \mathcal{P}(\mathcal{X})$
- $(Q \circ W)(x, y) = Q(x)W(y|x)$, it's the joint distribution of X, Y
- $\sum_x Q(x)W(y|x) = 1$, it's the pmf of Y

3.1 Channel Coding

To counter the naive intuition of sending bits via *repetition codes*, Shannon came up with the idea of sending blocks of codes.

[1] the factorization is due to the memoryless property

3.1.1 Definitions

Definition 3.4. *A (M, n) code for a channel $W_{Y|X}$ consists of*

1. *a **message set** $\mathcal{M}$*
2. *an **encoder** $f^n : \{1, 2, ..., \mathcal{M}\} \to \mathcal{X}^n$ which yields codewords of length n*
3. *a **decoder** $\phi^n : \mathcal{Y}^n \to \{1, 2..., \mathcal{M}\}$*

Definition 3.5. *A **codebook** is the image of an encoder over all possible messages, denoted*

$$x^n(1), x^n(2), ..., x^n(\mathcal{M}) \tag{3.1}$$

We typically think of a codebook as a $|\mathcal{M}| \times n$ matrix, where the m^{th} row is $x^n(m)$.

Definition 3.6. *The **rate** of a channel is defined by*

$$R = \frac{k}{n}[\frac{bits}{channel\ uses}] \tag{3.2}$$

Remark 12. *Notice that if we are considering a message space of $|\mathcal{M}|$ messages, then*

$$R = \frac{\log_2(|\mathcal{M}|)}{n} \tag{3.3}$$

The key idea here is to have a **fixed rate**, and then fit as much as we can given that constraint.

Definition 3.7 (Probability of error)**.** *Let λ_m denote the error of transmission of information, i.e.*

- $\lambda_m = \sum_{y:\phi(y)\neq m} \Pi_{i=1}^n W(y_i|x_i(m)) = \sum_{y:\phi(y)\neq m} W(Y^n|X^n)$
- $\lambda_{max} = \max_m\{\lambda_m\}$

i.e. λ_m is the sum of probabilities that W will map $X(m)$ to another Y.

Definition 3.8 (Arithmetic Average Probability of Error)**.** *is denoted with $P_e^{(n)}$, and is defined by:*

$$P_e^{(n)} = \frac{1}{|M|}\sum_{i=1}^{n} \lambda_i \tag{3.4}$$

Definition 3.9. *A rate R is said **achievable** on a discrete memoryless channel if $\exists$ a sequence of codes (f_n, ϕ_n) indexed by the blocklength of rate $R = \frac{\log(M)}{n} = \frac{\lceil 2^{nR} \rceil}{n}$ such that $\lambda_{max} \to 0$. I.e. the maximal probability of error tends to zero as $n \to \infty$.*

Definition 3.10. *The **capacity** of a discrete memoryless channel W is the supremum of achievable rates, denoted C.*

Theorem 3.11. *The **information capacity** of a channel W, when fed an input sequence of distribution Q is given by:*

$$C^{(I)} = \max_Q\{I(Q, W)\} \tag{3.5}$$

which can also be written as

$$C = \max_Q\{I(X;Y) = H(X) - H(X|Y)\} \tag{3.6}$$

for input X and output Y.

Example: [Binary Symmetric Channel]
Assuming the crossover probability is ϵ,

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x Q(x)H(Y|X=x) \tag{3.7}$$

$$= H(Y) - \sum_x Q(x)H_b(\epsilon) = H(Y) - H_b(\epsilon) \leq 1 - H_b(\epsilon)$$

The mutual information is maximized when $X \sim Ber(\frac{1}{2})$.

Definition 3.12 (Weakly Symmetric Channel)**.** *A weakly symmetric channel is a channel whose matrix' rows are permutations of each other, and all columns sum to the same value. For a weakly symmetric channel, the capacity is given by:*

$$C_{WSC} = \log(|\mathcal{Y}|) - H(row) \tag{3.8}$$

Proof. We start with any Q.

$$\mathcal{I}(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x Q(x)H(Y|X=x) = H(Y) - \sum_x Q(x)H(rowX=x)$$
$$\implies H(Y) - H(row) \leq_i \log(|\mathcal{Y}|) - H(row)$$

To hold i) with equality we have to choose Y to be uniform. If we choose X uniform we get

$$P(Y=y) = \sum_x Q(x)W(Y=y|X=x) = \sum_x \frac{1}{|\mathcal{X}|}W(Y=y|X=x) = \frac{1}{|\mathcal{X}|}sum(of\ column\ y) = \frac{t}{|\mathcal{X}|}$$

Another argument would be that due to symmetry it is clearly that $Q^*(0) = Q^*(1) = \frac{1}{2}$ achieves capacity. To prove this assume that for any a $Q_1 = (a, 1-a)$ achieves capacity. Then, due to symmetry $Q_2 = (1-a, a)$ achieves also capacity. Thus $C = \mathcal{I}(Q_1, W) = \mathcal{I}(Q_2, W)$. Mutual information is concave and thus we get that

$$\mathcal{I}(\frac{1}{2}Q_1 + \frac{1}{2}Q_2; W) \geq \frac{1}{2}\mathcal{I}(Q_1;W) + \frac{1}{2}\mathcal{I}(Q_2;W) = \frac{1}{2}C + \frac{1}{2}C = C.$$

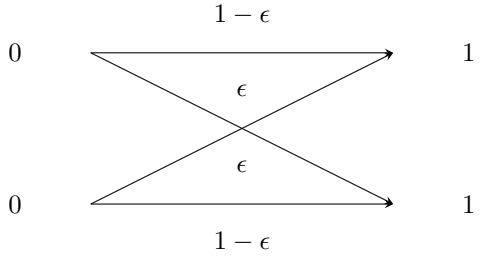
But $\frac{1}{2}Q_1 + \frac{1}{2}Q_2 = (\frac{1}{2}, \frac{1}{2})$ □

3.2 Basic Channel Capacity Examples

Here we will derive useful formulas for the capacity of a few simple channels.

3.2.1 Binary Symmetric Channel

Consider a binary symmetric channel, represented by The following diagram.



Its channel matrix law is thus:

$$\begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix} \tag{3.9}$$

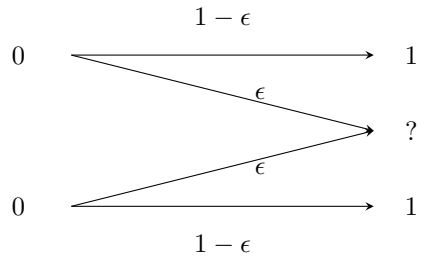
To calculate the capacity, we will maximize the mutual information $\mathcal{I}(X;Y)$:

$$\begin{aligned} \mathcal{I}(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x\in\mathcal{X}} Q(x)H(Y|X=x) \\ = H(Y) - \sum_{x\in\mathcal{X}} Q(x)H_b(\epsilon) = H(Y) - H_b(\epsilon) \leq \log_2|\mathcal{Y}| - H_b(\epsilon) = 1 - H_b(\epsilon) \end{aligned} \tag{3.10}$$

which is maximized for $Q = (\frac{1}{2}, \frac{1}{2})$.

3.2.2 Binary Erasure Channel

The binary erasure channel is a bit different, as it models the 'loss' of a bit during transmission, rather than a corruption. It looks like this:



A fraction ϵ of the bits are erased. The capacity is computed as follows:

$$\mathcal{I}(X;Y) = H(Y) - H(Y|X) = H(Y) - H_b(\epsilon) \tag{3.11}$$

$H(Y)$ is however not that simple to compute. Consider$P(X=1) = \pi$, then Y is distributed along the vector $QW = [(1-\pi)(1-\epsilon), \epsilon, \pi(1-\epsilon)]$. Thus

$$\begin{aligned} H(Y) = H(QW) \\ = -(1-\pi)(1-\epsilon)\log((1-\pi)(1-\epsilon)) - \epsilon\log(\epsilon) - \pi(1-\epsilon)\log(\pi(1-\epsilon)) \\ = -(1-\pi)(1-\epsilon)(\log(1-\pi) + \log(1-\epsilon)) - \epsilon\log(\epsilon) - \pi(1-\epsilon)(\log(\pi) + \log(1-\epsilon)) \\ = -(1-\epsilon)(\pi\log(\pi) + (1-\pi)\log(1-\pi)) - (\epsilon\log(\epsilon) + (1-\epsilon)\log(1-\epsilon)) \\ = (1-\epsilon)H_b(\pi) + H_b(\epsilon) \end{aligned} \tag{3.12}$$

Thus, plugging it into the mutual information,

$$I(X;Y) = H(Y) - H_b(\epsilon) = (1-\epsilon)H(\pi) \tag{3.13}$$

which is maximized at $C = 1-\epsilon$ by $\pi = \frac{1}{2}$.

3.3 Symmetric Channels

Definition 3.13. *A **symmetric channel** is a channel whose matrix law satsfies the following conditions:*

1. *the rows of the probability transition matrix are permutations of each other*

2. *the columns of the probability transition matrix are permutations of each other*

Example 2.

$$W = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix} \tag{3.14}$$

describes a symmetric channel.

Proposition 3.14 (Capacity Bound for Symmetric Channels)**.** *Letting r denote a row of the probability transition matrix, we know that*

$$\mathcal{I}(X;Y) = H(Y) - H(Y|X) \leq \log|\mathcal{Y}| - H(r) \tag{3.15}$$

with equality if X is uniformly distributed, since uniform distribution on $\mathcal{X}$ yields uniform distribution on $\mathcal{Y}$.

$$p(y) = \sum_{x \in \mathcal{X}} W(y|x)Q(x) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} W(y|x) = \frac{1}{|\mathcal{Y}|} \tag{3.16}$$

Thus a uniform distribution on the input maximizes the capacity of a symmetric channel.

3.3.1 Weakly Symmetric Channels

Definition 3.15. *A weakly symmetric channel is a relaxation of the symmetric channel, and must have a matrix law satisfying:*

1. *the rows of the probability transition matrix are permutations of each other*

2. *all columns sum to the same amount*

Example 3.

$$W = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{3} \end{bmatrix} \tag{3.17}$$

describes a weakly symmetric channel.

In other words, a weakly symmetric channel is a convex combination of (strongly) symmetric channels (see Fig 1)!

Theorem 3.16 (Capacity of Weakly Symmetric Channels)**.** *Let r be a row of a weakly symmetric channel's matrix law, then its capacity is given by*

$$C = \log|\mathcal{Y}| - H(r) \tag{3.18}$$

and it is achieved by a uniform distribution on the input alphabet.

3.4 Karush-Kuhn-Tucker Conditions

The Karush-Kuhn-Tucker conditions establish a method to maximize a concave function over a probability vector.

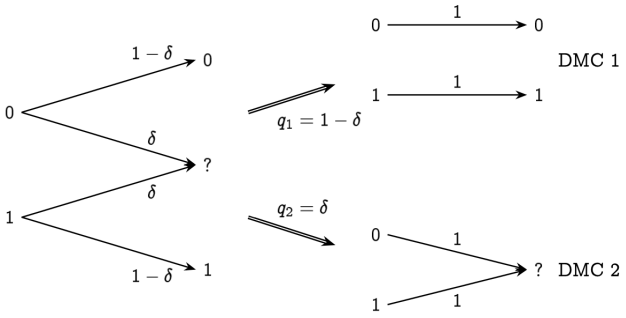


Figure 1: The BEQ as a convex combination of BSCs

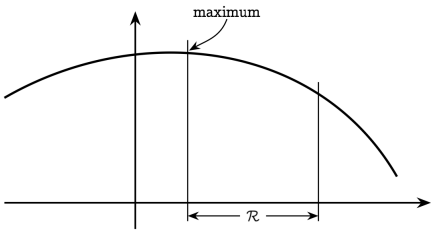


Figure 2: A typical concave function

3.4.1 General Conditions

Let f be a concave function over $\mathcal{R}$, a convex region of $\mathbb{R}^n$. We wish to find $q \in \mathcal{R}$ such that $f(q)$ is maximal.

Using calculus arguments, we know that $q^* \in \mathbb{R}^n$ maximizes f over $\mathbb{R}^n$ if:

$$\frac{\partial f(q^*)}{\partial q_l^*} = 0, \quad \forall l \in \{1, 2, .., n\} \tag{3.19}$$

However, q^* is not guaranteed to lie in $\mathcal{R}$. To find the maximum $q^* \in \mathcal{R}$, we will use the following lemma:

Lemma 3.17. *If there exists a $q^* \in \mathcal{R}$ such that*

$$\frac{\partial f(q^*)}{\partial q_l^*} = 0 \quad \forall l$$

then q^* maximizes f.

Proof. By contradiction.. TODO □

3.4.2 For Probability Vectors

The fact that q must be a probability vector can be translated to the following condition:

$$\sum_{l=1}^{L} q_l = 1 \iff \sum_{l=1}^{L} q_l - 1 = 0 \implies \lambda(\sum_{l=1}^{L} q_l - 1) = 0 \quad \forall \lambda$$

Thus, any q that maximizes f also maximizes

$$F(q) = f(q) - \lambda(\sum_{l=1}^{L} q_l - 1) \tag{3.20}$$

where λ is called the **Lagrange multiplier**.

Theorem 3.18 (KKT Condition Theorem)**.**

$$\begin{aligned} \frac{\partial f(q)}{\partial q_l} &= \lambda \quad \forall l : q_l \geq 0 \\ \frac{\partial f(q)}{\partial q_l} &\leq \lambda \quad \forall l : q_l = 0 \end{aligned} \tag{3.21}$$

*are **necessary** and **sufficient conditions** on q to maximize f on $\mathcal{R}$. λ is chosen so that the condition on q is satisfied.*

Proof. Suppose we have β such that $f(\beta) \leq f(\alpha)$ where α is the maximiser of $f(\cdot)$. From concavity, we know that

$$f(\beta) - f(\alpha) \leq \frac{f(\theta\beta + \bar{\theta}\alpha) - f(\alpha)}{\theta} \to \sum_{l=1}^{L} \frac{\partial f(\alpha)}{\partial \alpha_l}(\beta_l - \alpha_l) \leq \sum_{l=1}^{L} \lambda(\beta_l - \alpha_l) = 0 \tag{3.22}$$

as $\theta \to 0$ since $\sum_{l=1} \beta_l = 1 = \sum_{l=1} \alpha_l$. Thus, for any β, $f(\beta) \leq f(\alpha)$.

For necessity, again suppose that α maximises $f(\cdot)$. W.l.o.g. suppose that $\alpha_1 > 0$. We can define β as

$$\beta = \alpha + \epsilon e_k - \epsilon e_1 \tag{3.23}$$

for some fixed $k \in \{1, 2, ..., L\}$. We choose

$$\lambda := \frac{\partial f(\alpha)}{\partial \alpha_1} \tag{3.24}$$

and by the above part of the proof, we know that

$$\begin{aligned} \frac{f(\theta\beta + \bar{\theta}\alpha) - f(\alpha)}{\theta} \leq 0 \implies \sum_{l=1}^{L} \frac{\partial f(\alpha)}{\partial \alpha_l}(\beta_l - \alpha_l) \leq 0 \\ = \epsilon \frac{\partial f(\alpha}{\partial \alpha_k} - \epsilon \frac{\partial f(\alpha)}{\partial \alpha_1} = \epsilon \frac{\partial f(\alpha}{\partial \alpha_k} - \epsilon\lambda \leq 0 \end{aligned} \tag{3.25}$$

yielding our first condition. Now if $\alpha_k > 0$, there is an ϵ which can be negative and still satisfy our conditions, so

$$-\alpha_k < \epsilon < 0 \tag{3.26}$$

plugged into our above inequality will yield

$$\frac{\partial f(\alpha)}{\partial \alpha_k} \geq \lambda \tag{3.27}$$

Both conditions are only satisfied at the same time if

$$\frac{\partial f(\alpha)}{\partial \alpha_k} = \lambda \tag{3.28}$$

□

3.4.3 KKT for Channel Capacity

A useful reminder would be that for $Q \in \mathcal{P}(\mathcal{X})$ and $W(y|x)$ a channel,

- $Q \mapsto \mathcal{I}(Q, W)$ is **concave**
- $W \mapsto \mathcal{I}(Q, W)$ is **convex**

Also, we will denote $\max\{\mathcal{I}(Q, W)\}$ by $C^{(I)}$. By convention in Probability Theory, we will consider Q to be a row vector. This will help later on. For Q and W as given above, if Q and $\lambda \in \mathbb{R}$ are such that

$$D(W(\cdot, x)||(QW)(\cdot)) \leq \lambda, \quad \forall x \in \mathcal{X} \tag{3.29}$$

$$D(W(\cdot, x)||(QW)(\cdot)) = \lambda, \quad \forall x : Q(x) > 0 \tag{3.30}$$

then Q achieves capacity $C^{(I)} = \lambda$.

Conversely, if Q^* achieves $\max\{\mathcal{I}(Q, W)\}$, then

$$D(W(\cdot, x)||(Q^*W)(\cdot)) \leq \lambda, \quad \forall x \in \mathcal{X} \tag{3.31}$$

$$D(W(\cdot, x)||(Q^*W)(\cdot)) = \lambda, \quad \forall x : Q^*(x) > 0 \tag{3.32}$$

Proof. We know that $Q \mapsto \mathcal{I}(Q, W)$ is concave, so we apply Theorem 3.18, with q corresponding to Q and q_l corresponding to $Q_k = Q(x_k)$. To compute the partial derivatives, we express the mutual information between the input and the output DMC as

$$\begin{aligned}
\mathcal{I}(Q, W) &= \sum_x \sum_y Q(x)W(y|x) \ln\left(\frac{Q(x)W(y|x)}{Q(x)(QW)(y)}\right) \\
&= \sum_x \sum_y Q(x)W(y|x) \ln\left(\frac{W(y|x)}{\sum_{x'} Q(x')W(y|x')}\right)
\end{aligned}$$

By the product rule and by the chain rule, we have

$$\begin{aligned}
\frac{\partial \mathcal{I}(Q,W)}{\partial Q_k} &= \sum_x \sum_y I\{x = x_k\} W(y|x) \ln\left(\frac{W(y|x)}{\sum_{x'} Q(x')W(y|x')}\right) \\
&+ \sum_x \sum_y Q(x)W(y|x) \frac{\sum_{x'} Q(x')W(y|x')}{W(y|x)} \frac{-W(y|x)W(y|x_k)}{(\sum_{x'} Q(x')W(y|x'))^2} \\
&= \sum_y W(y|x_k) \ln\left(\frac{W(y|x)}{\sum_{x'} Q(x')W(y|x')}\right) - \sum_y \frac{W(y|x_k}{\sum_{x'} Q(x')W(y|x')} \sum_x Q(x)W(y|x) \\
&= \sum_y W(y|x_k) \ln\left(\frac{W(y|x)}{\sum_{x'} Q(x')W(y|x')}\right) - \sum_y W(y|x_k) \\
&= \sum_y W(y|x_k) \ln\left(\frac{W(y|x)}{\sum_{x'} Q(x')W(y|x')}\right) - 1 \\
&= D(W(\cdot|x_k)||(QW)(\cdot)) - 1
\end{aligned}$$

This also allows to check that the partial derivatives fulfill the conditions of theorem 3.18. Since (1) and (2) are satisfied we can invoke Theorem 3.18 with $\lambda' = \lambda - 1$ to conclude that Q maximizes $\mathcal{I}(\cdot|W)$ over all input distributions, i.e, that Q achieves capacity. Then,

$$\begin{aligned}
C = \mathcal{I}(Q,W) &= \sum_x \sum_y Q(x)W(y|x) \ln\left(\frac{W(y|x)}{\sum_{x'} Q(x')W(y|x')}\right) \\
&= \sum_x Q(x) D(W(\cdot|x_k)||(QW)(\cdot)) = \lambda \quad (t)
\end{aligned}$$

We finish by proving the Part b). Because Q maximizes $\mathcal{I}(\cdot, W)$ over all input distributions, we know by theorem 2 that there exists a λ' such that

$$D(W(\cdot, x)||(Q^*W)(\cdot)) \leq \lambda' + 1, \quad \forall x \in \mathcal{X} \tag{3.33}$$

$$D(W(\cdot, x)||(Q^*W)(\cdot)) = \lambda' + 1, \quad \forall x : Q^*(x) > 0 \tag{3.34}$$

From the same computation as in (t) we obtain $\mathcal{I}(Q,W) = \lambda' + 1$. Because we know that $\mathcal{I}(Q,W) = C$ it follows that $C = \lambda' + 1$ □

Example 4 (BEC Channel Capacity)**.** *We ask ourselves whether* $(\frac{1}{2}, \frac{1}{2})$ *is an optimal input distribution to the BEC channel. We start off by writing what the output distributions are given* Q*.*

- $(QW)(0) = \frac{1}{2}(1-\rho)$
- $(QW)(?) = \rho$
- $(QW)(1) = \frac{1}{2}(1-\rho)$

We now compute the relative entropy:

$$\begin{aligned}
D(W(\cdot,0)||(QW)(\cdot)) &= (1-\rho)\log(\frac{1-\rho}{\frac{1-\rho}{2}}) + \rho\log(\frac{\rho}{\rho}) + 0\log(\frac{0}{\frac{1-\rho}{2}}) \\
&= (1-\rho)\log(2) = 1-\rho
\end{aligned}$$

Analogously, $D(W(\cdot,1)||(QW)(\cdot)) = 1-\rho$*. Thus, the capacity of the binary erasure channel is* $1-\rho$*.*

3.5 Data Processing

Proposition 3.19. *$D(\cdot||\cdot)$ is convex.*

Proof. Consider $D(\lambda P_1 + \bar{\lambda}P_2||\lambda Q_1 + \bar{\lambda}Q_2)$. First, notice that both $\lambda P_1 + \bar{\lambda}P_2$ and $\lambda Q_1 + \bar{\lambda}Q_2$ are indeed pmfs.

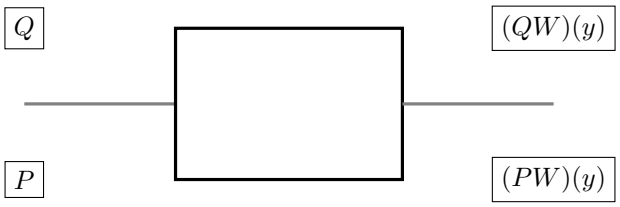
$$\begin{aligned} D(\lambda P_1 + \bar{\lambda}P_2||\lambda Q_1 + \bar{\lambda}Q_2) &= \sum_x (\lambda P_1 + \bar{\lambda}P_2) \log(\frac{\lambda P_1 + \bar{\lambda}P_2}{\lambda Q_1 + \bar{\lambda}Q_2}) \\ &\leq \sum_x \lambda P_1 \log(\frac{\lambda P_1}{\lambda Q_1}) + \bar{\lambda}P_2 \log(\frac{\bar{\lambda}P_2}{\bar{\lambda}Q_2}) \\ &= D(\lambda P_1||\lambda Q_1) + D(\bar{\lambda}P_2||\bar{\lambda}Q_2) \end{aligned}$$

□

3.5.1 Data Processing Inequality for $D(P||Q)$

Consider two input distributions Q, P being 'processed' by a channel W. Then

$$D(PW||QW) \leq D(P||Q) \tag{3.35}$$



Proof.

$$\begin{aligned} D(PW||QW) = \sum_y PW(y) \log(\frac{PW(y)}{QW(y)}) = \sum_y \sum_x P(x)W(y|x) \log(\frac{\sum_x P(x)W(y|x)}{\sum_x Q(x)W(y|x)}) \\ =_{i)} \sum_x P(x) \log(\frac{\sum_x P(x)}{\sum_x Q(x)}) \leq_{ii)} \sum_x P(x) \log(\frac{P(x)}{Q(x)}) = D(P||Q) \end{aligned}$$

Where Step i) cancels out the sum over y and the $W(y|x)$ since they sum to 1, and ii) uses the log-sum inequality. □

Data processing can be expressed as a **Markov chain**. Let X, Y, Z for a Markov chain, which we will denote as

$$X \multimap Y \multimap Z \tag{3.36}$$

3.5.2 Data Processing Inequality for $\mathcal{I}(X;Y)$

Proposition 3.20.

$$X \multimap Y \multimap Z \implies \mathcal{I}(X;Z) \leq \mathcal{I}(X;Y) \tag{3.37}$$

Proof. We first decompose $\mathcal{I}(X;(Y,Z))$ into two different ways

$$\begin{aligned} \mathcal{I}(X;(Y,Z)) &= \mathcal{I}(X;Y) + \mathcal{I}(X;Z|Y) = \mathcal{I}(X;Z) + \mathcal{I}(X;Y|Z) \\ &\implies {}_{i)}\mathcal{I}(X;Y) = \mathcal{I}(X;Z) + \mathcal{I}(X;Y|Z) \\ &\implies {}_{ii)}\mathcal{I}(X;Y) \geq \mathcal{I}(X;Z) \end{aligned} \tag{3.38}$$

where $i)$ denotes $\mathcal{I}(X;Z|Y)$ by Markov property and $ii)$ is non-negativity of $\mathcal{I}(X;Y|Z)$ □

3.6 Jointly Typical Sequences

Definition 3.21 (Jointly Typical Sequences)**.** *$\mathcal{A}_\epsilon^{(n)}$ denotes the set of jointly typical sequences $\{(x^n, y^n)\}$ of length n with respect to a joint distribution p_{xy}. This is formally defined as:*

$$\begin{aligned} \mathcal{A}_\epsilon^{(n)} = \{(\xi,\eta) \in \mathcal{X}^n \times \mathcal{Y}^n : & 2^{-n(H(P_{XY})+\epsilon)} < \Pi_{i=1}^n P_{xy}(\xi_i,\eta_i) < 2^{-n(H(P_{XY})-\epsilon)}, \\ & 2^{-n(H(P_X)+\epsilon)} < \Pi_{i=1}^n P_{xy}(\xi_i) < 2^{-n(H(P_X)-\epsilon)}, \\ & 2^{-n(H(P_Y)+\epsilon)} < \Pi_{i=1}^n P_{xy}(\eta_i) < 2^{-n(H(P_Y)-\epsilon)}\} \end{aligned} \tag{3.39}$$

Lemma 3.22 (Properties of Jointly AEP)**.** *Assumr $(X^n, Y^n) \sim_{i.i.d.} P_{XY}$*

- $|\mathcal{A}_\epsilon^{(n)}| < 2^{-n(H(X,Y)+\epsilon)}$
- $P((X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}) \to_{n\to\infty} 1$
- $|\mathcal{A}_\epsilon^{(n)}| > 2^{n(H(X,Y)-\epsilon}$ *for n large enough*

Now consider $(x_1,y_1),...,(x_n,y_n) \sim_{i.i.d.} P_X \times P_Y$, i.e. $X^n \perp\!\!\!\perp Y^n$, then

$$P((X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}) < 2^{-n(\mathcal{I}(X;Y)-3\epsilon)} \tag{3.40}$$

Proof.

$$\begin{aligned} P((X^n,Y^n)) &= \sum_{(\xi,\eta\in\mathcal{A}_\epsilon^{(n)})} P[X^n = \xi^n, Y^n = \eta^n] = \sum_{(\xi,\eta\in\mathcal{A}_\epsilon^{(n)})} P_X(X^n = \xi^n) P_Y(Y^n = \eta_n) \\ &\leq \sum_{(\xi,\eta\in\mathcal{A}_\epsilon^{(n)})} 2^{-n(H(P_X)-\epsilon)} 2^{-n(H(P_Y)-\epsilon)} |\mathcal{A}_\epsilon^{(n)}| = 2^{-n(\mathcal{I}(X;Y)-3\epsilon)} \end{aligned}$$

□

Remark 13. *We know that there are about $2^{nH(P_X)}$ (weakly) typical sequences in $\mathcal{X}^n$, and about $2^{nH(P_Y)}$ (weakly) jointly typical sequences in $\mathcal{Y}^n$. However, the above lemma shows that there only about $2^{nH(P_{XY})}$ jointly typical sequences in $(\mathcal{X}\times\mathcal{Y})^n$. Thus not all pairs of typical sequences in $\mathcal{X}^n$ and $\mathcal{Y}^n$ are jointly typical.*

4 Channel Coding Theorem

This will cover the entire process of proving the Channel Coding Theorem, laying down all prerequisites and establishing a proof.

Theorem 4.1. *(Channel Coding Theorem) All rates below C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda_n \to 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda_n \to 0$ must satisfy $R \leq C$.*

4.1 Proving the Converse

The setup: Given

1. a **message set** $\mathcal{M}$
2. an **encoder** $f^n : \{1, 2, ..., \mathcal{M}\} \to \mathcal{X}^n$ which yields codewords of length n
3. a **decoder** $\phi^n : \mathcal{Y}^n \to \{1, 2..., \mathcal{M}\}$
4. $\lambda_m = \sum_{y:\phi(y)\neq m} \Pi_{i=1}^n W(y_i|x_i(m)) = \sum_{y:\phi(y)\neq m} W(Y^n|X^n)$
5. $P_e^{(n)} = \frac{1}{|M|} \sum_{i=1}^n \lambda_i$

We will prove that if $P_e^{(n)}$ tends to zero then $R \leq C^{(I)} = \max_Q\{I(Q, W)\}$ *i.e. R is necessarily upper bounded.* Generate message $M \sim Unif(M)$, then $X(M) = (x_1, x_2, ..., x_n)$ and $Y^n = (Y_1, Y_2, ..., Y_n) \to \hat{M}$. Fano's Ineq states that

$$H(M|Y^n) \leq H_b(P_e^n) + P_e^n \log(|M|-1)$$

We should prove that the error $P_e^{(n)} = P(M \neq \hat{M})$

$$P(M \neq \hat{M}) = \sum_m P(M = m)P(\hat{M} \neq m|M = m) = \frac{1}{M} \sum_m \lambda_m = P_e^n$$

Fano's Ineq more loose can be stated also as $H(M|Y^n) \leq 1 + P_e^n \cdot (nR)$

Lemma 4.2. *Let $X^n \sim_{i.i.d.} P_X$ arbitrary, and Y_i is the result of feeding X_i to channel W, then*

$$\mathcal{I}(X^n; Y^n) \leq nC^{(I)} \tag{4.1}$$

Proof.

$$\begin{aligned} \mathcal{I}(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) = \sum_i H(Y_i|Y^{i-1}) - \sum_i H(Y_i|Y^{i-1}, X^n) \\ \leq \sum_i H(Y_i) - \sum_i H(Y_i|X_i) = \sum_i \mathcal{I}(X_i; Y_i) \leq \sum_i C^{(I)} = nC^{(I)} \end{aligned} \tag{4.2}$$

□

We pick M at random thus we have $H(M) = \log|M| = nR$. So, we can write

$$nR = H(M) = \mathcal{I}(M, Y^n) + H(M|Y^n) \leq \mathcal{I}(M, Y^n) + 1 + P_e^n nR$$

$$R \leq \frac{1}{n}\mathcal{I}(M, Y^n) + 1 + P_e^n R$$

Taking into consideration that n tends to infinity, that M is sampled iid, and that the error tends to 0 and lemma 4.1 we get that

$$R \leq C^{(n)} + \frac{1}{n} + P_e^n R \implies R \leq C^{(n)}$$

- use Fano's inequality, DPI & the previous lemma
- construct a weird encoder that samples a codeword with Q
- then show that $nR = H(M)$, use Fano, DPI to show it's smaller than $nC^{(I)}$

4.2 Proving the Direct

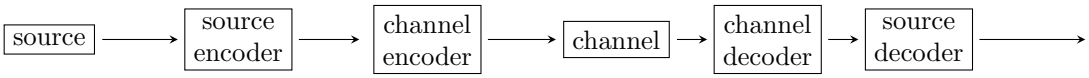
- Construct a Weak Typicality Decoder & drive average error prob. to zero via AEP
- show that $\lambda_{max} \to 0$ by selecting half of the codewords, which would reduce the rate,

4.3 Source Channel Separation

Now that the Channel Capacity Theorem has been proven, we can safely use the fact that

$$C = C^{(I)} \tag{4.3}$$

In the concept of source channel separation, we ask ourselves whether encoding the source and transmitting should **completely** separated tasks or not, i.e. if knowledge of the source encoding aids in efficient transmission of information of not. Our communication system will thus look like:



4.3.1 Fano's Inequality for Sequences

Proposition 4.3. *Let $U^k \in \mathcal{U}^k$ be a sequence with Y^n a sequence of observations. $\hat{U}(Y^n)$ is a k-length function of Y^n. Let $P_{e,i}$ denote $P[U_i \neq \hat{U}_i]$, and $P_{avg} = \frac{1}{k}\sum_{l=1}^{k} P_{e,l}$ be the average error probability, then*

$$H(U^k|\hat{U}) \leq_{Fano} H_b(P_{avg}) + P_{avg}\log(|\mathcal{U}|-1) \tag{4.4}$$

Proof.

$$\begin{aligned} \frac{1}{k}H(U^k|\hat{U}) &= \frac{1}{k}\sum_i H(U_i|U^{i-1}, \hat{U}) \leq \frac{1}{k}\sum_i (H_b(P_{e,i}) + P_{e,i}\log(|\mathcal{U}|-1)) \\ &= \frac{1}{k}\sum_i H_b(P_{e,i}) + \sum_i P_{e,i}\log(|\mathcal{U}|-1) = \frac{1}{k}\sum_i H_b(P_{e,i}) + P_{avg}\log(|\mathcal{U}|-1) \\ &\leq_{i)} H_b(P_{avg}) + P_{avg}\log(|\mathcal{U}|) \end{aligned} \tag{4.5}$$

where i) exploits concavity of the binary entropy function □

Let C be in units $[\frac{bits}{channel\ use}]$, $H(U)$ be in units $[\frac{bits}{source\ symbols}]$, then $\rho = \frac{k}{n}[\frac{source\ symbols}{channel\ use}]$.

We will show that if $H(U) < C$, then **the source can be communicated reliably with the separation approach.**

Consider that the source is compressed optimally using AEP, so the source encoder simply sends the 'address' of the source, which is an element in $\mathcal{A}_\epsilon^{(n)}(P_U)$. On top of that, we consider the channel to satisfy $\lambda_{max} < \epsilon$.

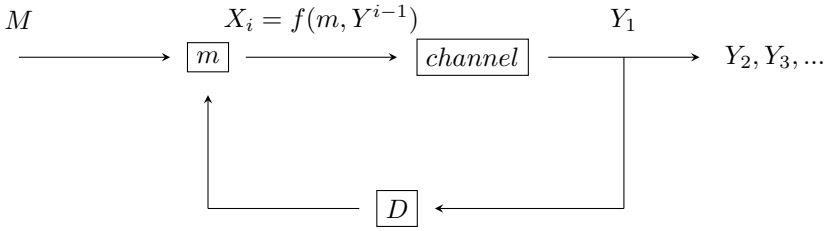
Logically, an error occurs **if the source sequence is not typical** or **if the channel encoder messed up**. Both of these events occus with a probability smaller than ϵ, thus by the union bound of event, the probability of both occurring is upper bounded by 2ϵ.
Even with combined source channel coding you cannot have $P(U \neq \hat{U}) \to 0$ if $H(U)\rho > C$

Proof.

$$\begin{aligned}\frac{1}{k}H(U^k) &= \frac{1}{k}(H(U^k) - H(U^k|\hat{U}^k) + H(U^k|\hat{U}^k)) \\ &\leq \frac{1}{k}\mathcal{I}(U^k;\hat{U}^k) + H_b(P_{avg}) + P_{avg}log|U| \\ &\leq \frac{1}{k}\mathcal{I}(X^n;Y^n) + H_b(P_{avg}) + P_{avg}log|U| \\ &\leq \frac{1}{k}nC \Rightarrow H(u) \leq \frac{1}{\rho}C\end{aligned}$$

□

4.3.2 Feebdack Communication



It is a fair argument to believe that feedback communication would allow us to achieve higher rates, but we will show why it is not true.

Theorem 4.4. *(Feedback capacity) The feedback capacity C_{FB} is equal to the channel capacity with no feedback C.*

Proof. Since a non-feedback code is a particular version of a feedback-code (just with no feedback), then it is safe to assume that

$$C_{FB} \geq C \tag{4.6}$$

The other way around is a bit more tricky. We need to assume that the messages are uniformly distributed, so that we have

$$H(M) = nR \tag{4.7}$$

which we can extend to

$$nR = H(M) - H(M|Y^n) + H(M|Y^n) = H(M|Y^n) + \mathcal{I}(M;Y^n) \tag{4.8}$$

which by Fano's inequality for sequences

$$nR \leq 1 + nRP_e^{(n)} + \mathcal{I}(M;Y^n) = 1 + nRP_e^{(n)} + H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{4.9}$$

Using the conditional entropy bound, we get that

$$\mathcal{I}(M;Y^n) \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i=1}^{n} \mathcal{I}(X_i;Y_i) \leq nC \tag{4.10}$$

Thus our final inequality is

$$nR \ \leq 1 + nRP_e^{(n)} + nC \implies_{n\to\infty} \quad nR \leq nC \tag{4.11}$$

Thus proving that

$$C_{FB} = C \tag{4.12}$$

□

Example 5 (Typical Decoder of a BSC)**.** *Consider a BSC with crossover probability α, and a fixed codebook. We also have $P_X(0) = P_X(1) = \frac{1}{2}$.*

Remember that a typical decoder will search for pairs $(x^n(m), y^n)$ such that they are in $\mathcal{A}_\epsilon^{(n)(P_{XY})}$. If exactly one $\hat{m}$ fits the criterion, then the decoder outputs this $\hat{m}$.

We will have a look at $\mathcal{A}_\epsilon^{(n)}(P_{XY})$. Note that

$$P_{XY}(x,y) = P_X(x)W_{Y|X}(y|x) \tag{4.13}$$

i.e.

$$P_{XY}(0,0) = \frac{1}{2}(1-\alpha) = P_{XY}(1,1), \quad P_{XY}(0,1) = \frac{\alpha}{2} = P_{XY}(1,0) \tag{4.14}$$

This also yields that $P_Y(0) = P_Y(1) = \frac{1}{2} \implies H(P_X) = H(Y) = 1$ bit. Thus, H(X,Y) = H(Y) + H(Y—X) = 1 + $H_b(\alpha)$. BY the AEP, we get

$$|-\frac{1}{n}\log\prod_{i=1}^{n} p(x_i,y_i) - H(P_{XY})| \leq \epsilon \tag{4.15}$$

For a fixed (x^n, y^n), we denote the Hamming distance between x^n and y^n to be d, thus

$$\prod_{i=1}^{n} P_{XY}(x_i,y_i) = \prod_{i=1}^{n} P_X(x)W_{Y|X}(y|x) = 2^{-n}\prod_{i=1}^{n} W_{Y|X}(y|x) = 2^{-n}\alpha^d(1-\alpha)^{n-d} - 1 - H_b(\alpha) \tag{4.16}$$

Plugging this into our conditions, we get

$$\begin{aligned} |-\frac{1}{n}\log\prod_{i=1}^{n} p(x_i,y_i) - H(P_{XY})| &= |-\frac{1}{n}\log(2^{-n}\alpha^d(1-\alpha)^{n-d})| \\ &= |1 + \frac{d}{n}\log(\frac{1}{\alpha}) + \frac{n-d}{n}\log(\frac{1}{1-\alpha}) - 1 - H_b(\alpha)| \\ &= |(\frac{d}{n} - \alpha)\log(\frac{1}{\alpha}) - (\frac{d}{n} - \alpha)\log(\frac{1}{1-\alpha})| \\ &= |(\frac{d}{n} - \alpha)\log(\frac{1-\alpha}{\alpha})| \leq \epsilon \end{aligned} \tag{4.17}$$

yielding

$$\alpha - \frac{\epsilon}{|\log(\frac{1-\alpha}{\alpha})|} \leq \frac{d}{n} \leq \alpha + \frac{\epsilon}{|\log(\frac{1-\alpha}{\alpha})|} \tag{4.18}$$

which makes sense, as the expected crossovers for n channel uses is α.

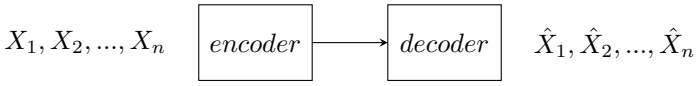
5 Rate Distortion Theory

Also known as **vector quantization** or **lossy data compression**, this theory aims to help define the 'goodness' of a representation of a source. We will see that this is achieved by defining a **distortion measure**, which will quanitify the 'distance' between the random variable and its representation.

Example 6. *Representing an **arbitrary** real number will require an infinite amount of bits. How can we minimize the mean squared error of the representation?*

5.1 Formal Problem Definition

Formally, we presume to have a source that produces a sequence $\mathcal{X}^n \ni X_1, X_2, ..., X_n \sim_{i.i.d.} P_X$. Our encoding scheme will look like



where $\hat{\mathcal{X}}$ is the **representation alphabet**.

Definition 5.1. *A **distortion function** is a mapping*

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+ \tag{5.1}$$

Definition 5.2. *A **distortion function between two sequences** $x^n, \hat{x}^n$ is defined by*

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_i d(x_i, \hat{x}_i) \tag{5.2}$$

i.e. it is the average per-symbol distortion.

Definition 5.3. *Consider an encoder $f : \mathcal{X} \to \{1, 2, ..., 2^{nR}\}$ and a decoder $\phi : \{1, 2, ..., 2^{nR}\} \to \mathcal{X}$. Then the value*

$$d(X^n, \phi(f(x^n))) \tag{5.3}$$

*is a random variable. This forms a **$(2^{nR}, n)$ rate distortion code**.*

Most of the time, we will need to consider the value

$$E[d(X^n, \phi(f(x^n)))] \tag{5.4}$$

If this value is bounded by $D < \infty$, then D is called the **maximal allowed distortion**.

Definition 5.4. *(R,D) is **achievable** if for any $\delta > 0$ and large enough n, $\exists f, \phi$ such that*

$$E[d(X^n, \phi(f(x^n)))] \leq D + \delta \tag{5.5}$$

Definition 5.5. *The **rate distortion function** $R(D)$ is defined as*

$$R(D) = \inf\{R\} \tag{5.6}$$

such that (R, D) is achievable. I.e. it is the smallest rate allowing reconstruction with distortion D.

Definition 5.6. *The **distortion rate function** $D(R)$ is defined as*

$$D(R) = \inf\{D\} \tag{5.7}$$

such that (R, D) is still achievable.

Definition 5.7. *The **information rate distortion function** is defined as*

$$R^{(I)}(D) = \min_{p(\hat{x}|x:\sum_{(}\hat{x},x)p(x)p(\hat{x}|x)d(x,\hat{x})\leq D)} \{\mathcal{I}(X, \hat{X})\} \tag{5.8}$$

This minimization is over all conditionals $p(\hat{x}|x)$ such that an expected distortion rate over the joint $p(x, \hat{x}) = p(x)p(\hat{x}|x)$ is lower than D.

5.2 Rate Distortion Theorem

This section will cover the main result of rate distortion theory, as well as its proofs and accompanying examples.

Theorem 5.8 (**Rate Distortion Theorem**)**.** *The rate distortion for an i.i.d source X with prior distribution $p(X)$ and bounded distortion function $d(x, \hat{x})$ is equal to its information rate distortion function. Thus,*

$$R(D) = R^{(I)}(D) \tag{5.9}$$

is the minimal achievable rate at distortion D.

The proof will be done in the following subsections. Analogously to the channel coding theorem, we will prove it in two parts (converse + direct).

5.2.1 Rate Distortion of a Binary Source

In a memoryless binary source setup, we have $X \sim Ber(p)$ for $p \in [0, 1]$ and we will consider the *Hamming distortion* d_H which counts the number of discrepancies between X and $\hat{X}$.

Taking our formula, we have

$$R^{(I)}(D) = \min_{p(\hat{x}|x:\sum_{(}\hat{x},x)p(x)p(\hat{x}|x)d(x,\hat{x})\leq D)} \{\mathcal{I}(X, \hat{X})\} \tag{5.10}$$

which we will minimize by

1. Defining $\mathcal{I}(X, \hat{X})$
2. Finding a lower bound to $\mathcal{I}(X, \hat{X})$
3. Proving the lower bound is achievable (by construction)

Here we go. Notice that $X \oplus \hat{X} = 1 \iff X \neq \hat{X}$

$$\begin{aligned} \mathcal{I}(X, \hat{X}) = H(X) - H(X|\hat{X}) &=_i H(X) - H(X \oplus \hat{X}|\hat{X}) \\ &\geq_{ii} H(X) - H(X \oplus \hat{X}) \geq_{iii} H_b(p) - H_b(D) \end{aligned} \tag{5.11}$$

i $(X \oplus \hat{X})|\hat{X}$ is a deterministic one-to-one function of X, thus it has the same entropy as X

ii conditioning reduces entropy

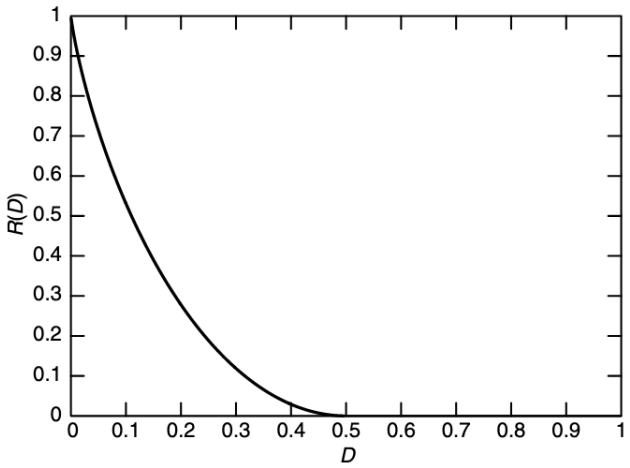


Figure 3: Rate of a Ber($\frac{1}{2}$) random source

iii $(\hat{X} \oplus \hat{X})$ is a binary random variable, and $\hat{X} \oplus \hat{X} = 1$ with probability $E[\hat{X} \oplus \hat{X}] \leq D$

To match the lower bound with equality, we will need to satisfy both *ii* and *iii*. For *ii*, it is enough to say that $(X \oplus \hat{X} \perp\!\!\!\perp \hat{X}) \iff X = \hat{X} \oplus Z, \quad Z \perp\!\!\!\perp \hat{X}$. To match *iii*, we can simply state that $E[X \oplus \hat{X}] = D$.

Since this implies $X \oplus \hat{X} = Z$, we get that $Z \sim Ber(D)$, thus $X = \hat{X} \oplus Ber(D) \sim Ber(p)$. This 'forces' $\hat{X} \sim Ber(q)$, and we get that

$$q = \frac{p - D}{1 - 2D} \tag{5.12}$$

Our rate distortion function thus looks like (Fig 3):

Remark 14. *The maximal rate distortion value for binary source is*

$$R(D)|_{D=0} = H(P) \tag{5.13}$$

5.2.2 Converse of the Rate Distortion Theorem

The converse of the rate distortion theorem can be expressed as: for any rate-R encoder-decoder pair (f, ϕ), we have

$$E[d(x, \phi(f(x)))] \leq D \implies R \geq R^{(I)}(D) \tag{5.14}$$

It can be restated as

$$R < R^{(I)}(D) \implies E[d(x, \phi(f(x)))] > D \tag{5.15}$$

or in words, **if we describe X at a rate less than $R^{(I)}(D)$, then we cannot achieve a distortion of less than D.**[2] The following proposition gives us the tools to prove this converse.

Proposition 5.9. *Here are some useful properties of $R^{(I)}(D)$ (resp $R(D)$):*

[2]In RD Theory, we wish to **minimize the rate**, since it is the 'amount' we require to describe the source

1. *monotonically decreasing*
2. *convexity*
3. *continuous*

Proof. For monotonicity, it is clear that

$$R^{(I)}(D+\delta) \leq R^{(I)}(D) \quad \forall \delta > 0 \tag{5.16}$$

since $R^{(I)}(D)$ is the infimum of the feasible set.

For convexity, it is a bit more complicated. We will want to prove that for two values D_0, D_1

$$R(\lambda D_0 + \bar{\lambda} D_1) \leq \lambda R(D_0) + \bar{\lambda} R(D_1), \quad \lambda \in [0,1] \tag{5.17}$$

We consider D_i such that

$$\sum_x \sum_{\hat{x}} P_X(x) P^i_{\hat{X}|X}(\hat{x}|x) d(x,\hat{x}) \leq D_i \tag{5.18}$$

Since $\lambda \in [0,1]$, then $\lambda P^0_{\hat{X}|X} + \bar{\lambda} P^1_{\hat{X}|X}$ is also a probability distribution (mixture), so the constraints are also satisfied, i.e. we have

$$\sum_x \sum_{\hat{x}} P_X(x)(\lambda P^0_{\hat{X}|X}(\hat{x}|x) + \bar{\lambda} P^1_{\hat{X}|X}(\hat{x}|x)) d(x,\hat{x}) \leq \lambda D_0 + \bar{\lambda} D_1 \tag{5.19}$$

Now, considering $\mathcal{I}(\hat{X};X)$ with respect to the joint $P_X(\lambda P^0_{\hat{X}|X} + \bar{\lambda} P^1_{\hat{X}|X})$, and that it is convex, then we have

$$\mathcal{I}(\hat{X};X)_{P_X(\lambda P^0_{\hat{X}|X} + \bar{\lambda} P^1_{\hat{X}|X})} \leq \lambda \mathcal{I}(\hat{X};X)_{P_X P^0_{\hat{X}|X}} + \bar{\lambda} \mathcal{I}(\hat{X};X)_{P_X P^1_{\hat{X}|X}} \tag{5.20}$$

Finally, since $R^{(I)}(D)$ is the minimum of $\mathcal{I}(\hat{X};X)$, then we can squeeze the mutual information in our desired inequality:

$$\begin{aligned} R(D) &\leq \mathcal{I}(X^n, \phi(f(X^n))) \leq_{DPI} \mathcal{X}^{\setminus}; \{(\mathcal{X}^{\setminus}) \\ &= H(f(X^n)) - H(f(X^n)|X^n) = H(f(X^n)) \leq nR \end{aligned} \tag{5.21}$$

□

Using this proof of convexity, we can now prove the converse. Let $D_i = E[d(x_i, \hat{x}_i)]$& we promise that $\frac{1}{n}\sum_i D_i \leq D$

$$\begin{aligned} nR &\geq H(X^n) - H(X^n|\hat{X}^n) = \sum_i H(X_i) - \sum_i H(X_i|X^{i-1}, \hat{X}^n) \\ &\geq \sum_i H(X_i) - \sum_i H(X_i|\hat{X}^i) = \sum_i \mathcal{I}(X_i, \hat{X}_i) \geq \sum_i R^{(I)}(D_i) \end{aligned} \tag{5.22}$$

which proves our converse bound.

5.2.3 Direct Part of Rate Distortion Theorem

This part will focus on proving that

$$\begin{aligned} \forall \delta > 0, \tilde{\epsilon} > 0 \quad \text{if } P_{\hat{X}|X} : E_{P_X P_{\hat{X}|X}}[d(X,\hat{X})] \leq D \text{ then} \\ \exists \text{ codes of rate } R = \mathcal{I}(X;\hat{X}) + \tilde{\epsilon} \text{ s.t. } E[d(X,\phi(f(X)))] \leq D + \delta \end{aligned} \tag{5.23}$$

this basically means that $E_{P_X P_{\hat{X}|X}}[d(X,\hat{X})] \leq D \implies R^{(I)}(D) \geq R$

Lemma 5.10 (Never Give Up)**.** *For n independent Bernoulli trials of probability p, the probability of success tends to 1 and the probability of failure tends to 0 as $np \to \infty$*[3]

Proof.

$$P(fail) = 1 - P(success) = (1-p)^n, \quad 1 - p \leq e^{-p} \implies (1-p)^n \leq e^{-pn} \to_{np\to\infty} 0 \tag{5.24}$$

□

Lemma 5.11 (Strong Typicality)**.** *Recall that the strongly typical set of a given probability distribution P is*

$$T_\epsilon^{(n)} = \left\{ x \in \mathcal{X}^n : |\frac{1}{n} N(a|x) - P(a)| < \epsilon P(a) \right\} \tag{5.25}$$

Let $g : \mathcal{X} \to \mathbb{R}_+$.

$$x \in T_\epsilon^{(n)} \implies \frac{1}{n} \sum_{a \in \mathcal{X}} g(a) \leq (1+\epsilon) E_P[g(X)] \tag{5.26}$$

Proof.

$$\frac{1}{n} \sum_{i=1}^{n} g(x_i) = \frac{1}{n} \sum_{a \in \mathcal{X}} N(a|x) g(a) \leq \frac{1}{n} \sum_{a \in \mathcal{X}} P(a)(1+\epsilon) g(a) \tag{5.27}$$

□

Lemma 5.12. *Fix $0 < \epsilon' < \epsilon$. Then for a sufficiently large n, if $x \in T_\epsilon^{(n)}$ & $Y_i \sim_{iid} P_Y$, then*

$$Pr[(X,Y) \in T_\epsilon^{(n)}(P_{XY})] \geq 2^{-n(\mathcal{I}(X;Y)) + 4\delta_{XY}} \tag{5.28}$$

where $4\delta_{XY} \to 0$ as $\epsilon \to 0$

Proof. no proof □

Now to our proof of the direct. We first **fix the conditional** $P_{\hat{X}|X}$ & consider an encoder/decoder pair as follows:

$$\begin{cases} f : \mathcal{X}^n \to \{1, 2, ..., 2^{nR}\} : & \text{outputs the first index of a cwd that is strongly typical with the seq.} \\ \phi : \{1, 2, ..., 2^{nR}\} \to \hat{\mathcal{X}}^n : j \mapsto \hat{x}(j) \in \hat{\mathcal{X}}^n \end{cases} \tag{5.29}$$

If no codeword is strongly typical with the source sequence x, we just sent $j = 1$ & it will be reconstructed with allowed distortion. We now want to find rate-R codes such that the success probability is diven to 1.

Since the conditional $P_{\hat{X}|X}$ and the marginal P_X are given, we can compute

[3] success is when at least one success occurs

1. the joint $P_{\hat{X},X} = P_X P_{\hat{X}|X}$
2. the marginal $P_{\hat{X}} = \sum_x P_X(x) P_{\hat{X}|X}(\cdot|x)$

We first create a codebook $\mathcal{C}$ where the rows are chosen according to $P_{\hat{X}}$. Next up we analyse the probability of a successful encoding with our two sources of randomness: **the codebook and the source sequence**.

$$P(success) = \sum_{\mathcal{C}} P(\mathcal{C}) P(success|\mathcal{C}) = \sum_{x \in \mathcal{X}^n} P_X(x) P(success|X = x) \tag{5.30}$$

By showing that $P(success|X = x) \to 1$, we will prove our claim, since if the average success probability tends to 1, then there *must* exist a codebook $\mathcal{C}^*$ such that $P(success|\mathcal{C}^*) \to 1$.

$$\begin{aligned} \sum_{x \in \mathcal{X}^n} P_X(x) P(success|X = x) = & \sum_{x \in T_\epsilon^{(n)}(P_X)} P_X(x) P(success|X = x) + \\ & \sum_{x \notin T_\epsilon^{(n)}(P_X)} P_X(x) P(success|X = x) \\ \geq & \sum_{x \in T_\epsilon^{(n)}(P_X)} P_X(x) P(success|X = x) \end{aligned} \tag{5.31}$$

By the last lemma, we have a probability of success with 2^{nR} attempts to match a codeword that is strongly typical with x. This match occurs with probability $2^{-n(\mathcal{I}(X;\hat{X})+4\delta)}$.

By the never give up lemma, we have that if

$$2^{nR} 2^{-n(\mathcal{I}(X;\hat{X})+4\delta)} \to \infty \tag{5.32}$$

then $P(success|X = x) \to 1$.

This allows us to conclude that

$$R > \mathcal{I}(X; \hat{X}) + 4\delta_{X\hat{X}} \tag{5.33}$$

6 Multi-Terminal Information Theory

This section will cover a brief introduction to distributed data compression, more specifically the **Slepian-Wolf Coding Scheme.**

6.1 Distributed Source Codes

Our setup is a joint sequence of random variables

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) \sim_{i.i.d.} P_{XY} \tag{6.1}$$

We will consider a scheme where

1. The (X^n, Y^n) sequence is 'correlated'
2. X^n and Y^n are separately encoded
3. (X^n, Y^n) are jointly decoded

Definition 6.1. *A $((2^{nR_1}, 2^{nR_2}), n)$ distributed source code for th ejoint sources (X, Y) consists of **two** encoders*

$$\begin{aligned} f_1 &: \mathcal{X}^n \to \{1, 2, ..., 2^n nR_1\} \\ f_2 &: \mathcal{Y}^n \to \{1, 2, ..., 2^n nR_2\} \end{aligned} \tag{6.2}$$

and a single decoder

$$g : \{1, 2, ..., 2^n nR_1\} \times \{1, 2, ..., 2^n nR_2\} \ \to \mathcal{X}^n \times \mathcal{Y}^n \tag{6.3}$$

Definition 6.2. *The probability of error of a distributed source code is defined as*

$$P_e^(n) = P(g(f_1(X^n), f_2(Y^n)) \neq (X^n, Y^n)) \tag{6.4}$$

i.e. a mismatch in decoding the encoding.

Definition 6.3. *A rate pair $(R-_1, R_2)$ is said **achievable** for a distributed source if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with probability error P_e such that it is driven to 0. This defines an **achievability region** in $\mathbb{R}^2$.*

6.2 Slepian Wolf Theorem

Intuitively, we would tend to believe that to decode a separately encoded pair (X^n, Y^n), we would require a minimum rate of

$$R_1 + R_2 \geq H(X) + H(Y) \tag{6.5}$$

This is however not the case, in a paper written by Slepian and Wolf.[4]

Theorem 6.4 (Slepian & Wolf Coding)**.** *For.a distributed source coding of the source $(X, Y) \sim_{i.i.d.} P_{XY}$, the achievable region is given by the following conditions:*

$$\begin{aligned} R_1 &\geq H(X|Y) \\ R_2 &\geq H(Y|X) \\ R_1 + R_2 &\geq H(X, Y) \end{aligned} \tag{6.6}$$

Again, the proof will be decomposed in two parts: achievability and converse.

[4]Fun fact, Lapidoth met Jack Wolf

6.2.1 Achievability of Slepian-Wolf Coding

Our encoding and decoding scheme follows differet semantics. We use a **binning** technique, in which the encoder will assign a sequence to one of 2^{nR_i} 'bins' at random. The index of that bin is then given to the decoder.

$$f_1 : \mathcal{X}^n \to \cup_1 \quad \cup_2 \quad \cup_3 \quad ... \quad \cup_{2^{nR_1}} \tag{6.7}$$

For a **single source**, the decoder will correctly decode if there exists a single typical sequence assigned to the index received. We can thus re-prove our bound of $R > H(X)$ with this scheme.

$$\begin{aligned} P(g(f(x)) \neq x) &\leq P(x \notin \mathcal{A}_\epsilon^{(n)}) + \sum_x \sum_{x' \neq x, x' \in \mathcal{A}_\epsilon^{(n)}} P(f(x') = f(x))p(x) \\ &\leq \epsilon + \sum_{x' \in \mathcal{A}_\epsilon^{(n)}} 2^{-nR} \leq \epsilon + 2^{-nR}|\mathcal{A}_\epsilon^{(n)}| \\ &= \epsilon + 2^{-nR}2^{n(H(X)+\epsilon)} \leq 2\epsilon \end{aligned} \tag{6.8}$$

as long as $R > H(X)$. We can now extend our argument to distributed sources.

Analysing the error probability, we distinguish 4 cases of error:

1. $E_0 = \{(X,Y) \notin \mathcal{A}_\epsilon^{(n)}(P_{XY})\}$
2. $E_1 = \{\exists x' \neq X : f_1(x') = f_1(X) \wedge (x', Y) \in \mathcal{A}_\epsilon^{(n)}(P_{XY})\}$
3. $E_2 = \{\exists y' \neq Y : f_2(y') = f_2(Y) \wedge (X, y') \in \mathcal{A}_\epsilon^{(n)}(P_{XY})\}$
4. $E_{12} = \{\exists (x', y') : x' \neq X, y' \neq Y, f_1(x') = f_1(X), f_2(y') = f_2(Y) \wedge (x', y') \in \mathcal{A}_\epsilon^{(n)}(P_{XY})\}$

Using the union bound, we can upper bound our probability error

$$P_e^{(n)} = P(E_0 \cup E_1 \cup E_2 \cup E_{12}) \leq P(E_0) + P(E_1) + P(E_2) + P(E_{12}) \tag{6.9}$$

and we will consider each case separately.

By the AEP, $P(E_0) \to 0$ as $n \to \infty$, so we can assume $P(E_0) < \epsilon$.

The following lemma must be introduced to analyse the next bounds

Lemma 6.5. *For any $\epsilon > 0$, $\mathcal{A}_\epsilon(X|y)$ denotes the set of sequences x that are jointly typical with a fixed sequence y. For a sufficiently large n, we have*

$$|\mathcal{A}_\epsilon(X|y)| \leq 2^{n(H(X|Y)+2\epsilon)} \tag{6.10}$$

With this lemma, we will no analyse the error probability for E_1, E_2, E_{12}.

$$\begin{aligned} P(E_1) \leq \sum_{(x,y)} p(x,y) \sum_{x' \neq x, (x'y) \in \mathcal{A}_\epsilon^{(n)}(P_{XY})} P(f_1(x') = f_1(x)) &= \sum_{(x,y)} P(x,y)2^{-nR_1}|\mathcal{A}_\epsilon(X|y)| \quad [5] \\ &\leq 2^{-nR_1}2^{n(H(X|Y)+\epsilon)} \end{aligned} \tag{6.11}$$

By symmetry, we have

$$P(E_1) \leq 2^{-nR_2}2^{n(H(Y|X)+\epsilon)} \tag{6.12}$$

both are driven to 0 if $R_1 > H(X|Y)$ and $R_2 > H(Y|X)$. Moreover, $P(E_{12}) \to 0$ if $R_1 + R_2 > H(X,Y)$. The average error probability is thus upper bounded by 4ϵ, which is driven to 0 as $n \to \infty$.

[5] by the above lemma

6.2.2 Converse for Slepian-Wolf Coding

Similarly to the single-source case, we will use Fano's inequality. Let $I_0 = f_1(X^n)$ and $J_0 = f_2(Y^n)$. Then we have

$$H(X^n, Y^n | I_0, J_0) \leq P_e^{(n)} n(\log|\mathcal{X}| + \log|\mathcal{Y}|) + 1 = n\epsilon_n \tag{6.13}$$

For the conditional probabilities, we have

$$H(X^n | Y^n, I_0, J_0) \leq P_e^{(n)} n\epsilon_n \tag{6.14}$$

and

$$H(Y^n | X^n, I_0, J_0) \leq P_\epsilon^{(n)} n\epsilon_n \tag{6.15}$$

So using the chain rule,

$$\begin{aligned} n(R_1 + R_2) \geq H(I_0, J_0) = I(X^n, Y^n; I_0, J_0) + H(I_0, J_0 | X^n, Y^n) \\ = I(X^n, Y^n; I_0, J_0) = H(X^n, Y^n) - H(X^n, Y^n | I_0, J_0) \geq nH(X, Y) - n\epsilon_n \end{aligned} \tag{6.16}$$