# Information Theory and Coding - Prof. Emere Telatar

Jean-Baptiste Cordonnier, Sebastien Speierer, Thomas Batschelet

October 17, 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* ***uniquily 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*

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

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.

**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

$$\begin{aligned} &\Rightarrow kraftsum(\mathscr{C}^n) \leq 1 + \max_{u_1,\dots,u_n} l(\mathscr{C}^n) \\ &\Rightarrow kraftsum(\mathscr{C})^n \leq 1 + nL, \quad L = \max l(\mathscr{C}(n)) \end{aligned}$$

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 $\sum_u i^{\mathscr{C}(u)} \leq 1$, then, there exist a prefix-free code $\mathscr{C} : \mathcal{U} \to \{0,1\}$ s.t. $\forall l(\mathscr{C}(u)) = \mathscr{C}(u)$*

*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 u.d., 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 $\mathcal{U}$ 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

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

**Theorem 2.1.** *if $\mathscr{C}$ is u.d., 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 u.d.

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

□

**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

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

□

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

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

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

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 $p(u,v) = P(U = u, V = v)$, 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 $\ln z \leq z - 1 \ \forall z$ (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 $U$s 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)**.**

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

**Theorem 3.3.**

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

**Theorem 3.4.**

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

**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

$$H(U_1, U_2, \ldots U_n) = H(U_1) + H(U_2|U_1) + \cdots + H(U_n|U_1, U_2 \ldots U_{n-1})$$

**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.**

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

**Notation 1.**

$$U^n \triangleq (U_1, U_2, \dots U_n)$$

**Theorem 3.6.**

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

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

*Proof.*

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

□

# 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

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

□

**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,

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

□

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

**Definition 4.1** (Entropy rate of a stochastic process)**.** *$\lim_{n\to\infty} \frac{1}{n} H(U^n)$ if the limit exists.*

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

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

*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

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

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

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

□

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

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

*then*

1. *for every source coding scheme*

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

*the expected number of bits / letter is given by*

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

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

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

*Proof.* 1. we already know

$$\frac{1}{n}E[l(\mathscr{C}_n(\mathcal{U}^n))] \ge \frac{1}{n}H(\mathcal{U}_1 \ldots \mathcal{U}_n)$$

and the right term is decreasing

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

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

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

□

# 5 Typicality and typical set

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

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

**Theorem 5.1.** *$u^n$ is $(\epsilon, p)$-typical then*

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

*Proof.*

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

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 5.1** (Typical set)**.**

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

**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.*

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

*4.*

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

*Proof.* TODO: □

**Definition 5.2** (Kullback-Leiber divergence (information gain))**.** *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 number of average additional bits per datum necessary for compression.*

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

$$\begin{aligned} Pr\{U^n = u^n\} &= \left[\prod q(u)^{p(u)}\right]^{n(1+\epsilon)} \\ &= 2^{-n(1\pm\epsilon)} \sum_u p(u) \log \frac{1}{q(u)} \end{aligned}$$

$U_1, U_2, \dots$ iid $\sim p$
$Pr\{U^n \in T(n,\epsilon,p)\} \to 1$ as $n \to \infty$

$$(1-\epsilon)2^{nH(U)(1-\epsilon) \leq |T(n,\epsilon,p)| \leq 2^{nH(U)(1+\epsilon)}}$$

Suppose $U_1 \dots U_n$ are iid following $q$ and $u^n \in T(n,\epsilon,p)$
Observe:

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

and

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

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

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

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

*Proof.*

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

We have

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

□

**Theorem 5.4.**

$$D(p||q) = \sum_u p(u) \log\left(\frac{p(u)}{q(u)}\right) \geq 0 \text{ with equality iff } p = q$$

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

$$D(p||q) = \frac{1}{2}\log\frac{1}{1-2\delta} + frac12\log\frac{1}{1+2\delta} = \frac{1}{2}\log\frac{1}{1-4\delta^2} = -\frac{1}{2}\log(1-4\delta^2) \approx \frac{1}{2}4\delta^2 + o(\delta^4)$$

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

**Example 4.** *Suppose we are told that* $U$ *is* $p$ *distributed and* $p(u)$ *are powers of 2 and we design a prefix-free code* $\mathscr{C}$ *to minimize* $\sum_u p(u)l(\mathscr{C}(u))$.

*We have been misinformed and* $U \sim q$

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

## 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 $E\left[h_2\left(\frac{K}{n}\right)\right] \leq h_2 E\left[\frac{K}{n}\right] = h_2(\theta)$

Suggestion for $\mathscr{C}$

Because the probability of a bit string is only dependant of the number of 1 (or 0), 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 $p_0 = 1 - p_1$

Observe for any $0 \leq \alpha \leq 1$

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

Then for all $\alpha$

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

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

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

with this bound we have

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

$$E\left[\frac{1}{n}l(\mathscr{C}(U))\right] \ leqo(n) + E\left[h_2\left(\frac{k}{n}\right)\right], \text{ with } K = \sum u_i$$

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

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

*consequently this scheme is asymptotically optimal.*

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

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

For this let $U$ and $V$ be random variables with $U \in \{0,1\}, V \in \{1, \ldots, k\}$ with $Pr\left\{V = i\right\} = q_i$, $Pr\left\{U = 1 | V = i\right\} = \beta_i$ and $Pr\left\{U = 0 | V = i\right\} = 1 - \beta_i$

$Pr\left\{U = 1\right\} = \sum_i q_i \beta_i$, $H(U) = h_2\left(\sum_i q_i \beta_i\right)$ and $H(U|V) = \sum_i q_i h_2(\beta_i)$

We already know $H(U) \geq H(U|V)$ $\square$

# 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.*