Information Theory and Coding - Prof. Emere Telatar

Jean-Baptiste Cordonnier, Sebastien Speierer, Thomas Batschelet

November 15, 2017

1 Data compression

Definition 1.1 (Information)**.** *Abstractly, **information** can be thought of as the resolution of uncertainty.*

Given an alphabet $\mathcal{U}$ (e.g. $\mathcal{U} = \{a, ..., z, A, ..., Z, ...\}$), we want to assign binary sequences to elements of $\mathcal{U}$, i.e.

$$\mathscr{C} : \mathcal{U} \to \{0,1\}^* = \{\emptyset, 0, 1, 00, 01, ...\}$$

For $\mathcal{X}$ a set

$$\mathcal{X}^n \equiv \{(x_0 ... x_n), x_i \in \mathcal{X}\}$$
$$\mathcal{X}^* \equiv \bigcup_{n \geq 0} \mathcal{X}^n$$

Definition 1.2. *A code $\mathscr{C}$ is called **singular** if*

$$\exists (u, v) \in \mathcal{U}^2, u \neq v \quad s.t. \quad C(u) = C(v)$$

Non singular code is defined as opposite

Definition 1.3. *A code $\mathscr{C}$ is called **uniquely decodable** if*

$$\forall u_1, ..., u_n, v_1, ..., v_n \in \mathcal{U}^* \quad s.t. \quad u_1, ..., u_n \neq v_1, ..., v_n$$

we have

$$\mathscr{C}(u_1)...\mathscr{C}(u_n) \neq \mathscr{C}(v_1)...\mathscr{C}(v_n)$$

i.e, $\mathscr{C}$ is non-singular

Definition 1.4. *Suppose $\mathscr{C} : \mathcal{U} \to \{0,1\}^*$ and $\mathcal{D} : \mathcal{V} \to \{0,1\}^*$ we can define*

$$\mathscr{C} \times \mathcal{D} : \mathcal{U} \times \mathcal{V} \to \{0,1\}^* \quad as \quad (\mathscr{C} \times \mathcal{D})(u, v) \to \mathscr{C}(u)\mathcal{D}(v)$$

Definition 1.5. *Given $\mathscr{C} : \mathcal{U} \to \{0,1\}^*$, define*

$$\mathscr{C}^* : \mathcal{U}^* \to \{0,1\}^* \quad as \quad \mathscr{C}^*(u_1 ... u_n) = \mathscr{C}(u_1)...\mathscr{C}(u_n)$$

Definition 1.6. *A code $\mathcal{U} \to \{0,1\}^*$ is **prefix-free** is for no $u \neq v$ $\mathscr{C}(u)$ is a prefix of $\mathscr{C}(v)$.*

Theorem 1.1. *If $\mathscr{C}$ is prefix-free then $\mathscr{C}$ is uniquely decodable.*

Definition 1.7. *$l(\mathscr{C}(u))$ is the length of the code word $\mathscr{C}(u)$ and $l(\mathscr{C})$ is the expected length of the code:*

$$l(\mathscr{C}) = \sum_u l(\mathscr{C}(u))p(u)$$

Definition 1.8 (Kraft sum)**.** *Given $\mathscr{C} : \mathcal{U} \to \{0,1\}^*$*

$$kraftsum(\mathscr{C}) = \sum_u 2^{-l(\mathscr{C}(u))}$$

Lemma 1.2. *if $\mathscr{C} : \mathcal{U} \to \{0,1\}^*$ and $\mathcal{D} : \mathcal{V} \to \{0,1\}^*$ then*

$$kraftsum(\mathscr{C} \times \mathcal{D}) = kraftsum(\mathscr{C}) \times kraftsum(\mathcal{D})$$

Proof.

$$\begin{aligned} kraftsum(\mathscr{C} \times \mathcal{D}) &= \sum_{u,v} 2^{-(l(\mathscr{C})*l(\mathcal{D}))} \\ &= \sum_u 2^{-l(\mathscr{C})} \sum_v 2^{-l(\mathcal{D})} \end{aligned}$$

□

Corollary 1.2.1. *$kraftsum(\mathscr{C}^n) = (kraftsum(\mathscr{C}))^n$*

Proposition 1.1. *if $\mathscr{C}$ is non-singular, then*

$$kraftsum(\mathscr{C}) \leq 1 + \max_u l(\mathscr{C}(u))$$

In coding theory, the **Kraft-McMillan inequality** gives a necessary and sufficient condition for the existence of a uniquely decodable code for a given set of codeword lengths.

Wikipedia. *Kraft's inequality limits the lengths of codewords in a prefix code: if one takes an exponential of the length of each valid codeword, the resulting set of values must look like a probability mass function, that is, it must have total measure less than or equal to one. Kraft's inequality can be thought of in terms of a constrained budget to be spent on codewords, with shorter codewords being more expensive.*

Theorem 1.3. *if $\mathscr{C}$ is uniquely decodable, then $kraftsum(\mathscr{C}) \leq 1$*

Proof. $\mathscr{C}$ is uniquely decodable $\equiv \mathscr{C}^*$ is non singular

$$\Rightarrow kraftsum(\mathscr{C}^n) \leq 1 + \max_{u_1,\ldots,u_n} l(\mathscr{C}^n)$$

$$\Rightarrow kraftsum(\mathscr{C})^n \leq 1 + nL, \quad L = \max l(\mathscr{C}(n))$$

A growing exp cannot be bounded by a linear function

$$\Rightarrow kraftsum(\mathscr{C}) \leq 1$$

□

Theorem 1.4. *Suppose $\mathscr{C} : \mathcal{U} \to \mathcal{N}$ is such that $\sum_u 2^{-l(\mathscr{C}(u))} \leq 1$, then, there exists a prefix-free code $\mathscr{C}' : \mathcal{U} \to \{0,1\}$ s.t. $\forall u, l(\mathscr{C}'(u)) = l(\mathscr{C}(u))$*

Proof. Let $\mathcal{U} = \{u_1, ..., u_n\}$ and $\mathscr{C}(u_1) \leq \mathscr{C}(u_2) \leq ... \leq \mathscr{C}(u_k) = \mathscr{C}_{max}$. Consider the complete binary tree up to depth $\mathscr{C}_{max}$ initially all nodes are available to be used as codewords. For $i = 1, 2, ..., n$, place $\mathscr{C}(u_i)$ at an available node at level $\mathscr{C}(u_i)$ remove all descendant of $\mathscr{C}(u_i)$ from the available list.

Corollary 1.4.1. *Suppose $\mathscr{C} : \mathcal{U} \to \{0,1\}^*$ is uniquely decodable, then there exist an $\mathscr{C}' : \mathcal{U} \to \{0,1\}^*$ which is prefix-free and $l(\mathscr{C}'(n)) = l(\mathscr{C}(n))$*

□

Example 1. $\mathcal{U} = \{a, b, c, d\}$, $\mathscr{C} : \{0, 01, 011, 111\}$ and $\mathscr{C}' : \{0, 10, 110, 111\}$
In this case, decoding $\mathscr{C}$ may require delay, while decoding $\mathscr{C}'$ is instanteneous.

2 Alphabet with statistics

Suppose we have an alphabet $\mathcal{U}$, and suppose we have a random variable U taking values in $\mathcal{U}$. We denote by $p(u) = Pr(U = u), u \in \mathcal{U}$ with $p(u) \geq 0$ and $\sum_u p(u) = 1$.

Suppose we have a code $\mathscr{C} : \mathcal{U} \to \{0,1\}^*$. We then have $\mathscr{C}(u)$ a random binary string and $l(\mathscr{C}(u))$ a random integer.

Example 2. $\mathcal{U} = \{a, b, c, d\}$
$p : \{0.5, 0.25, 0.125, 0.125\}$
$\mathscr{C} : \{0, 01, 110, 111\}$

then we have

$$l(\mathscr{C}(u)) = \begin{cases} 1, & p = 0.5 \\ 2, & p = 0.25 \\ 3, & p = 0.125 + 0.125 + 0.25 \end{cases}$$

We can measure how efficient $\mathscr{C}$ represents $\mathcal{U}$ by considering

$$E[l(\mathscr{C}(u))] = \sum_u p(u) l(\mathscr{C}(u)) \quad \text{with} \quad \mathscr{C}(u) = l(\mathscr{C}(u))$$

Theorem 2.1. *if $\mathscr{C}$ is uniquely decodable, then*

$$E[l(\mathscr{C}(u))] \geq \sum_u p(u) \log(\frac{1}{p(u)})$$

Proof. let $\mathscr{C}(u) = l(\mathscr{C}(u))$, we know $\sum_u 2^{-\mathscr{C}(u)} \leq 1$ because $\mathscr{C}$ is uniquely decodable. We write $q(u) = 2^{-\mathscr{C}(u)}$ and get

$$\begin{aligned} E[l(\mathscr{C}(u))] = \sum_u p(u)\mathscr{C}(u) &= \sum_u p(u) \log_2 \frac{1}{q(u)} \\ &\equiv \sum_u p(u) \log \frac{q(u)}{p(u)} \leq 0 \\ &\equiv \sum_u p(u) \ln \frac{q(u)}{p(u)} \leq 0 \\ &\leq \sum_u p(u) \left[\frac{q(u)}{p(u)} - 1\right] = \underbrace{\sum_u q(u)}_{\leq 1} - \underbrace{\sum_u p(u)}_{=1} \leq 0 \end{aligned}$$

□

Theorem 2.2. *For any $\mathcal{U}$, there exists a prefix-free code $\mathscr{C}$ s.t.*

$$E[l(\mathscr{C}(u))] < 1 + \sum_{u \in \mathcal{U}} p(u) \log \frac{1}{p(u)}$$

Proof. Given $\mathcal{U}$, let

$$\begin{aligned} &\mathscr{C}(u) = \lceil \log \frac{1}{p(u)} \rceil < 1 + \log \frac{1}{p(u)} \\ &\Rightarrow \sum_u 2^{-\mathscr{C}(u)} \leq \sum_u p(u) = 1 \\ &\Rightarrow \sum_u p(u)\mathscr{C}(u) < \sum_u p(u) \log(\frac{1}{p(u)}) + \underbrace{1}_{\sum p(u)} \end{aligned}$$

□

Definition 2.1 (Entropy)**.** *Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process.*

Theorem 2.3. *The entropy of a random variable $U \in \mathcal{U}$ is*

$$H(U) = \sum_{u \in \mathcal{U}} p(u) \log(\frac{1}{p(u)})$$

with $p(u) = Pr(U = u)$

Wikipedia. *The entropy is a lower bound on the optimal expected length*

$$H(U) \leq \mathbb{E}\, l(\mathscr{C}(u))$$

In fact, one can show that there exists a uniquely decodable code such that

$$H(U) \leq \mathbb{E}\, l(\mathscr{C}(u)) < H(U) + 1$$

Note that $H(U)$ is a fonction of the distribution $\mathscr{C}_u(.)$ of the random variable U, it isn't a function of U.

$$H(U) = E[f(U)] \quad \text{where} \quad f(U) = \log(\frac{1}{p(u)})$$

How to design optimal codes (in the sense of minimizing $E[l(\mathscr{C}(u))]$)?
Formally, given a random variable U, find $\mathscr{C}(u) \to \mathcal{N}$ s.t.

$$\sum_{u \in U} 2^{\mathscr{C}(u)} \leq 1 \quad \text{that minimizes} \quad \sum_{u \in U} p(u)\mathscr{C}(u)$$

Properties of optimal prefix-free codes

- if $p(u) < p(v)$ then $\mathscr{C}(u) \geq \mathscr{C}(v)$
- The two longest codewords have the same length
- There is an optimal code such that the two least probable letters are assigned codewords that differ in the last bit.

Observe that if $\{\mathscr{C}(u_1), ..., \mathscr{C}(u_{k-1}), \mathscr{C}(u_k)\}$ is a prefix-free collection of the property that

$$\begin{array}{ll} \mathscr{C}(u_{k-1}) & = \alpha 0 \\ \mathscr{C}(u_k) & = \alpha 1 \end{array} \quad \text{with} \quad \alpha \in \{0,1\}^*$$

then $\{\mathscr{C}(u_1), ..., \mathscr{C}(u_{k-2}), \alpha\}$ is also a prefix-free collection.
Also

$$\begin{aligned} \sum_{u \in \mathcal{U}} p(u) l(\mathscr{C}(u)) &= p(u_1) l(\mathscr{C}(u_1)) + ... + p(u_{k-2}) l(\mathscr{C}(u_{k-2})) + [p(u_{k-1}) + p(u_k)](l(\alpha) + 1) \\ &= (p(u_{k-1}) + p(u_k)) + \sum_{v \in \mathcal{V}} p(v) l(\mathscr{C}'(v)) \end{aligned}$$

So we have shown that with

$$E[l(\mathscr{C}(U)] = p(u_{k-1}) + p(u_k) + E[l(\mathscr{C}'(V))]$$

if $\mathscr{C}$ is optimal for U, then $\mathscr{C}'$ is optimal for V

3 Entropy and mutual information

Definition 3.1 (Joint entropy). *Suppose U, V are random variables with $p(u,v) = Pr\{U = u, V = v\}$, the joint entropy is*

$$H(UV) = \sum_{u,v} p(u,v) \log \frac{1}{p(u,v)}$$

Theorem 3.1.

$$H(UV) \leq H(U) + H(V)$$

with equality iff U and V are independants.

Proof. We want to show that

$$\sum_{u,v} p(u,v) \log \frac{1}{p(u,v)} \leq \sum_{u} p(u) \log \frac{1}{p(u)} + \sum_{v} p(v) \log \frac{1}{p(v)} \iff \sum_{u,v} p(u,v) \log \frac{p(u)p(v)}{p(u,v)} \leq 0$$

We use $\ln z \leq z - 1$ for all z (with equality iff $z = 1$):

$$\sum_{u,v} p(u,v) \log \frac{p(u)p(v)}{p(u,v)} \leq \sum_{u,v} p(u,v) \left[\frac{p(u)p(v)}{p(u,v)} - 1\right] = \sum_{u,v} p(u)p(v) - \sum_{u,v} p(u,v) = 1 - 1 = 0$$

□

Same definitions of entropy holds for n symbols.

Definition 3.2 (Joint Entropy). *Suppose $U_1, U_2, \ldots, U_n$ are RVs and we are given $p(u_1 \ldots u_n)$, the joint entropy is*

$$H(U_1, \ldots, U_n) = \sum_{u_1 \ldots u_n} p(u_1 \ldots u_n) \log \frac{1}{p(u_1 \ldots u_n)}$$

Theorem 3.2.

$$H(U_1 \ldots U_n) \leq \sum_{i=1}^{n} H(U_i)$$

with equality iff Us are independants

Corollary 3.2.1. *if $U_1, \ldots, U_n$ are i.i.d. then $H(U_1 \ldots U_n) = nH(U_1)$*

Definition 3.3 (Conditional entropy).

$$\begin{aligned} H(U|V) &= \sum_{u,v} p(u,v) \log \frac{1}{p(u|v)} \\ &= \sum_{v} H(U|V = v) Pr\{V = v\} \end{aligned}$$

Theorem 3.3.

$$H(UV) = H(U) + H(V|U) = H(V) + H(U|V)$$

Theorem 3.4.

$$H(U) + H(V) \geq H(UV) = H(V) + H(U|V)$$

Definition 3.4 (Mutual information). *Mutual information measures the amount of information that can be obtained about one random variable by observing another.*

$$\begin{aligned} I(U;V) = I(V;U) &= H(U) - H(U|V) \\ &= H(V) - H(V|U) \\ &= H(U) + H(V) - H(UV) \end{aligned}$$

We can apply the chain rule on the entropy as follow

$$H(U_1U_2\dots U_n) = H(U_1) + H(U_2|U_1) + \dots + H(U_n|U_1U_2\dots U_{n-1}) = \sum_{i=1}^{n} H(U_i|U^{i-1})$$

Definition 3.5 (Conditional mutual information)**.**

$$\begin{aligned} I(U;V|W) &= H(U|W) - H(U|VW) \\ &= H(V|W) - H(V|UW) \\ &= \mathbb{E}_{u,v,w}\left[\log \frac{p(uv|w)}{p(u|w)p(v|w)}\right] \end{aligned}$$

Theorem 3.5.

$$I(V;U_1\dots U_n) = I(V;U_1) + I(V;U_2|U_1) + \dots + I(V;U_n|U_1\dots U_{n-1})$$

We can apply the chain rule on the mutual information as follows

$$I(U_1, U_2, ...; V) = I(U_1;V) + I(U_2;V|U_1) + ...$$

Theorem 3.6. *Data processing inequality Let $X \to Y \to Z$ be a Markov chain, then*

$$I(X;Y) \geq I(X;Z)$$

Notation 1.

$$U^n \triangleq (U_1U_2\dots U_n)$$

Theorem 3.7.

$$I(U;V|W) \geq 0$$

equality iff conditioned on w, u and v are independant, that is iff $U - V - W$ is a Markov chain.

Proof.

$$\begin{aligned} I(U;V|W) &= \frac{1}{\ln 2}\sum_{u,v,w} p(uvw) \ln \frac{p(u|w)p(v|w)}{p(uv|w)} \\ &\geq \frac{1}{\ln 2}\sum_{u,v,w} p(uvw)\left[\frac{p(u|w)p(v|w)}{p(uv|w)} - 1\right] \\ &= \frac{1}{\ln 2}\sum_{u,v,w} (p(w)p(u|w)p(v|w) - p(uvw)) \\ &= \frac{1}{\ln 2}(1-1) \\ &= 0 \end{aligned}$$

□

4 Data processing

Theorem 4.1. *$U - V - W$ is a MC $\iff I(U;W|V) = 0$*

Corollary 4.1.1. *$I(U;V) \geq I(U;W)$ and by symetry of MC $I(W;V) \geq I(U;W)$*

Proof.

$$I(U;VW) = I(U;V) + I(U;W|V) = I(U;V)$$

and

$$I(U;VW) = I(U;W) + I(U;V|W) \geq I(U;W)$$

$\square$

Theorem 4.2. *Given U a RV taking values in $\mathcal{U}$ then $0 \leq H(U) \leq \log|\mathcal{U}|$. $H(U) = 0$ iff U is constant, $H(U) = \log|\mathcal{U}|$ iff U is $p(u) = 1/|\mathcal{U}|$ for all u.*

Proof. For the lower bound,

$$H(U) = \sum_u \underbrace{p(u)}_{\geq 0} \underbrace{\log \frac{1}{p(u)}}_{\geq 0} \geq 0$$

For the upper bound,

$$\begin{aligned}
H(U) - \log|\mathcal{U}| &= \sum_u p(u) \log \frac{1}{p(u)} - \sum_u p(u) \log|\mathcal{U}| \\
&= \frac{1}{\ln 2} \sum_u p(u) \ln \frac{1}{|\mathcal{U}|p(u)} \\
&\leq \frac{1}{\ln 2} \sum_u p(u) \left(\frac{1}{|\mathcal{U}|p(u)} - 1 \right) \\
&= \frac{1}{\ln 2} \left[\sum_u \frac{1}{|\mathcal{U}|} - \sum_u p(u) \right] \\
&= 0
\end{aligned}$$

$\square$

Theorem 4.3. *$I(U;V) = 0 \iff U \perp V$*

Definition 4.1 (Entropy rate of a stochastic process)**.**

$$r = \lim_{n\to\infty} \frac{1}{n} H(U^n) \quad \textit{if the limit exists}$$

Theorem 4.4. *For stationary stochastic process U^n, the sequences*

$$a_n = \frac{1}{n} H(U^n) \textit{ and } b_n = H(U_n|U^{n-1})$$

are positive and non increasing. Then $a = \lim_{n\to\infty} a_n$ and $b = \lim_{n\to\infty} b_n$ exists and $a = b$.

Proof.

$$\begin{aligned}
b_{n+1} &= H(U_{n+1}|U_1, U_2, \ldots, U_n) \\
&\leq H(U_{n+1}|U_2, \ldots, U_n) \\
&= H(U_n|U_1, U_2, \ldots, U_{n-1}) \\
&= b_n \text{ , because } U_1 \ldots U_n \sim U_2 \ldots U_{n+1} \text{ (Stationarity).}
\end{aligned}$$

Hence, it is non-increasing.

For the $\{a_n\}$, observe that

$$a_n = \frac{1}{n}H(U^n) = \frac{1}{n}\Big[H(U_1) + H(U_2|U_1) + H(U_3|U^2) + \cdots + H(U_n|U^{n-1})\Big]$$
$$= \frac{1}{n}\Big[b_1 + b_2 + \cdots + b_n\Big]$$

and by the "Lemma", whenever $b_n \to b$, $a_n \to b$ □

Lemma 4.5 (Cesaro)**.** *Suppose $b_n \to b$,*

then,

$$a_n = \frac{1}{n}\Big[b_1 + b_2 + \cdots + b_n\Big] \textit{ also converges and to 1.}$$

Proof. Since $b_n \to b$, $\Big(\equiv \forall\epsilon > 0 \text{ , } \exists\ n(\epsilon) \text{ s.t } \forall n > n(\epsilon)\ |b_n - b| < \epsilon\Big)$

$\exists B$ s.t. $|b_n| < B$ for all n.

Take $n > n_1(\epsilon) \triangleq \ldots$ then

$$|a_n - b| \leq \frac{|b_1 - b| + |b_2 - b| + |b_3 - b| + \cdots + |b_n - b|}{n}$$

$$\text{so } |a_n - b| \leq \frac{1}{n}\Bigg[\sum_{i=1}^{n_0(\epsilon)} \underbrace{|b_i - b|}_{2B} + \sum_{i=n_0(\epsilon)+1}^{n} \underbrace{|b_i - b|}_{\leq\epsilon}\Bigg] \leq \frac{n_0(\epsilon)2B}{n} + \epsilon < 2\epsilon$$

$$\text{for } n > n_1(\epsilon) \triangleq \max, \{n_0(\epsilon)\frac{1}{\epsilon}n_0(\epsilon)2B\}$$

□

Theorem 4.6. *Given a stationary process with entropy rate r:*

$$r = \lim_{n\to\infty}\frac{1}{n}H(U^n)$$

then

1. *for every source coding scheme*

$$\mathscr{C}_n : U^n \to \{0,1\}^*$$

the expected number of bits / letter is given by

$$\frac{1}{n}E[l(\mathscr{C}(U^n))] \geq r$$

2. *for any $\epsilon > 0$, there exists a source coding scheme $\mathscr{C}_n : U^n \to \{0,1\}^*$ s.t.*

$$\frac{1}{n}E[l(\mathscr{C}_n(U^n))] < r + \epsilon$$

Proof. 1. we already know

$$\frac{1}{n}E[l(\mathscr{C}_n(U^n))] \geq \frac{1}{n}H(U^n)$$

and the right term is decreasing

2. we also know that for each $n, \exists \mathscr{C}_n$ that is prefix-free s.t.

$$E[l(\mathscr{C}_n(U^n))] < \underbrace{\frac{1}{n}H(U^n)}_{r} + \underbrace{\frac{1}{n}}_{0}$$

we can find n large enough s.t. the right hand side $< r + \epsilon$

□

5 Typicality and typical set

Definition 5.1 (Typicality)**.** *Suppose we have a sequence $U_1, U_2, ...$ of i.i.d. random variables taking values in an alphabet $\mathcal{U}$. Suppose we observe $u_1, u_2 ..., u_n$. We will call it to be typical-(ϵ, p) if*

$$p(u)(1-\epsilon) \leq \frac{\#\text{ of times } u \text{ appears in } u_1, ..., u_n}{n} \leq p(u)(1+\epsilon)$$

Theorem 5.1. *u^n is (ϵ, p)-typical then*

$$2^{-nH(u)(1+\epsilon)} \leq Pr(U^n = u^n) \leq 2^{-nH(u)(1-\epsilon)}$$

Proof.

$$Pr(U^n = u^n) = \prod_{i=1}^{n} Pr(U_i = u_i) = \prod_{i=1}^{n} p(u_i) = \prod_{u \in \mathcal{U}} p(u)^{\#_u}$$

with $\#_u$ the number of times u appears in $u_1, ..., u_n$ where

$$n(1-\epsilon)p(u) \leq \#_u \leq n(1+\epsilon)p(u)$$

consequently

$$p(u)^{(np(u)(1-\epsilon))} \geq p(u)^{\#_u} \geq p(u)^{np(u)(1+\epsilon)}$$

then

$$(\prod_{n} p(u)^{p(u)})^{(1-\epsilon)n} \geq Pr(U^n = u^n) \geq (\prod_{n} p(u)^{p(u)})^{(1+\epsilon)n}$$

but

$$p(u)^{p(u)} = 2^{-p(u)\log(\frac{1}{p(u)})} \Rightarrow \prod p(u)^{p(u)} = 2^{-H(u)}$$

□

Definition 5.2 (Typical set)**.**

$$T(n, \epsilon, p) = \{u^n \in \mathcal{U}^n : u^n \textit{ is } (\epsilon, p)\textit{-typical}\}$$

Wikipedia. *Typical sets provide a theoretical means for compressing data, allowing us to represent any sequence X^n using $nH(X)$ bits on average, and, hence, justifying the use of entropy as a measure of information from a source.*

Theorem 5.2. *1. if $u^n \in T(n,\epsilon,p)$ then*

$$p(u^n) = Pr(U^n = u^n) = 2^{-nH(u)(1\pm\epsilon)}$$

when U_i i.i.d.

2.

$$\lim_{n\to\infty} Pr(U^n \in T(n,\epsilon,p)) = 1$$

3.

$$|T(n,\epsilon,p)| \le 2^{n(H(u)(1+\epsilon))}$$

4.

$$|T(n,\epsilon,p)| \ge (1-\epsilon)2^{nH(u)(1-\epsilon)}$$

Proof. 1. Fix $u \in \mathcal{U}$ let $X_i = 1$ if $U_i = u$ and 0 otherwise

$$\frac{\#\text{ of times } u \text{ appears in } U_1 ... U_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

observe that $\{X_i\}$ are i.i.d.

$$X_i = \begin{cases} 1 & \text{w.p. } p(u) \\ 0 & \text{w.p. } 1-p(u) \end{cases}$$
$$\Rightarrow E[X_i] = p(u) \quad \text{and} \quad Var[X_i] = p(u) - p(u)^2$$

$$\underbrace{Pr\left\{|\frac{1}{n}\sum_{i+1}^{n} X_i - p(u)|\right\} \ge \epsilon p(u)}_{u^n \text{ fails the test for letter } u} \le \frac{Var(\frac{1}{n}\sum X_i)}{(\epsilon p(u))^2} = \frac{(1-p(u))}{\epsilon^2 p(u)}$$

2.

$$\begin{aligned} Pr\{U^n \notin T(n,\epsilon,p)\} &= Pr\left\{\bigcup_{u\in U}\{u^n \text{ fails the test for u}\}\right\} \\ &\le \sum_{u\in U} Pr\{U^n \text{ fails the test for } u\} \\ &\le \frac{1}{n}\sum_{u\in U}\frac{(1-p(u))}{p(u)\epsilon^2} \qquad \text{which goes to 0 as } n \text{ gets large} \end{aligned}$$

3.

$$\begin{aligned} 1 \ge Pr\{U^n \in T(n,\epsilon,p)\} &= \sum_{u^n\in T(n,\epsilon,p)} Pr\{U^n = u^n\} \\ &\ge \sum_{u^n\in T(n,\epsilon,p)} 2^{-n(1+\epsilon)H(u)} \\ &= 2^{-n(1+\epsilon)H(u)}|T(n,\epsilon,p)| \end{aligned}$$

4.
$$\begin{aligned} 1-\epsilon \le Pr\left\{U^n \in T(n,\epsilon,p)\right\} &= \sum_{u^n \in T(n,\epsilon,p)} Pr\left\{U^n = u^n\right\} \\ &\le \sum_{u^n \in T(n,\epsilon,p)} 2^{nH(u)(1-\epsilon)} \\ &= 2^{-nH(u)(1-\epsilon)} |T(n,\epsilon,p)| \end{aligned}$$

□

Observation 5.1. *$Pr\left\{U^n \in T(n,\epsilon,p)\right\} \to 1$ as $n \to \infty$*

Definition 5.3 (Kullback Leibler divergence)**.**

$$D(p||q) = \sum_u p(u) \log \frac{p(u)}{q(u)} \ge 0 \textit{ with equality iff } p = q$$

If we compress data in a manner that assumes $q(u)$ is the distribution underlying some data, when, in reality, $p(u)$ is the correct distribution, the Kullback-Leiber divergence is the average number of additional bits per datum necessary for compression. It is also called **relative entropy** and is a measure of how one probability distribution diverges from a second probability distribution.

Lemma 5.3. *if $U_1 \dots U_n$ are i.i.d. with distribution q and $u_1 \dots u_n$ is (ϵ, p)-tipycal, then*

$$2^{-n[H(p)+D(p||q)](1+\epsilon)} \le Pr\left\{U^n = u^n\right\} \le 2^{-n[H(p)+D(p||q)](1-\epsilon)}$$

Proof. Follows from

$$\left[\prod_u q(u)^{p(u)}\right]^{n(1+\epsilon)} \le Pr\left\{U^n = u^n\right\} \le \left[\prod_u q(u)^{p(u)}\right]^{n(1-\epsilon)}$$

$$\prod_u q(u)^{p(u)} = 2^{-\sum p(u) \log \frac{1}{q(u)}}$$

and

$$\sum_u p(u) \log \frac{1}{q(u)} = \underbrace{\sum_u p(u) \log \frac{1}{p(u)}}_{H(p)} + \underbrace{\sum_u p(u) \log \frac{p(u)}{q(u)}}_{D(p||q)}$$

□

Corollary 5.3.1. *if $U_1 \dots U_n$ are i.i.d. following distribution q, then*

$$2^{-n[(1+\epsilon)D(p||q)+2\epsilon H(p)]} \le Pr\left\{U^n \in T(n,\epsilon,p)\right\} \le 2^{-n[(1-\epsilon)D(p||q)-2\epsilon H(p)]}$$

Proof.

$$Pr\left\{U^n \in T(n,\epsilon,p)\right\} = \sum_{u^n \in T(n,\epsilon,p)} Pr\left\{U^n = u^n\right\}$$

We have

$$\begin{aligned} 2^{-n[H(p)+D(p||q)](1+\epsilon)} &\le Pr\left\{U^n = u^n\right\} \le 2^{-n[H(p)+D(p||q)](1-\epsilon)} \\ 2^{nH(p)(1-\epsilon)} &\le |T(n,\epsilon,p)| \le 2^{nH(p)(1+\epsilon)} \end{aligned}$$

□

Example 3. $U \in \{0,1\}$, $p = \frac{1}{2}, \frac{1}{2}$, $q = \frac{1}{2} - \delta, \frac{1}{2} + \delta$

$$D(p||q) = \frac{1}{2}\log\frac{1}{1-2\delta} + \frac{1}{2}\log\frac{1}{1+2\delta} = \frac{1}{2}\log\frac{1}{1-4\delta^2} = -\frac{1}{2}\log(1-4\delta^2) \approx \frac{1}{2}4\delta^2 + o(\delta^4)$$

So if we want $2^{-nD(p||q)}$ small, we must pick $n = \Omega(1/\delta^2)$

Example 4. Suppose we are told that U is p distributed and $p(u)$ are powers of 2. We design a prefix-free code $\mathscr{C}$ to minimize $\sum_u p(u)l(\mathscr{C}(u))$. We have been misinformed and $U \sim q$, then:

$$\begin{aligned} E\left[l(\mathscr{C}(u))\right] &= \sum_u q(u)\log\frac{1}{p(u)} \\ &= \underbrace{H(q)}_{\text{length for optimal code}} + \underbrace{D(q||p)}_{\text{penalty for misbelief}} \end{aligned}$$

5.1 Universal data compression

Suppose we know that the distribution p of U is either p_1, p_2 ... p_k, can we design a code $\mathscr{C} : U \to \{0,1\}^*$

$$E\left[l(\mathscr{C}(U))\right] \leq H(U) + \text{small for every } p$$

$$E\left[\frac{1}{n}l(\mathscr{C}(U))\right] \leq o(n) + E\left[h_2\left(\frac{K}{n}\right)\right]$$

with $K = \sum_{i=1}^n u_i$

We have $\frac{E[K]}{n} = \theta_1$ and $E\left[h_2\left(\frac{K}{n}\right)\right] \leq h_2\left(E\left[\frac{K}{n}\right]\right) = h_2(\theta)$

Design $\mathscr{C}$ Because the probability of a bit string is only dependant of the number of 1s (or 0s), it makes sense to encode two strings with the same numbers of 1 with code words of same lengths. Given $u_1 \dots u_n \in \{0,1\}^n$, first count the number of 1, call it k.

$$\mathscr{C}(u_1 \dots u_n) = \underbrace{\text{describe } k}_{\lceil \log(n+1) \rceil}\underbrace{\text{describe } u_1 \dots u_n}_{\lceil \log \binom{n}{k} \rceil}$$

We now want to evaluate

$$\frac{1}{n}E\left[l(\mathscr{C}(U))\right]$$

when $U_1 \dots U_n$ are i.i.d with $p_1 = \theta$ and $p_0 = 1 - p_1$

Observation 5.2. *for any* $0 \leq \alpha \leq 1$

$$\begin{aligned} 1 = 1^n = (\alpha + (1-\alpha))^n &= \sum_{i=0}^n \binom{n}{i}\alpha^i(1-\alpha)^{n-i} \\ &\geq \binom{n}{k}\alpha^k(1-\alpha)^{n-k} \end{aligned}$$

Then for all α

$$\binom{n}{k} \leq \alpha^{-k}(1-\alpha)^{-(n-k)} = 2^{n(\frac{k}{n}\log\frac{1}{\alpha} + (1-\frac{k}{n})\log\frac{1}{1-\alpha})}$$

We pick $\alpha = \frac{k}{n}$, *and we get*

$$\binom{n}{k} < 2^{nh_2\left(\frac{k}{n}\right)}$$

Using this bound we have

$$\frac{1}{n} l(\mathscr{C}(u_1 \dots u_n)) \le \frac{2}{n} + \frac{\log(n+1)}{n} + h_2\left(\frac{k}{n}\right)$$

$$E\left[\frac{1}{n} l(\mathscr{C}(U))\right] \le o(n) + E\left[h_2\left(\frac{k}{n}\right)\right]$$

Claim 5.1. *Suppose U_i are i.i.d. with $Pr\{U_1 = 1\} = \theta$. We have $E\left[\frac{k}{n}\right] = \theta$ and $E\left[h_2\left(\frac{k}{n}\right)\right] \le h_2(E\left[\frac{k}{n}\right]) = h_2(\theta)$. So*

$$\lim_{n\to\infty} \frac{1}{n} E\left[l(\mathscr{C}(u_1 \dots u_n))\right] \le h_2(\theta)$$

consequently this scheme is asymptotically optimal.

Proof. To prove the claim we need to show that if $\beta_1 \dots \beta_k$ are in $[0,1]$ and $q_1 \dots q_k$ are non negative numbers that sum to 1 then

$$\sum_{i=1}^{k} q_i h_2(\beta_i) \le h_2\left(\sum_{i=1}^{k} q_i \beta_i\right)$$

Let U and V be random variables with $U \in \{0,1\}$ and $V \in \{1,\dots,k\}$ with

$$\begin{aligned} Pr\{V = i\} &= q_i \\ Pr\{U = 1 | V = i\} &= \beta_i \\ Pr\{U = 0 | V = i\} &= 1 - \beta_i \end{aligned}$$

Then,

$$\begin{aligned} Pr\{U = 1\} &= \sum_i q_i \beta_i \\ H(U) &= h_2\left(\sum_i q_i \beta_i\right) \\ H(U|V) &= \sum_i q_i h_2(\beta_i) \end{aligned}$$

And we already know that $H(U) \ge H(U|V)$ □

TODO: Thomas scribes here

Suppose we have an infinite string $u_1 u_2 \dots, u_i \in \mathcal{U}$, and

$$u_1 u_2 \dots = v_1 v_2 \dots \text{ with } v_i \in \mathcal{U}^*, v_i \ne v_j \text{ when } i \ne j$$

for any k we have

$$\lim_{m\to\infty} \frac{length(v_1 \dots v_m)}{m} \ge k \Rightarrow \lim_{m\to\infty} \frac{length(v_1 \dots v_m)}{m} = \infty$$

Definition 5.4. *Given an infinite string $u_1 u_2 \dots$ and a machine M, let*

$$\rho_M(u_1 u_2 \dots) = \overline{\lim_{n\to\infty}} \frac{\textit{length of the output } M \textit{ after reading } u_1 u_2 \dots}{n}$$

also given $s > 0$, define

- *The compressibility of $\mathcal{U}^*$ be s-state machines*

$$\rho_s(u_1u_2...) = \min_M \rho_M(u_1u_2...)$$

  with M an s'-state machine with $s' \leq s$

- *Compressibility of $\mathcal{U}^*$ by finite state machines*

$$\rho_{FSM}(u_1u_2...) = \lim_{s\to\infty} \rho_s(u_1u_2...)$$

Definition 5.5. *Suppose $u_1u_2...$ an infinite sequence, define $m(n)$ as the largest m for which $u_1...u_n = v_1...v_m$ with distinct $v_1...v_m$*

Example 5.

$$u = aaaaaaaaa, \quad \underbrace{\emptyset}_{v_1}\underbrace{a}_{v_2}\underbrace{aa}_{v_3}\underbrace{aaa}_{v_4}\underbrace{aaaa}_{v_5} \quad \Rightarrow m(10) = 5$$

So far we know that

$$\frac{\text{length of the output of any s-state IL machine when it reads } u_1u_2...}{n} \geq \frac{m(n)\log(\frac{m(n)}{8s^2})}{n}$$

with

$$\frac{m(n)\log(\frac{m(n)}{8s^2})}{n} = \frac{m(n)\log(m(n))}{n} - \frac{m(n)\log(8s^2)}{length(v_1...v_m)}$$

hence if M is a s-state machine

$$\rho_M(u_1u_2...) \geq \overline{\lim_{n\to\infty}}\frac{m(n)\log(m(n))}{n} \quad \text{then} \quad \rho_{FSM}(u_1u_2...) \geq \overline{\lim_{n\to\infty}}\frac{m(n)\log m(n)}{n}$$

6 Lemple-Ziv data compression method

Given some alphabet $\mathcal{U}$ to both encoder and decoder, they also agree an order on $\mathcal{U}$:

1. Start with a dictionary $\mathcal{D} = \mathcal{U}$
2. To each word $w \in \mathcal{D}$, assign a $\lceil \log |\mathcal{D}| \rceil$-bit binary description in the dictionary order
3. Parse the first word w in $u_1 u_2 ...$ in the dictionary, output its binary description
4. replace w in $\mathcal{D}$ by $\{wu, \forall u \in \mathcal{U}\}$.
5. Go to 2.

Example 6. Define an alphabet $\mathcal{U} = \{a, b, c\}$ with $a \leq b \leq c$ and an input message

$$u = abacac$$

- Create the dictionary $\mathcal{D} = \{a, b, c\}$ and its corresponding binary description $\mathcal{D}_{bin} = \{00, 01, 10\}$
- The first word in the message is $'a'$, output its binary description

$$output = 01$$

- Update the dictionary:

$$\mathcal{D} = \{a, ba, bb, bc, c\} \quad \mathcal{D}_{bin} = \{000, 001, 010, 011, 100\}$$

- Parse the next word $'ba'$ and output its binary description

$$output = 01001$$

- Update the dictionary

$$\mathcal{D} = \{a, baa, bab, bac, bb, bc, c\} \quad \mathcal{D}_{bin} = \{000, 001, ...\}$$

- Continue until the end of the input data...

The decoder can proceed in a similar way to iteritavely update the dictionary while decoding the message.

6.1 Analysis of LZ

Observe that LZ parses the string $u_1 u_2 ...$ into $v_1 v_2 ...$ with $v_i \in \mathcal{U}^*$ or $v_i \in \mathcal{D}_i$ where $\mathcal{D}_i$ is the dictionary at step i.

When going from iteration $i \to i+1$, v_i is removed from $\mathcal{D}$, consequently v_1, v_2, v_3 are distinct.

The length of the output of LZ after reading $u_1 ... u_m$ is given by

$$\text{LZ output's length} = \lceil \log |\mathcal{U}| \rceil + \lceil \log(2|\mathcal{U}| - 1) \rceil + \lceil \log(3|\mathcal{U}| - 2) \rceil + ... + \lceil \log(m|\mathcal{U}| - m + 1) \rceil$$

we observe that

$$\text{LZ output's length} < m(\log(m|\mathcal{U}|) + 1) = m \log(2m|\mathcal{U}|)$$

Also we have

$$\begin{aligned} \text{\# bits / letter} &< \frac{m \log(2m|\mathcal{U}|)}{length(u_1 ... u_m)} \\ &= \frac{m \log(m)}{length(u_1 ... u_m)} + \frac{m \log(2|\mathcal{U}|)}{length(u_1 ... u_m)} \end{aligned}$$

therefore

$$\rho_{LZ}(u_1 u_2 ...) = \lim_{m \to \infty} \frac{\text{\# bits}}{\text{letter}} \leq \lim_{m \to \infty} \frac{m \log(m)}{length(u_1 ... u_m)} \leq \lim_{n \to \infty} \frac{m(n) \log(m(n))}{n} \leq \rho_{FSM}(u_1 u_2 ...)$$

So we have proved the following theorem:

Theorem 6.1. *for every $u_1u_2...$*

$$\rho_{LZ}(u_1u_2...) \leq \rho_{FSM}(u_1u_2...)$$

Corollary 6.1.1. *if $u_1u_2...$ is stationary*

$$\rho_{LZ}(u_1u_2...) = \text{ entropy rate of } u_1u_2...$$

7 Transmission of data

Interesting in the case of unreliable transmission media.

Definition 7.1 (Communication channel). *A communication channel W is a device with an input alphabet $\mathcal{X}$ and an output alphabet $\mathcal{Y}$. Its behabvior is described by*

$$W_i(y_i|x^i, y^{i-1}) = Pr\left\{Y_i = y_i|X^i = x^i, Y^{i-1} = y^{i-1}\right\}$$

Definition 7.2 (Memoryless channel). *a channel W is said to be memoryless if*

$$W_i(y_i|x^i, y^{i-1}) = W(y_i|x_i)$$

Definition 7.3 (Stationary channel). *a channel W is said to be stationary if*

$$W_i(y|x) = W(y|x)$$

Example 7 (Binary erasure channel - BEC). $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, ?\}$, then

$$\begin{aligned} W(0|0) &= 1 - p \\ W(?|0) &= p \\ W(1|0) &= 0 \end{aligned}$$

and same for $x_i = 1$.

Example 8 (Binary symetric channel - BSC).

$$\begin{aligned} W(0|0) &= 1 - p = W(1|1) \\ W(1|0) &= p = W(0|1) \end{aligned}$$

The input $X_1, X_2 \dots X_n$ to a channel might have memory

$$Pr\left\{X^n = x^n\right\} = p(x_1)p(x_2|x_1)\dots p(x_i|x^{i-1})\dots p(x_n|x^{n-1})$$

$$\begin{aligned} Pr\left\{X^n = x^n, Y^n = y^n\right\} &= p(x_1)W_1(y_1|x_1)p(x_2|x_1, y_1)W(y_2|x_1, x_2, y_1)\dots \\ &= \prod_i p(x_i|\underbrace{x^{i-1}}_{\text{feedback}}\underbrace{y^{i-1}}_{\text{memory}})W_i(y_i|x^iy^{i-1}) \end{aligned}$$

Lemma 7.1. *if there is no feedback and the channel is memoryless and stationary, then*

$$Pr\left\{Y^n = y^n|X^n = x^n\right\} = \prod_{i=1}^{n} W(y_i|x_i)$$

Proof.

$$\begin{aligned} Pr\left\{Y^n = y^n, X^n = x^n\right\} &= \prod_{i=1}^{n} p(x_i|x^{i-1}y^{i-1})W_i(y_i|x^iy^{i-1}) \\ &= \prod_{i=1}^{n} p(x_i|x^{i-1})W(y_i|x_i) \\ &= \prod_{i=1}^{n} W(y_i|x_i)Pr\left\{X^n = x^n\right\} \end{aligned}$$

$\square$

Example 9. Suppose W is BSC(1/2) but we have feedback, defined by $X_1 = 0$ and $X_i = Y_{i-1}$.

$$Pr\left\{Y^2 = 00|X^2 = 01\right\} = 0$$
$$W(0|0)W(0|1) = \frac{1}{4}$$

Lemma 7.2. *if W is stationary memoryless and there is no feedback, then*

$$H(Y^n|X^n) = \sum_{i=1}^{n} H(Y_i|X_i)$$

Proof.

$$H(Y^n|X^n) = E\left[\log \frac{1}{Pr\left\{Y^n|X^n\right\}}\right] = E\left[\log \prod_{i=1}^{n} \frac{1}{Pr\left\{Y_i|X_i\right\}}\right] = \sum_{i=1}^{n} E\left[\log \frac{1}{Pr\left\{Y_i|X_i\right\}}\right] = \sum_{i=1}^{n} H(Y_i|X_i)$$

□

For a memoryless stationary channel $W(Y|X)$ we can compute, for any distribution $p(x)$, $p(x,y) = p(x)W(y|x)$ and $I(X;Y)$, we can also compute

$$C(W) = \max_{p(x)} I(X;Y)$$

Lemma 7.3. *for a stationary memoryless W without feedback, we have*

$$I(X^n;Y^n) \le nC(W)$$

Proof.

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &= H(Y^n) - \sum_i H(Y_i|X_i) \\ &\le \sum_i H(Y_i) - \sum_i H(Y_i|X_i) \\ &= \sum_i I(X_i;Y_i) \end{aligned}$$

Note that the joint distribution $Pr\left\{X_i, Y_i\right\}$ is of the form $p(x)W(y|x)$, then $I(X_i;Y_i) \le C(W)$ □

Notation 2. *for simplicity*

$$p * q = (1-q)p + q(1-p)$$

Example 10. Let W be a BSC(p), $Pr\left\{X = 0\right\} = 1 - q$ and $Pr\left\{X = 1\right\} = q$. Then

$$Pr\left\{Y = 0\right\} = (1-q)(1-p) + qp$$
$$Pr\left\{Y = 1\right\} = (1-q)p + q(1-p)$$

$$H(Y|X = 0) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$
$$H(Y|X = 1) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$
$$H(Y|X) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= (p*q)\log\frac{1}{p*q} + (1-(p*q))\log\frac{1}{1-(p*q)} - \left[p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}\right]
\end{aligned}
$$

We maximize $I(X;Y)$ for $q = 1/2$

$$C(W) = \log 2 - h_2(p)$$

Example 11. Let W be BEC(p) and $Pr\{X=1\} = q$

$$
\begin{aligned}
H(X) &= h_2(q) \\
H(X|Y=0) &= 0 \\
H(X|Y=1) &= 0 \\
H(X|Y=?) &= h_2(q)
\end{aligned}
$$

$$
\begin{aligned}
I(X;Y) &= h_2(q) - p h_2(q) = (1-p)h_2(q) \\
C(W) &= (1-p)\log 2
\end{aligned}
$$

7.1 Fano's inequality

Suppose U and V take values in the same alphabet $\mathcal{U}$, then

$$H(U|V) \leq p_e \log(|\mathcal{U}|-1) + h_2(p_e)$$

with

$$p_e = Pr\{U \neq V\} \quad \text{and} \quad h_2(p) = p\log(\frac{1}{p}) + (1-p)\log(\frac{1}{(1-p)})$$

Proof. Define

$$Z = \begin{cases} 1 & U \neq V \\ 0 & U = V \end{cases}, \quad H(Z) = h_2(p_e)$$

$$
\begin{aligned}
H(UZ|V) &= H(U|V) + H(Z|UV) \\
&= H(Z|V) + H(U|VZ) \\
&\leq H(Z) + H(U|VZ)
\end{aligned}
$$

but

$$H(U|VZ) = \underbrace{H(U|V,Z=0)}_{0} Pr\{Z=0\} + \underbrace{H(U|V,Z=1)}_{\leq \log(|\mathcal{U}|-1)} \underbrace{Pr\{Z=1\}}_{p_e}$$

□

So if $H(U|V) > \lambda \Rightarrow \exists f(\lambda) > 0$, $p_e > f(\lambda)$

Corollary 7.3.1. *Suppose U^L, V^L are random sequences with common alphabet $\mathcal{U}$, define :*

$$p_{e,i} = Pr\{U_i \neq V_i\}, \quad \bar{p_e} = \frac{1}{L}\sum_{i=1}^{L} p_{e,i}$$

then

$$\frac{1}{L}H(U^L|V^L) \leq h_2(\bar{p_e}) + \bar{p_e}\log(|\mathcal{U}|-1)$$

Proof.

$$\begin{aligned}
\frac{1}{L}H(U^L|V^L) &= \frac{1}{L}\sum_{i=1}^{L} H(U_i|U^{i-1}V^L) \\
&\leq \frac{1}{L}\sum_{i=1}^{L} H(U_i|V_i) \\
&\leq \frac{1}{L}\sum_{i=1}^{L}(p_{e,i}\log(|\mathcal{U}|-1)+h_2(p_{e,i})) \\
&= \bar{p_e}\log(|\mathcal{U}|-1)+\frac{1}{L}\sum_{i=1}^{L} h_2(p_{e,i}) \\
&\leq \bar{p_e}\log(|\mathcal{U}|-1)+h_2(\frac{1}{L}\sum_{i=1}^{L} p_{e,i}) \\
&= \bar{p_e}\log(|\mathcal{U}|-1)+h_2(\bar{p_e})
\end{aligned}$$

□

Theorem 7.4. *"Bad news" theorem, converse to the coding theorem*

- *Suppose we have a stationary source $U_1U_2...$ with entropy rate H and produces a letter every τ_s seconds.*
- *Suppose also that we have a channel W that accepts input $X_1X_2...$ once every τ_c seconds.*
- *Suppose also*

$$\frac{H}{\tau_s} > \frac{C(W)}{\tau_c}$$

then there is a $\lambda > 0$ such that $\bar{p_e} > \lambda$

Definition 7.4. *stable suppose the encoder works by taking blocks of L letters*

$$(U_1...U_L)(U_{L+1}...U_{2L})...$$

and outputs

$$(X_1...X_n)(X_{n+1}...U_{2n})...$$

then the encoder is stable if

$$L\tau_s \geq n\tau_c$$

Proof. Recall that for a stationary source $\frac{H(U_1...U_L)}{L}$ tends to H so

$$H(U_1...U_L) \geq LH$$

We also have

$$I(U^2;V^2) \leq nC(W)$$

therefore, since $\frac{n}{L} \leq \frac{\tau_s}{\tau_c}$

$$\begin{aligned}
H(U^2|V^2) = \frac{1}{L}(H(U^2)-I(U^2;V^2)) &\geq H-\frac{n}{L}C(W) \\
&\geq H-\frac{\tau_s}{\tau_c}C(W) \\
&= \tau_s(\frac{H}{\tau_s}-\frac{C(W)}{\tau_c})
\end{aligned}$$

The right hand side is

$$\epsilon(\tau_c, \tau_s, H, C) > 0$$

so for every stable encoder, decoder, we have

$$\bar{p_e} \log(|\mathcal{U}| - 1) + h_2(\bar{p_e}) > \epsilon(\tau_s, \tau_c, H, C)$$

then

$$\bar{p_e} \geq \epsilon(\tau_s, \tau_c, H, C, |\mathcal{U}|)$$

□

Example 12. Suppose $\mathcal{U} = \{0, 1\}$ and $U_1U_2...$ is a Markov process with

$$U_1 = \begin{cases} 0 & \text{with } p = 0.5 \\ 1 & \text{with } p = 0.5 \end{cases}, \quad p(U_{n+1}|U_n) = \begin{cases} 1-p & u_{n+1} = u_n \\ p & u_{n+1} \neq u_n \end{cases},$$

$$\begin{aligned} H &= \lim_{n\to\infty} H(U_n|U^{n-1}) \\ &= \lim_{n\to\infty} H(U_n|U_{n_1}) \\ &= H(U_2|U_1) = h_2(p) \end{aligned}$$

suppose $w = BEC(q), c(w) = (1-q)\log(2)$ and $\tau_s = \tau_c = 1$

$$h_2(\bar{p_e}) \geq h_2(p) - (1-q)\log(2) \Rightarrow \bar{p_e} \geq \lambda$$

What we want to do next is to show a matching "Good news" theorem:

We could show that if $\frac{H}{\tau_s} \leq \frac{c(w)}{\tau_c}$ then for any $\lambda > 0$, we can find a stable encoder and decoder such that $p_e < \lambda$. Instead, we will show stronger results:

1. **Separation theorem** The encoder can be designed in a modular way:
   - A **source encoder** which encoder message words in bits. The design of this encoder is strongly dependent of the type of the input.
   - A **channel encoder** which encoder the bits to maximize the performance with a specific channel.
2. We will show that

   $$Pr\left\{U^L \neq V^L\right\} < \lambda$$

   using the fact that

   $$(U_i \neq V_i) \Rightarrow (U^L \neq V^L) \quad \text{so} \quad p_{e,i} \leq Pr\left\{U^L \neq V^L\right\} \Rightarrow \bar{p_e} \leq Pr\left\{U^L \neq V^L\right\}$$

We will now show that good channel encoders and channel decoders exist

Definition 7.5. *Given a channel W with input alphabet $\mathcal{X}$, a block encoder is a function*

$$Enc : \{1, ..., M\} \to \mathcal{X}^n$$

with n the block length.
$Enc(1), ..., Enc(M)$ are each called codewords and M is equal to the number of codewords.
The rate of the code can be defined by

$$R = \frac{\log M}{n}$$

Definition 7.6. *Given a channel W with ouput alphabet Y, a block decoder is a function*

$$Dec : \mathcal{Y}^n \to \{?, 1, ..., M\}$$

Definition 7.7.

$$p_{error}(m) = Pr\left\{\hat{m} \neq m | m\right\}$$

$$\bar{p}_{error}(m) = \frac{1}{M}\sum_{m=1}^{M} p_{error}(m)$$

$$\hat{p}_{error}(m) = \max_{m} p_{error}(m)$$

7.2 Computational consideration for $C(W)$

We have an optimization problem

$$\max_{p_X} f(p_X) \quad \text{where} \quad f(p_X) = I(X;Y)$$

See section C for further information on convex optimization.

Claim 7.1. *f is a concave function*

We want to compute

$$\frac{\partial I(X;Y)}{\partial p(x)}$$

We have

$$I(X;Y) = \sum_{x,y} p(x)W(y|x)\log\frac{W(y|x)}{p_Y(y)}$$

$$p_Y(y) = \sum_{x} p(x)W(y|x)$$

$$\begin{aligned}
\frac{\partial I}{\partial p(x_0)} &= \sum_{x,y}\frac{\partial}{\partial p(x_0)}\left\{p(x)W(y|x)\log\frac{W(y|x)}{p_Y(y)}\right\} \\
&= \sum_{x,y}\left\{I_{x=x_0}W(y|x)\log\frac{W(y|x)}{P_Y(y)} - p(x)W(y|x)\frac{W(y|x_0)}{p_Y(y)}\log e\right\} \\
&= \sum_{y} W(y|x_0)\log\frac{W(y|x_0)}{p_Y(y)} - \sum_{y} p_Y(y)\frac{W(y|x_0)}{p_Y(y)}\log e \\
&= \sum_{y} W(y|x_0)\log\frac{W(y|x)}{P_Y(y)} - \log e
\end{aligned}$$

Theorem 7.5. *p_X maximizes $I(X;Y)$ iff there exists λ such that for all x*

$$\sum_{y} W(y|x)\log\frac{W(y|x)}{P_Y(y)} \leq \lambda$$

with equality when $p_X(x) = 0$. Furthermore $\lambda = C(W)$.

Proof. We only need to prove the furthermore part. Observe that for all x

$$p_X(x)\sum_y W(y|x)\log\frac{W(y|x)}{P_Y(y)} = p_X(x)\lambda$$

and then

$$\sum_x p_X(x)\sum_y W(y|x)\log\frac{W(y|x)}{P_Y(y)} = \sum_x p_X(x)\lambda$$

$\square$

Example 13 (Z channel)**.** W is a normal binary channel that maps a 1 input to a 0 output with probability ϵ. Applying theorem 7.5 with $x = 0$ and $x = 1$:

$$\begin{aligned}
& W(0|0)\log\frac{W(0|0)}{p_Y(0)} = W(0|1)\log\frac{W(0|1)}{p_Y(1)} + W(1|1)\log\frac{W(1|1)}{p_Y(1)} \\
\iff & \log\frac{1}{p_Y(y)} = \epsilon\log\frac{\epsilon}{p_Y(0)} + (1-\epsilon)\log\frac{1-\epsilon}{p_Y(1)} = h_2(\epsilon) + \epsilon\log\frac{1}{p_Y(0)} + (1-\epsilon)\log\frac{1}{p_Y(1)} \\
\iff & \log\frac{p_Y(1)}{p_Y(0)} = -\frac{h_2(\epsilon)}{1-\epsilon} \triangleq -\alpha \\
\implies & p_Y(1) = \frac{2^{-\alpha}}{1+2^{-\alpha}} \text{ and } p_Y(0) = \frac{1}{1+2^{-\alpha}}
\end{aligned}$$

$$C(W) = \log(1 + 2^{-\alpha})$$

Lemma 7.6. *For any circle with red segments of cumulative length strictly less than 1/4, there exists a square whose all corners are on the circle but not on the red segments.*

Proof. By random construction. Place the first corner of the square uniformly at random on the circle (also makes the 3 other uniform).

$$\begin{aligned}
Pr\{\text{1st corner lands on red}\} &< \frac{1}{4} \\
Pr\{i\text{th corner lands on red}\} &< \frac{1}{4} \\
Pr\left\{\bigcup_{i=1} i\text{th corner lands on red}\right\} &< 1 \\
Pr\{\text{none of the corners land on red}\} &> 0
\end{aligned}$$

$\square$

Theorem 7.7 (Channel coding - good news)**.** *Given a channel W (discrete, memoryless, stationary), a rate $R < C(W)$ and $\epsilon > 0$, there exists a n large enough and encoding/decoding functions $Enc : \{1 \dots M\} \to \mathcal{X}^n$ with $M \geq 2^{nR}$ and $Dec : \mathcal{Y}^n \to \{1 \dots M\}$ such that for all $m \in \{1 \dots M\}$*

$$Pr\{Dec(Y^n) \neq m | X^n = Enc(m)\} < \epsilon$$

In other words we can communicate reliably at rate greater or equal to R on channel W.

Proof. Given W and $R < C(W)$, fix a p_X such that $I(X;Y) > R$. Pick $\delta > 0$, n large enough (to be determined later) and set $M' = \lceil 2 \cdot 2^{nR}\rceil$. Define the encoding function

$$\begin{aligned}
Enc(1) =& \ X(1)_1 \dots X(1)_n \\
\dots =& \ \dots \\
Enc(M') =& \ X(M')_1 \dots X(M')_n
\end{aligned}$$

choosing $\{X(m)_i : 1 \leq i \leq n, 1 \leq m \leq M'\}$ i.i.d. $\sim p_X$.

For the decoder fix

$$T(n, \delta, p_{XY}) = \left\{ (x^n, y^n) : (1-\delta) p_{XY}(x,y) \leq \frac{\#\{(x_i, y_i) = (x,y)\}}{n} \leq (1+\delta) p_{XY}(x,y) \right\}$$

$Dec(y^n)$: check for each m if $(Enc(m), y^n) \in T(n, \delta, p_{XY})$, if there is only a single m for which the pair is in the typical set then $Dec(y^n) = m$ otherwise (if there is none or more than one) $Dec(y^n) = 0$.

We now compute the probability of error $p_{e,m} \triangleq Pr\left\{Dec(Y^n) \neq m | X^n = Enc(m)\right\}$. $p_{e,m}$ depends on the choice of $Enc(1) \dots Enc(M)$ and since $Enc(1) \dots Enc(M)$ are randomly chosen, $p_{e,m}$ is a random variable. Supposing m is sent, an error will happen if and only if $(Enc(m), y^n) \notin T$ or for some $m' \neq m:\ (Enc(m'), y^n) \in T$

$$\begin{aligned}
E\left[p_{e,m}\right] &= E_{Enc}[E_y[I\left\{\text{error has happened} \mid m \text{ is sent}\right\}]] \\
&= E_{Enc}[I\{(Enc(m), y) \notin T, \exists m' \neq m (Enc(m', Y^n) \in T\} | \; m \text{ is sent}] \\
&\leq E[I\{(Enc(m), Y^n) \notin T\}] + \sum_{m' \neq m} I\{(Enc(m'), Y^n) \in T\} | \; m \text{ is sent}] \\
&= Pr\left\{(Enc(m), Y^n) \notin T | m \text{ is sent}\right\} + \sum_{m' \neq m} Pr\left\{(Enc(m'), Y^n) \in T | m \text{ is sent}\right\}
\end{aligned}$$

We have

$$\begin{aligned}
Pr\left\{Enc(m) = x_1 \dots x_n, Y^n = y_1 \dots y_n | \; m \text{ is sent}\right\} &= p_x(x_1) p_x(x_2) \dots p_x(x_n) W(y_1|x_1) W(y_2|x_2) \dots W(y_n|x_n) \\
&= p_X(x_1) p_X(x_2) \dots p_X(x_n) p_Y(y_1) p_Y(y_2) \dots p_Y(y_n)
\end{aligned}$$

and as n gets large

$$Pr\left\{(Enc(m), Y^n) \notin T(p_{XY}, n, \delta)\right\} = Pr\left\{\text{iid sequence} \sim p_{XY} \notin T(p_{XY}, n, \delta)\right\} \to 0$$

because $(Enc(m), Y^n)$ is iid $\sim p_{XY}$. Recall from typicallity that if U^n is iid p_U, then

$$\lim_{n \to \infty} Pr\left\{U^n \notin T(n, p_U, \delta)\right\} = 0$$

and if U^n is in reality iid $\sim\ q_U$

$$Pr\left\{U^n \in T(n, p, \delta)\right\} \leq 2^{-n[D(p||q) - o(\delta)]}$$

Then,

$$\begin{aligned}
Pr\left\{(Enc(m), Enc(m'), Y^n) = (x^n, (x')^n, y^n)\right\} &= p_X(x^n) p_X((x')^n) W(y^n | x^n) \\
Pr\left\{(Enc(m), y^n) = (x^n, y^n)\right\} &= p_X(x^n) W(y^n | x^n) \\
Pr\left\{(Enc(m'), y^n) = ((x')^n, y^n)\right\} &= p_X((x')^n) \underbrace{\sum_{x^n} p(x^n) W(y^n | x^n)}_{p_Y(y)}
\end{aligned}$$

$$\begin{aligned}
&\iff\ (Enc(m'), y') \text{ is iid } \sim q_{XY} = p_X p_Y \\
&\Rightarrow Pr\left\{(Enc(m'), y^n) \in T(p_{XY}, n, \delta)\right\} \leq 2^{-n[D(p||q) - o(\delta)]}
\end{aligned}$$

Also

$$D(p||q) = \sum_{xy} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} = I(X;Y)$$

Remember $M' = \lceil 2 \cdot 2^{nR} \rceil \leq 2 \cdot 2^{nR} + 1$ then $M' - 1 \leq 2 \cdot 2^{nR}$

$$E\left[p_{e,m}\right] \leq o_n(1) + (M'-1)2^{-n[I(X;Y)-o(\delta)]} \leq o_n(1) + 2 \cdot 2^{-n(I(X;Y)-R-o(\delta))}$$

We choose δ small enough to have a negative exponent. Then it will go to 0 as n gets large. So we have shown that for n large enough we can make for every m:

$$\begin{aligned} & E\left[p_{e,m}\right] < \frac{\epsilon}{2} \\ \Rightarrow & E\left[\sum_{m=1}^{M'} p_{e,m}\right] \leq \frac{M'}{2}\epsilon \\ \Rightarrow & \exists \text{an encoder such that } \sum_{m=1}^{M'} p_{e,m} \leq \frac{M'}{2}\epsilon \end{aligned}$$

How many terms in the sumation can be greater or equal to ϵ ? At most $M'/2$, so remaining must be stricly smaller than ϵ but

$$M' - \frac{1}{2}M' = \frac{1}{2}\lceil 2 \cdot 2^{nR} \rceil \geq \frac{1}{2}2 \cdot 2^{nR} = 2^{nR}$$

We throw away the one smaller than ϵ and we have a code with rate greater than R for

$$\max_m p_{e,m} < \epsilon$$

□

Example 14. Suppose $\mathcal{X} = \{a, b, c\}$, $C(W) = 1.3$ and $R = 1.25$, then $Enc(1 \dots 32) \to \mathcal{X}^4$ is a valid encoding function for this channel, while $Enc(1 \dots 32) \to \mathcal{X}^5$ would not allow reliable transmission.

Example 15. Suppose we want to design a code with $n = 1000$, $R = \frac{1}{2}$. The encoding table will have 1000×2^{500} elements, more than 10^{153} elements. "C'est impossible M'sieur !"

To illustrate the proof technique that we used to prove the coding theorem, we take an example.

Example 16. Assume W is a BEC channel (probability p to have an erasure symbol ?).

$$C(W) = 1 - p = \max_{p_x} I(X;Y)$$

achieved when $p_X(0) = p_X(1) = \frac{1}{2}$. Our coding theorem says that when $R < 1-p$, $\epsilon > 0$ we can find a code of rate R with error probability $< \epsilon$. In the proof of the theorem 7.7, we generate a $n \times M$ coding matrix with n large and $M = 2^{nR}$ according to p_X defining $C(W)$.

To send a nR-bit message $m \in \{1 \dots M\}$, we send the mth row of the table over the chanel. When we receive $y = (y_1 \dots\ y_n)$, we compare y to each row of the table and check the tipicality. In our case

$$\begin{array}{lll} \frac{1}{n}\{\#\text{ of }(0,0)\} \approx \frac{1-p}{2} & \frac{1}{n}\{\#\text{ of }(0,1)\} = 0 & \frac{1}{n}\{\#\text{ of }(0,?)\} \approx \frac{p}{2} \\ \frac{1}{n}\{\#\text{ of }(1,0)\} = 0 & \frac{1}{n}\{\#\text{ of }(1,1)\} \approx \frac{1-p}{2} & \frac{1}{n}\{\#\text{ of }(1,?)\} \approx \frac{p}{2} \end{array}$$

If there is exactly one row (i.e. $\hat{m}$) return $\hat{m}$, otherwise return 0.

- The correct codeword will pass the test with high probability, thanks to law of large numbers,
- What about an incorrect codeword ?

Recall the definition of typicality (definition 5.1) and suppose

$$y = \underbrace{0\dots0}_{n\frac{1-p}{2}}\underbrace{1\dots1}_{n\frac{1-p}{2}}\underbrace{?\dots?}_{np}$$

$m' = x_1x_2\dots x_n$ will be typical only if it is of the type

$$\underbrace{0\dots0}_{n\frac{1-p}{2}}\underbrace{1\dots1}_{n\frac{1-p}{2}}\underbrace{?\dots?}_{np}$$

$$Pr\left\{\begin{pmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{pmatrix} \text{ is typical}\right\} \leq \left(\frac{1}{2}\right)^{n(1-p)} = 2^{-n(1-p)}$$

Then, using that an upperbound to the number of incorrect codewords is 2^{nR},

$$Pr\{error\} < 2^{-n(1-p)}2^{nR} + Pr\{\text{correct } w \text{ fails the test}\}$$

and because $R < 1-p$

$$\lim_{n\to\infty} Pr\{error\} = 0$$

8 Differential entropy

Definition 8.1 (Differential entropy)**.** *Let X be a real valued random variable with probability density function $f(x)$ such that*

$$Pr\{x \leq X \leq x+\delta\} \approx \delta f(x)$$

The differential entropy of X is

$$h(X) \triangleq \int f(x)\log\frac{1}{f(x)}dx$$

Example 17. Uniform random variable in $[0,a]$ then

$$h(A) = loga = \begin{cases} <0 & \text{if } a<1 \\ 0 & \text{if } a=1 \\ >0 & \text{if } a>1 \end{cases}$$

Lemma 8.1. *Suppose $Y = X + a$, a is a constante then $h(Y) = h(X)$*

Proof. We have $f_Y(y) = f_X(y-a)$, then

$$h(Y) = \int f_X(y-a)\log\frac{1}{f_X(y-a)}dy = \int f_X(x)\log\frac{1}{f_X(x)}dx = h(X)$$

□

Lemma 8.2. *Suppose $Y = aX$, then $h(Y) = h(X) + \log|a|$*

Proof. Suppose $a > 0$,

$$f_Y(y) = Pr\{y \leq Y \leq y+\delta\} = Pr\left\{\frac{y}{a} \leq X < \frac{y}{a} + \frac{\delta}{a}\right\} \approx \frac{1}{a}f_X\left(\frac{y}{a}\right)$$

$$\log\frac{1}{f_Y(y)} = \log a + \log\frac{1}{f_X\left(\frac{y}{a}\right)}$$

$$h(Y) = \int f_Y(y)\log\frac{1}{f_Y(y)}dy = \log a + \int f_X\left(\frac{y}{a}\right)\left(\log\frac{1}{f_X\left(\frac{y}{a}\right)}\right)\frac{1}{a}dy = \log a + \underbrace{\int f_X(x)\log\frac{1}{f_X(x)}dx}_{h(X)}$$

□

Example 18. Suppose Y is a gaussian with mean μ and variance σ^2 then $Y = \sigma X + \mu$ where X is $N(0,1)$

$$h(Y) = h(\sigma X) = \log \sigma + h(X)$$

$$h(X) = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left[\log \sqrt{2\pi} + \frac{1}{2} x^2 \log e \right] dx \overset{(a)}{=} \frac{1}{2} \log 2\pi + \frac{1}{2} \log e = \frac{1}{2} \log 2\pi e$$

Where the second term of (a) follows from $E\left[X^2\right] = 1$.

Lemma 8.3. *Suppose X is a real value random variable with differentiable entropy $h(X)$. Consider a $\delta > 0$ and X_δ, the quantization of X in interval of width δ*

$$X_\delta = \delta \left\lfloor \frac{X}{\delta} \right\rfloor = n\delta \text{ if } n\delta \le X \le (n+1)\delta$$

then

$$\lim_{\delta \to 0} H(X_\delta) + \log \delta = h(X)$$

Proof.

$$\begin{aligned} H(X_\delta) &= \sum_n Pr\left\{X_\delta = n\delta\right\} \log \frac{1}{Pr\left\{X_\delta = n\delta\right\}} \\ &\approx \sum_n \delta f_X(n\delta) \log \frac{1}{\delta f_X(n\delta)} \\ &= \log \frac{1}{\delta} + \sum_n \left(f_X(n\delta) \log \frac{1}{f_X(n\delta)} \right) \delta \\ &\overset{(a)}{=} \log \frac{1}{\delta} + \int f(x) \log \frac{1}{f(x)} dx \end{aligned}$$

We recognize a Riemann sum for equality (a). □

Suppose $X_1 \ldots X_n$ are $\mathbb{R}$-valued RV's ($X^n \in \mathbb{R}^n$), we define

$$h(X^n) = h(X_1 \ldots X_n) = \underbrace{\int \int}_{\mathbb{R}^n} f_{X^n}(x_1 \ldots x_n) \log\left(\frac{1}{f_{X^n}(x_1 \ldots x_n)}\right) dx_1 \ldots dx_n$$

$$h(X|Y) = \int \int f_{XY}(x,y) \log\left(\frac{1}{f_{XY}(X|Y)}\right) dx dy = \mathbb{E} \log\left(\frac{1}{f_X Y(X|Y)}\right)$$

Theorem 8.4.

$$h(X^n) = \sum_{i=1}^{n} h(X_i | X^{i-1})$$

Proof.

$$f_{X^n}(x_1 \ldots x_n) = f_{x_1}(x_1) f_{x_2|x_1}(x_2|x_1) \ldots f_{x_n|x^{n-1}}(x_n|x^{n-1})$$

take log's, take expectation □

Definition 8.2. *Given two densities $f(x), g(x)$, let*

$$D(f||g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

Lemma 8.5.

$$D(f||g) \ge 0$$

with equality iff $f = g$

Proof. use $\ln(z) \leq z - 1$ to show that $-D(f||g) \leq 0$ $\square$

Definition 8.3. *For X, Y $\mathbb{R}$-valued RV's, define*

$$
\begin{aligned}
I(x;Y) &= \int f_{XY}(x,y)\log(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}) \\
&= D(f_{XY}||f_X(x)f_Y(y)) \\
&= h(X) + h(Y) - h(XY) \\
&= h(X) - h(X|Y) \\
&= h(Y) - h(Y|X)
\end{aligned}
$$

Proposition 8.1.

$$I(X;Y) \geq 0 \quad (= 0 \text{ iff } X \perp Y)$$

Proof.

$$I(X;Y) = D(f_{XY}||f_x f_y)$$

$\square$

Equivalently: $h(X|Y) \leq h(x)$, with equallity iff X and Y independent.

Lemma 8.6. *Given X, Y, $\mathbb{R}$-valued with joint pdf f_{XY}, for $\delta > 0$, define X_δ, Y_δ as $X_\delta = \delta\lfloor \frac{X}{\delta} \rfloor, Y_\delta = \delta\lfloor \frac{Y}{\delta} \rfloor$, then*

$$\underbrace{I(X_\delta;Y_\delta)}_{\text{discrete } I} \to I(X;Y) \text{ as } \delta \to 0$$

Proof. observe

$$
\begin{aligned}
Pr\{X_\delta = n\delta\} &= Pr\{X \in [n\delta,(n+1)\delta]\} \equiv \delta f_X(n\delta) \\
Pr\{Y_\delta = m\delta\} &= Pr\{Y \in [m\delta,(m+1)\delta]\} \equiv \delta f_Y(m\delta) \\
Pr\{X_\delta = n\delta, Y_\delta = m\delta\} &= Pr\{X \in [n\delta,(n+1)\delta], Y \in [m\delta,(m+1)\delta]\} \equiv \delta^2 f_{XY}(n\delta, m\delta)
\end{aligned}
$$

$$
\begin{aligned}
I(X_\delta;Y_\delta) &= \sum_{n,m} Pr\{X_\delta = n\delta, Y_\delta = m\delta\}\log(\frac{Pr\{X_\delta = n\delta, Y_\delta = m\delta\}}{Pr\{X_\delta = n\delta\}Pr\{Y_\delta = m\delta\}}) \\
&\equiv \sum_{n,m} \delta^2 f_{XY}(n\delta,m\delta)\log(\frac{\delta^2 f_{XY}(n\delta,m\delta)}{\delta f_X(n\delta)\delta f_Y(m\delta)}) \\
&\equiv \sum_{n,m} \delta^2 f_{XY}(n\delta,m\delta)\log(\frac{f_{XY}(n\delta,m\delta)}{f_X(n\delta)f_Y(m\delta)}) \\
&= \text{Riemann sum for} \quad \int\int f_{XY}(x,y)\log\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}dxdy = I(X;Y)
\end{aligned}
$$

$\square$

In general, define for $X^n \in \mathbb{R}^n, Y^m \in \mathbb{R}^m, Z^k \in \mathbb{R}^k$

$$I(X^n;Y^m|Z^k) = \underbrace{\int ... \int}_{n+m+k} f_{X^nY^mZ^k}(x^n,y^m,z^k)\log(\frac{f_{XY|Z}(x^ny^m|z^k)}{f_{X|Z}(x^n|z^k)f_{Y|Z}(Y^m|z^k)}$$

we then have

Theorem 8.7. *Chain Rule for I*

$$I(X^n;Y) = \sum_{i=1}^{n} I(X_i;Y|X^{i-1})$$

Proof. Same proof as in the discrete case $\square$

Example 19. X^n is a Gaussian Random Variable with $\mathbb{E}\, x^N = \bar{\mu}$ and variance matrix $K, K_{i,j} = \mathbb{E}\,(X_i - \mu_i)(X_j - \mu_j)$

$$h(X^n) = \underbrace{h(X^n - \bar{\mu})}_{\text{Gaussian with zero-mean with covariance} K}$$

consequently we may assume that $\bar{\mu} = \bar{0}$, recall that the joint pdf of a zero-mean gaussian is given by

$$f(\bar{x}) = \underbrace{\frac{1}{det(2\pi K)^{1/2}}}_{(2\pi)^{n/2}(det(k)^{1/2})} e^{0.5(X^T K^{-1} X)}$$

$$\log(\frac{1}{f(\bar{x})}) = \frac{1}{2}\log(det(2\pi K)) + \frac{1}{2} X^T K^{-1} X(\log(e))$$

$$\begin{aligned} h(X) &= \mathbb{E}\,\frac{1}{2}\log(det(2\pi K)) + \frac{\log(e)}{2} X^T K^{-1} X \\ &= \frac{1}{2}\log(det(2\pi K)) + \underbrace{\frac{\log(e)}{2} \underbrace{\mathbb{E}\, X^T K^{-1} X}_{tr(K^{-1}K) = tr(I_n) = n}}_{\frac{1}{2}\log(e^n)} \\ &= \frac{1}{2}\log(det(2\pi e K)) \end{aligned}$$

Side knowledge:

$$\mathbb{E}\, X^T A X = \mathbb{E} \sum_{i,j} X_i A_{ij} X_j = \sum_{i,j} A_{ij}\, \mathbb{E}\, X_i X_j = \sum_{i,j} A_{ij} K_{ij} = \sum_i (\sum_j A_{ij} K_{ij}) = tr(AK)$$

Theorem 8.8. *Suppose $X \in \mathbb{R}^n$ is a random vector with*

$$\mathbb{E}\, X_i X_j = k_{ij}$$

Then

$$h(X) \leq \frac{1}{2}\log(det(2\pi e K))$$

(Gaussians have maximum entropy among Random vectors with a given 2nd moment)

Proof. Let f be the density of X^n, let g be the gaussian density

$$g(x) = \frac{1}{det(2\pi K)^{1/2}} e^{1\frac{1}{2}(X^T K^{-1} X)}$$

observe that $\log(\frac{1}{g(x)}) = \frac{1}{2}\log(det(2\pi K) + \frac{1}{2}\log(2) X^T K_{-1} X$ so

$$\int f(\bar{x}) \log(\frac{1}{g(x)}) d\bar{x} = \frac{1}{2}\log(det(2\pi K)) + \frac{\log(e)}{2} \underbrace{E X^T K^{-1} X}_{n} = \int g(x) \log(\frac{1}{g(x)}) dx$$

how

$$0 \geq \int f(x) \log(\frac{g(x)}{f(x)}) dx = -\{\frac{1}{2}\log(det(2\pi e K))\} + h(x)$$

$\square$

Another example of maximum entropy

Suppose we know that $X \in [a, b]$ with probability 1. then $h(X) \leq \log(b - a)$ equality uff X is uniform on $[a, b]$.

Proof. Let

$$g(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{else} \end{cases}$$

$$\int f(x) \log(\frac{1}{g(x)}) dx = \log(b-a) = \int g(x) \log(\frac{1}{g(x)}) dx$$

$$0 \ge \int f(x) \log(\frac{g(x)}{f(x)} dx) = -\log(b-a) + h(X)$$

$\square$

Appendices

A Markov chains

$U_1 - U_2 - \cdots - U_n$ forms a Markov chain if the joint probability distribution of the RVs is

$$p(a,b,c,d) = p(a)p(b|a)p(c|b)p(d|c)$$

which is equivalent to $(U_1, \ldots, U_{k-1})$ are independant of $(U_{k+1}, \ldots, U_n)$ when conditionned on U_k for any k.

Theorem A.1. *The reverse of a MC is a MC*

B Stochastic processes

A stochastic process is a collection $U_1, U_2 \ldots U_n$ of RVs each taking values in $\mathcal{U}$. It is described by its joint probability

$$p(u^n) = P(U_1 \ldots U_n = u_1 \ldots u_n) = P(U^n = u^n)$$

Definition B.1 (Stationary stochastic process)**.** *A process $U_1, U_2, \ldots$ is called stationary if for every n and k and $u_1 \ldots u_n$, we have*

$$p(u^n) = p(U_1 \ldots U_n = u_1 \ldots u_n) = p(U_{1+k} \ldots U_{n+k} = u_1 \ldots u_n)$$

In other words, the process is time shift invariant.

C Concave/convex functions

A function $f : S \to \mathbb{R}$ is called convex if

$$\forall x, y \in S, 0 \leq \lambda \leq 1, f(\lambda x - (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

where S is a convex set.

Definition C.1. *A set $S \subseteq \mathbb{R}^k$ is called to be convex if*

$$\forall x, y \in S, 0 \leq \lambda \leq 1, \lambda x + (1-\lambda)y \in S$$

Definition C.2. *f is called concave if $-f$ is convex.*

Definition C.3. *k-simplex*

$$S_k = \{(p_1, ..., p_k) \in \mathbb{R}^k, p \geq 0, \sum_i p_i = 1\}$$

as the k-simplex (a $(k-1)$-dimentional subset of $\mathbb{R}^k$)

Remark: Given S_k a convex set and $p, q \in S_k$, let

$$r = \lambda p + (1-\lambda) q$$
$$r_i = \lambda p_i + (1-\lambda) q_i \geq 0$$

$$\sum r_i = \lambda + (1-\lambda) = 1$$

Example 20. Let $f : S_k \to \mathbb{R}$, with

$$f(p_1, ..., p_k) = \sum^k p_i \log_{\frac{1}{p_i}}$$

claim: f is concave

Proof. Given $p, q \in S_k, 0 \leq \lambda \leq 1$, define (U, V) with $U \in \{0, 1\}$ and $V \in \{1, ..., k\}$

$$P_{UV}(u, v) = \begin{cases} \lambda p_i, & u = 0, v = i \\ (1-\lambda) q_i, & u = 1, v = i \end{cases}$$

therefore we have

$$\begin{aligned} Pr\{V = i\} &= \lambda p_i + (1-\lambda) q_i \\ H(V) &= f(\lambda p + (1-\lambda) q) \\ H(V|U) &= \lambda f(p) + (1-\lambda) f(q) \end{aligned}$$

$\square$

Example 21. For $W(Y|X)$ let $f(p_X) = I(X;Y)$ when $p(x, y) = p_X(x) W(Y|X)$ **Claim:** f is concave,

$$I(X;Y) = H(Y) - H(X|Y)$$

and

$$H(Y|X) = \sum_x p_X(x) \sum_y W(Y|X) \log \frac{1}{W(Y|X)}$$

We see that $H(Y|X)$ is a linear function of $p_X(x)$.
$H(Y)$ is a concave function of $p_Y(y)$ with

$$p_Y(y) - \sum_x p_X(x) W(Y|X)$$

$$p_X \underbrace{\longrightarrow}_{\text{linear}} p_Y \underbrace{\longrightarrow}_{\text{concave}} H(Y) \Longrightarrow p_X \underbrace{\longrightarrow}_{\text{concave}} H(Y)$$

How to maximize a function on the simplex?

Theorem C.1. *Karush-Kuhn-Tucker conditions - (KKT)*
Suppose $f : S_k \to \mathbb{R}$, smooth ($\frac{df}{dp_i dp_j}$ exists), then if $p = \{p_1, ..., p_k\}$ maximizes f, then $\exists \lambda$ s.t.

$$\forall i, \frac{df}{dp_i} \leq \lambda$$

with equality $\forall i$ for which $p_i > 0$

Proof. Suppose $(p_1, ..., p_k)$ maximizes f, then suppose that $p_i > 0$. Then we can consider a $p' \in S_k$ as follow:
Pick $j \neq i$ and a small $\epsilon, 0 < \epsilon < p_i$

$$p'_k = \begin{cases} p_i - \epsilon, & k = i \\ p_j + \epsilon, & k = j \\ p_k, & \text{else} \end{cases}$$

$$\begin{aligned} f(p') &= f(p) + \frac{df(p)}{dp_i}(-\epsilon) + \frac{df(p)}{dp_j}(\epsilon) + O(\epsilon^2) \\ &= f(p) + \epsilon \left[\frac{df}{dp_j} - \frac{df}{dp_i} \right] + O(\epsilon^2) \end{aligned}$$

So for every i, j with $p_i > 0$ we have

$$\frac{df}{dp_j} \frac{df}{dp_i}$$

$\Rightarrow$ equality if i and j are such that $p_i > 0, p_j > 0$
$\Rightarrow$ for i's such that $p_i > 0$, $\frac{df}{dp_i} = \lambda$ and all the indices j have $\frac{df}{dp_j} \leq \lambda$ $\square$

Theorem C.2. *Suppose $f : S_k \to \mathbb{R}$, suppose f is concave and suppose for $p \in S_k$, the KKT condition hold. Then $\forall q \in S_k, f(q) \leq f(p)$*

Proof.

$$\begin{aligned} f(\epsilon q + (1-\epsilon)p) &\geq (1-\epsilon)f(p) + \epsilon f(q) \\ \frac{f(\epsilon q + (1-\epsilon)p) - f(p)}{\lambda} &\geq f(q) - f(p), \quad \forall 0 < \epsilon \leq 1 \end{aligned}$$

$$\Rightarrow f(q) - f(p) \leq \lim_{\epsilon \to 0} \frac{f(p + \epsilon(q-p)) - f(p)}{\epsilon}$$

$$\begin{aligned} f(p + \epsilon(q-p)) &= f(p) + \sum \epsilon(q_i - p_i)\frac{df(p)}{dp_i} + O(\epsilon^2) \\ \frac{f(p + \epsilon(q-p)) - f(p)}{\epsilon} &= \sum_i (q_i - p_i)\frac{df(p)}{dp_i} + O(\epsilon) \end{aligned}$$

So

$$\lim_{\epsilon \to 0} \frac{f(p + \epsilon(q-p)) - f(p)}{\epsilon} = \sum_i (q_i - p_i)\frac{df(p)}{dp_i}$$

with

$$(q_i - p_i)\frac{df}{dp_i} = \begin{cases} \lambda(q_i - p_i), & p_i > 0 \\ \underbrace{(q_i - p_i)}_{\geq 0}\underbrace{\frac{df}{dp_i}}_{\leq \lambda}, & p_i = 0 \end{cases} \leq \lambda(q_i - p_i)$$

$$\Rightarrow f(q) - f(p) \leq \lim_{\epsilon \to 0}[...] \leq 0$$

$\square$

Example 22. Supppose $f(p_1, p_2, p_3) = p_1 p_2^2 p_3^3$. We want to maximize it. If it isn't concave, we know that $\log(f(..))$ is concave. A try with KKT:

$$\frac{df}{dp_1} = \frac{1}{p_1}, \frac{df}{dp_2} = \frac{2}{p_2}, \frac{df}{dp_3} = \frac{3}{p_3}$$

setting then all λ yeild

$$(p_1, p_2, p_3) = \lambda(1, 2, 3) = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$$

Example 23.

$$f(p_1, p_2, p_3) = (1 + p_1)p_2 p_3$$

maximize f on the simplex by considering

$$\log(f) = \log(1 + p_1) + \log(p_2) + \log(p_3)$$

therefore:

$$\frac{df}{dp_1} = \frac{1}{1 + p_1}, \frac{df}{dp_2} = \frac{1}{p_2}, \frac{df}{dp_3} = \frac{1}{p_3}$$

suggest $p = (0, 0.5, 0.5)$ the $\frac{df}{dp} = (1, 2, 2) \to$ satisfy KKT with $\lambda = 2$