Information Theory
Part I. Shannon entropy

Damian Niwiński

University of Warsaw

Winter semester 2020/2021

Disclaimer. Credits to many authors. All errors are mine own.

Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte.
I have made this [letter] longer, because I have not had the time to make it shorter.

Blaise Pascal, *Lettres provinciales*, 1657

Can any message be made shorter?

Can any message be made shorter?

14159 26535 89793 23846 26433 83279 50288 41971 69399 37510 58209 74944 59230 78164 06286 20899 86280 34825 34211 70679

Can any message be made shorter?

14159 26535 89793 23846 26433 83279 50288 41971 69399 37510 58209 74944 59230 78164 06286 20899 86280 34825 34211 70679

100 digits (after comma) of π

Can any message be made shorter?

3.14159 26535 89793 23846 26433 83279 50288 41971 69399 37510 58209 74944 59230 78164 06286 20899 86280 34825 34211 70679

100 digits (after comma) of π

Can any message be made shorter?

Let n be the smallest integer that cannot be described in English with less than 1000 signs.

Can any message be made shorter?

Let n be the smallest integer that cannot be described in English with less than 1000 signs.
(☺ **Berry's paradox**).

Notation, what is it?

Any **1:1** function $\alpha : S \to \Sigma^*$, where Σ is a finite alphabet, is a **notation for** S.

Notation, what is it?

Any **1:1** function $\alpha : S \to \Sigma^*$, where Σ is a finite alphabet, is a **notation for** S.

Fact. If $\alpha : S \to \Sigma^*$ is notation for a finite set S, with $|S| \geq 1$ and $|\Sigma| = r \geq 2$ then, for some $s \in S$,

$$|\alpha(s)| \geq \lfloor \log_r |S| \rfloor.$$

Notation, what is it?

Any **1:1** function $\alpha : S \to \Sigma^*$, where Σ is a finite alphabet, is a **notation for** S.

Fact. If $\alpha : S \to \Sigma^*$ is notation for a finite set S, with $|S| \geq 1$ and $|\Sigma| = r \geq 2$ then, for some $s \in S$,

$$|\alpha(s)| \geq \lfloor \log_r |S| \rfloor.$$

Proof. The number of strings shorter than some $n \geq 1$ is

$$1 + r + r^2 + \ldots + r^{n-1} = \frac{r^n - 1}{r - 1} < r^n.$$

Therefore, if $|S| \geq r^n$ then there must be $s \in S$, such that $|\alpha(s)| \geq n$.

Notation, what is it?

Any **1:1** function $\alpha : S \to \Sigma^*$, where Σ is a finite alphabet, is a **notation for** S.

Fact. If $\alpha : S \to \Sigma^*$ is notation for a finite set S, with $|S| \geq 1$ and $|\Sigma| = r \geq 2$ then, for some $s \in S$,

$$|\alpha(s)| \geq \lfloor \log_r |S| \rfloor.$$

Proof. The number of strings shorter than some $n \geq 1$ is

$$1 + r + r^2 + \ldots + r^{n-1} = \frac{r^n - 1}{r - 1} < r^n.$$

Therefore, if $|S| \geq r^n$ then there must be $s \in S$, such that $|\alpha(s)| \geq n$.

Choose $r^n \leq |S| < r^{n+1}$. □

Numbers with long notation

Fact. If $\alpha : \mathbb{N} \to \Sigma^*$ is notation for natural numbers with $|\Sigma| = r \geq 2$ then, for infinitely many k's,

$$|\alpha(k)| > \log_r k.$$

Numbers with long notation

Fact. If $\alpha : \mathbb{N} \to \Sigma^*$ is notation for natural numbers with $|\Sigma| = r \geq 2$ then, for infinitely many k's,

$$|\alpha(k)| > \log_r k.$$

Proof. For $n \geq |\alpha(0)| + 1$, let

$$k_n = \min\{k \in \mathbb{N} : |\alpha(k)| \geq n\}.$$

Numbers with long notation

Fact. If $\alpha : \mathbb{N} \to \Sigma^*$ is notation for natural numbers with $|\Sigma| = r \geq 2$ then, for infinitely many k's,

$$|\alpha(k)| > \log_r k.$$

Proof. For $n \geq |\alpha(0)| + 1$, let

$$k_n = \min\{k \in \mathbb{N} : |\alpha(k)| \geq n\}.$$

Then $k_n > 0$, and for $i = 0, 1, \ldots, k_n - 1$, $|\alpha(i)| < n$.

Numbers with long notation

Fact. If $\alpha : \mathbb{N} \to \Sigma^*$ is notation for natural numbers with $|\Sigma| = r \geq 2$ then, for infinitely many k's,

$$|\alpha(k)| > \log_r k.$$

Proof. For $n \geq |\alpha(0)| + 1$, let

$$k_n = \min\{k \in \mathbb{N} : |\alpha(k)| \geq n\}.$$

Then $k_n > 0$, and for $i = 0, 1, \ldots, k_n - 1$, $|\alpha(i)| < n$.

Hence $k_n < r^n$, and consequently

$$\begin{aligned} \log_r k_n &< n \\ &\leq |\alpha(k_n)| \end{aligned}$$

□

Numbers with long notation

The above estimation is tight, for example, with $\Sigma = \{0, 1\}$,

n	0	1	2	3	4	5	6	7
$\alpha(n)$	ε	0	1	00	01	10	11	000

i.e., $\alpha(n) = \{0, 1\}^{-1}\text{bin}(n + 1)$, satisfies

$$|\alpha(n)| \leq \lceil \log_2 n \rceil,$$

for each $n \geq 2$.

Application

Fact (Euclid). There are infinitely many primes.

Application

Fact (Euclid). There are infinitely many primes.

Proof. Suppose there are only M primes: $p_1, \dots, p_M$. Define $\alpha : \mathbb{N} \to \{0, 1, \#\}$, for $n = p_1^{\beta_1} \dots p_M^{\beta_M}$,

$$\alpha(n) = bin(\beta_1)\#bin(\beta_2)\# \dots \#bin(\beta_M).$$

Application

Fact (Euclid). There are infinitely many primes.

Proof. Suppose there are only M primes: $p_1, \ldots, p_M$. Define $\alpha : \mathbb{N} \to \{0, 1, \#\}$, for $n = p_1^{\beta_1} \ldots p_M^{\beta_M}$,

$$\alpha(n) = bin(\beta_1)\#bin(\beta_2)\# \ldots \#bin(\beta_M).$$

Then

$$|\alpha(n)| \leq M(2 + \log_2 \log_2 n)$$

for all $n > 0$, which clearly contradicts that $|\alpha(n)| > \log_3 n$, for infinitely many n's. $\square$

Codes

For $\varphi : S \to \Sigma^*$, let $\hat{\varphi}(s_1 \dots s_\ell) = \varphi(s_1) \dots \varphi(s_\ell)$.

Codes

For $\varphi : S \to \Sigma^*$, let $\hat{\varphi}(s_1 \dots s_\ell) = \varphi(s_1) \dots \varphi(s_\ell)$.

A notation $\varphi : S \to \Sigma^*$ for a finite set S is a **code** if the mapping $\hat{\varphi}$ is **1:1**.

Codes

For $\varphi : S \to \Sigma^*$, let $\hat{\varphi}(s_1 \dots s_\ell) = \varphi(s_1) \dots \varphi(s_\ell)$.

A notation $\varphi : S \to \Sigma^*$ for a finite set S is a **code** if the mapping $\hat{\varphi}$ is **1:1**.

We call the set $\{\varphi(s) : s \in S\}$ a **code** as well.

Note. A set $C \subseteq \Sigma^*$ is a code if any word in C^* is a product of words in C in a **unique** way.

Codes

For $\varphi : S \to \Sigma^*$, let $\hat{\varphi}(s_1 \ldots s_\ell) = \varphi(s_1) \ldots \varphi(s_\ell)$.

A notation $\varphi : S \to \Sigma^*$ for a finite set S is a **code** if the mapping $\hat{\varphi}$ is **1:1**.

We call the set $\{\varphi(s) : s \in S\}$ a **code** as well.

Note. A set $C \subseteq \Sigma^*$ is a code if any word in C^* is a product of words in C in a **unique** way.

Examples. If no word in C is a prefix of another word, C is a **prefix-free code** code (sometimes called *prefix code*).

Codes

For $\varphi : S \to \Sigma^*$, let $\hat{\varphi}(s_1 \ldots s_\ell) = \varphi(s_1) \ldots \varphi(s_\ell)$.

A notation $\varphi : S \to \Sigma^*$ for a finite set S is a **code** if the mapping $\hat{\varphi}$ is **1:1**.

We call the set $\{\varphi(s) : s \in S\}$ a **code** as well.

Note. A set $C \subseteq \Sigma^*$ is a code if any word in C^* is a product of words in C in a **unique** way.

Examples. If no word in C is a prefix of another word, C is a **prefix-free code** code (sometimes called *prefix code*).

The set $\{a, ab, ba\}$ is **not** a code, e.g. $a \cdot ba = ab \cdot a$..

Codes

For $\varphi : S \to \Sigma^*$, let $\hat{\varphi}(s_1 \dots s_\ell) = \varphi(s_1) \dots \varphi(s_\ell)$.

A notation $\varphi : S \to \Sigma^*$ for a finite set S is a **code** if the mapping $\hat{\varphi}$ is **1:1**.

We call the set $\{\varphi(s) : s \in S\}$ a **code** as well.

Note. A set $C \subseteq \Sigma^*$ is a code if any word in C^* is a product of words in C in a **unique** way.

Examples. If no word in C is a prefix of another word, C is a **prefix-free code** code (sometimes called *prefix code*).

The set $\{a, ab, ba\}$ is **not** a code, e.g. $a \cdot ba = ab \cdot a$..

The set $\{aa, baa, ba\}$ **is** a code (not prefix-free).

Codes

For $\varphi : S \to \Sigma^*$, let $\hat{\varphi}(s_1 \dots s_\ell) = \varphi(s_1) \dots \varphi(s_\ell)$.

A notation $\varphi : S \to \Sigma^*$ for a finite set S is a **code** if the mapping $\hat{\varphi}$ is **1:1**.

We call the set $\{\varphi(s) : s \in S\}$ a **code** as well.

Note. A set $C \subseteq \Sigma^*$ is a code if any word in C^* is a product of words in C in a **unique** way.

Examples. If no word in C is a prefix of another word, C is a **prefix-free code** code (sometimes called *prefix code*).

The set $\{a, ab, ba\}$ is **not** a code, e.g. $a \cdot ba = ab \cdot a$..

The set $\{aa, baa, ba\}$ **is** a code (not prefix-free).

(Any word in $(aa)^+ + (aa)^* (ba^+)^+$ can be uniquely decoded.)

Codes

Property. A code φ is prefix iff, for any $v, w \in S^*$, $\hat{\varphi}(v) \leq \hat{\varphi}(w)$ implies $v \leq w$.

Codes

Property. A code φ is prefix iff, for any $v, w \in S^*$, $\hat{\varphi}(v) \leq \hat{\varphi}(w)$ implies $v \leq w$.

For this reason, a prefix-free code is also called **instantaneous**.

Codes

Property. A code φ is prefix iff, for any $v, w \in S^*$, $\hat{\varphi}(v) \leq \hat{\varphi}(w)$ implies $v \leq w$.

For this reason, a prefix-free code is also called **instantaneous**.

For a non-prefix code, e..g, $\{aa, baa, ba\}$, we may have

b	**a**	**a**	**a**	
b	**a**	**a**	**a**	**a**

Codes

Property. A code φ is prefix iff, for any $v, w \in S^*$, $\hat{\varphi}(v) \leq \hat{\varphi}(w)$ implies $v \leq w$.

For this reason, a prefix-free code is also called **instantaneous**.

For a non-prefix code, e..g, $\{aa, baa, ba\}$, we may have

b a a a
b a a a a

What can we say about the **length** of words in a code with m elements **?**

Kraft's inequality

Fact. If $C \subseteq \Sigma^*$ is an instantaneous code ($|\Sigma| = r \geq 2$) then

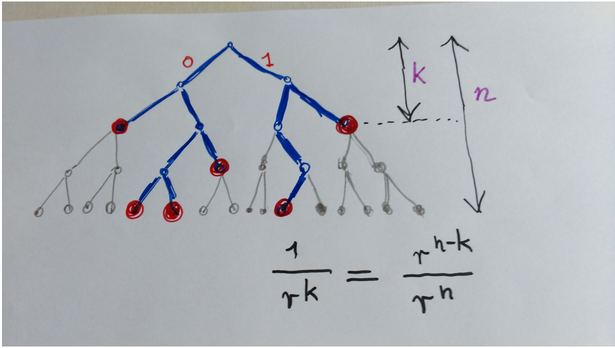
$$\sum_{w \in C} \frac{1}{r^{|w|}} \leq 1.$$

Kraft's inequality

Fact. If $C \subseteq \Sigma^*$ is an instantaneous code ($|\Sigma| = r \geq 2$) then

$$\sum_{w \in C} \frac{1}{r^{|w|}} \leq 1.$$

Proof by example. Take 00, 0100, 0101, 011, 1010, 11.



Kraft's inequality — characterization

Theorem. Let $2 \leq |S| < \infty$, and $\ell : S \to \mathbb{N}$. Then

$$\sum_{s \in S} \frac{1}{r^{\ell(s)}} \quad \leq \quad 1$$

if, and only if, $\ell = |\varphi|$, for some instantaneous code $\varphi : S \to \Sigma^*$, with $|\Sigma| = r$.

Kraft's inequality — characterization

Theorem. Let $2 \leq |S| < \infty$, and $\ell : S \to \mathbb{N}$. Then

$$\sum_{s \in S} \frac{1}{r^{\ell(s)}} \quad \leq \quad 1$$

if, and only if, $\ell = |\varphi|$, for some instantaneous code $\varphi : S \to \Sigma^*$, with $|\Sigma| = r$.

Proof (only if). W.l.o.g. $S = \{1, \ldots, m\}$, and $\ell(1) \leq \ldots \leq \ell(m)$. For $i = 0, 1, \ldots, m-1$, let $\varphi(i+1)$ be the **lexicographically first** word in $\Sigma^{\ell(i+1)}$ not extending any of $\varphi(1), \ldots, \varphi(i)$.

Kraft's inequality — characterization

Theorem. Let $2 \leq |S| < \infty$, and $\ell : S \to \mathbb{N}$. Then

$$\sum_{s \in S} \frac{1}{r^{\ell(s)}} \leq 1$$

if, and only if, $\ell = |\varphi|$, for some instantaneous code $\varphi : S \to \Sigma^*$, with $|\Sigma| = r$.

Proof (only if). W.l.o.g. $S = \{1, \ldots, m\}$, and $\ell(1) \leq \ldots \leq \ell(m)$. For $i = 0, 1, \ldots, m-1$, let $\varphi(i+1)$ be the **lexicographically first** word in $\Sigma^{\ell(i+1)}$ not extending any of $\varphi(1), \ldots, \varphi(i)$.
Can we always do it, i.e.

$$r^{\ell(i+1)-\ell(1)} + r^{\ell(i+1)-\ell(2)} + \ldots + r^{\ell(i+1)-\ell(i)} < r^{\ell(i+1)} \text{ ?}$$

Kraft's inequality — characterization

Theorem. Let $2 \leq |S| < \infty$, and $\ell : S \to \mathbb{N}$. Then

$$\sum_{s \in S} \frac{1}{r^{\ell(s)}} \quad \leq \quad 1$$

if, and only if, $\ell = |\varphi|$, for some instantaneous code $\varphi : S \to \Sigma^*$, with $|\Sigma| = r$.

Proof (only if). W.l.o.g. $S = \{1, \dots, m\}$, and $\ell(1) \leq \dots \leq \ell(m)$.
For $i = 0, 1, \dots, m-1$, let $\varphi(i+1)$ be the **lexicographically first** word in $\Sigma^{\ell(i+1)}$ not extending any of $\varphi(1), \dots, \varphi(i)$.
Can we always do it, i.e.

$$r^{\ell(i+1)-\ell(1)} + r^{\ell(i+1)-\ell(2)} + \dots + r^{\ell(i+1)-\ell(i)} < r^{\ell(i+1)} \text{ ?}$$

Yes, because

$$\frac{1}{r^{\ell(1)}} + \frac{1}{r^{\ell(2)}} + \dots + \frac{1}{r^{\ell(i)}} < 1. \qquad \square$$

McMillan's theorem

Theorem. For any code $\varphi : S \to \Sigma^*$, there is an instantaneous code φ' with $|\varphi| = |\varphi'|$.

(Thus **any** code satisfies Kraft's inequality.)

McMillan's theorem

Theorem. For any code $\varphi : S \to \Sigma^*$, there is an instantaneous code φ' with $|\varphi| = |\varphi'|$.

(Thus **any** code satisfies Kraft's inequality.)

Proof. **Suppose** $K = \sum_{s \in S} \frac{1}{r^{|\varphi(s)|}} > 1$.

McMillan's theorem

Theorem. For any code $\varphi : S \to \Sigma^*$, there is an instantaneous code φ' with $|\varphi| = |\varphi'|$.

(Thus **any** code satisfies Kraft's inequality.)

Proof. **Suppose** $K = \sum_{s\in S} \frac{1}{r^{|\varphi(s)|}} > 1$.
Let $Min = \min\{|\varphi(s)| : s \in S\}$, $Max = \max\{|\varphi(s)| : s \in S\}$.
Consider

$$K^n = \left(\sum_{s\in S} \frac{1}{r^{|\varphi(s)|}}\right)^n = \sum_{i=Min\cdot n}^{Max\cdot n} \frac{N_{n,i}}{r^i},$$

where $N_{n,i}$ is the number of sequences $q_1, \ldots, q_n \in S^n$, such that $i = |\varphi(q_1)| + \ldots + |\varphi(q_n)| = |\hat{\varphi}(q_1 \ldots q_n)|$.

McMillan's theorem

Theorem. For any code $\varphi : S \to \Sigma^*$, there is an instantaneous code φ' with $|\varphi| = |\varphi'|$.

(Thus **any** code satisfies Kraft's inequality.)

Proof. **Suppose** $K = \sum_{s \in S} \frac{1}{r^{|\varphi(s)|}} > 1$.
Let $Min = \min\{|\varphi(s)| : s \in S\}$, $Max = \max\{|\varphi(s)| : s \in S\}$.
Consider

$$K^n = \left(\sum_{s \in S} \frac{1}{r^{|\varphi(s)|}} \right)^n = \sum_{i=Min \cdot n}^{Max \cdot n} \frac{N_{n,i}}{r^i},$$

where $N_{n,i}$ is the number of sequences $q_1, \ldots, q_n \in S^n$, such that $i = |\varphi(q_1)| + \ldots + |\varphi(q_n)| = |\hat{\varphi}(q_1 \ldots q_n)|$. But **at most one** such sequence can be encoded by a word in Σ^i, hence

$$\frac{N_{n,i}}{r^i} \leq 1,$$

and

$$K^n \leq (Max - Min) \cdot n + 1, \qquad \textbf{impossible!}$$

Average length of a code

Let $p : S \to [0.1]$ be a **probability distribution** over S.

We wish to minimize

$$\sum_{s \in S} p(s) \cdot |\varphi(s)|,$$

for a code φ.

Average length of a code

Let $p : S \to [0.1]$ be a **probability distribution** over S.

We wish to minimize

$$\sum_{s \in S} p(s) \cdot |\varphi(s)|,$$

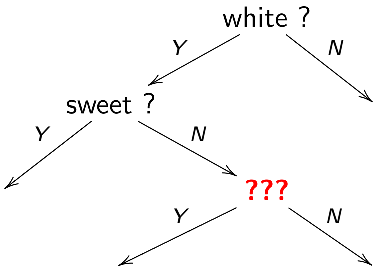
for a code φ.

Let $S = \{s_1, \ldots, s_m\}$, $p(s_i) = p_i$.

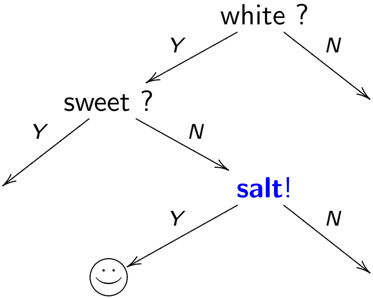
Task. Among all tuples $\ell_1, \ldots, \ell_m$, satisfying Kraft's inequality find a one with minimal

$$\sum_i p_i \cdot \ell_i.$$

Relation to 20 question game



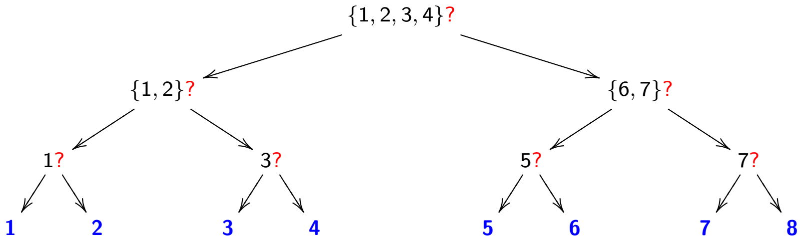
Relation to 20 question game



Relation to 20 question game

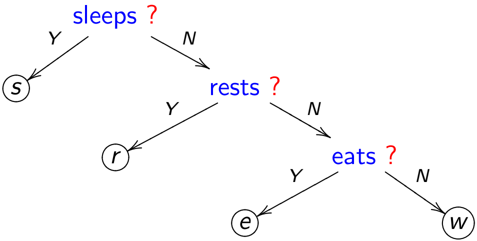
For n possibilities, $\lceil \log_2 n \rceil$ question suffices.

$S = \{1, 2, 3, 4, 5, 6, 7, 8\}$



Relation to 20 question game

But knowing the probability we can do better.
p (sleeps) $= \frac{1}{2}$, p (rests) $= \frac{1}{4}$, p (eats) $=$ p (works) $= \frac{1}{8}$.



Average number of questions:

$$1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \left(\frac{1}{8} + \frac{1}{8}\right) = \frac{7}{4} < 2 = \log_2 4.$$

Relation to 20 question game

We wish to find an object in S, knowing a probability distribution $p : S \to [0.1]$.

Task. Find a strategy that minimizes the average number of questions.

Note. Any strategy induces an instantaneous code over $\{0, 1\}$:
$\varphi(s) =$ the sequence of yes and no answers leading to s.

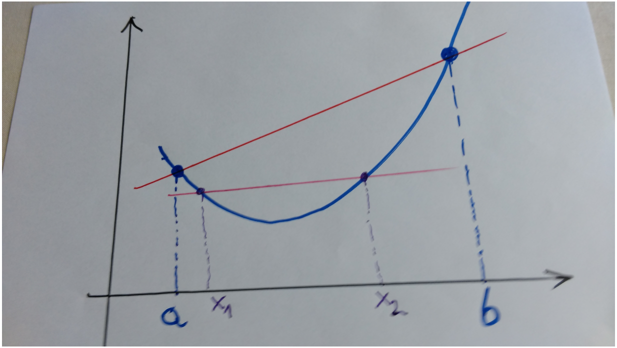
Conversely, an instantaneous code induces a strategy.

Calculus revisited — convex functions

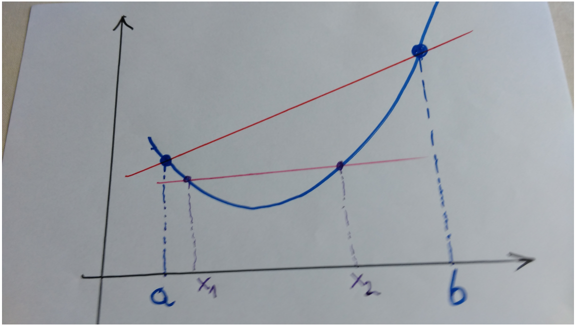
A function $f : [a, b] \to \mathbb{R}$ is **convex** (on $[a, b]$) if $\forall x_1, x_2 \in [a, b]$, $\forall \lambda \in [0, 1]$,

$$\lambda f(x_1) + (1 - \lambda) f(x_2) \quad \geq \quad f(\lambda x_1 + (1 - \lambda) x_2).$$

It is **strictly convex** if the inequality is strict, except for $\lambda \in \{0, 1\}$ and $x_1 = x_2$.



Calculus revisited — convex functions



Lemma. If f is continuous on $[a, b]$ and has a **second derivative** on (a, b) with $f'' \geq 0$ ($f'' > 0$) then it is convex (strictly convex).

Jensen's inequality

Let X be a **random variable** over a finite probability space S.

If $S = \{s_1, \ldots, s_m\}$, we let $p(s_i) = p_i$, $X(s) = x_i$.

Jensen's inequality

Let X be a **random variable** over a finite probability space S.

If $S = \{s_1, \ldots, s_m\}$, we let $p(s_i) = p_i$, $X(s) = x_i$.

X is **constant** if there are no $x_i \neq x_j$ with $p_i, p_j > 0$.

Jensen's inequality

Let X be a **random variable** over a finite probability space S.

If $S = \{s_1, \ldots, s_m\}$, we let $p(s_i) = p_i$, $X(s) = x_i$.

X is **constant** if there are no $x_i \neq x_j$ with $p_i, p_j > 0$.

The **expected value** of X is

$$EX = \sum_{s \in S} p(s) \cdot X(s) = p_1 x_1 + \ldots + p_m x_m.$$

Jensen's inequality

Let X be a **random variable** over a finite probability space S.

If $S = \{s_1, \ldots, s_m\}$, we let $p(s_i) = p_i$, $X(s) = x_i$.

X is **constant** if there are no $x_i \neq x_j$ with $p_i, p_j > 0$.

The **expected value** of X is

$$EX = \sum_{s \in S} p(s) \cdot X(s) = p_1 x_1 + \ldots + p_m x_m.$$

Theorem (Jensen's inequality)
If $f : [a, b] \to \mathbb{R}$ is a convex function then, for any random variable $X : S \to [a, b]$,

$$Ef(X) \geq f(EX).$$

If moreover f is strictly convex then the inequality is strict unless X is constant.

Thm . . . . . . $Ef(X) \geq f(EX)$.

Proof. By induction on $|S|$. For $|S| = 2$,
$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$, convexity.

Thm . . . . . . $Ef(X) \geq f(EX)$.

Proof. By induction on $|S|$. For $|S| = 2$,
$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$, convexity.
Let $|S| = m$, w.l.o.g. $p_m < 1$.

Let $p'_i = \frac{p_i}{1-p_m}$, for $i = 1, \ldots, m-1$.

Thm ...... $Ef(X) \geq f(EX)$.

Proof. By induction on $|S|$. For $|S| = 2$,
$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$, convexity.
Let $|S| = m$, w.l.o.g. $p_m < 1$.
Let $p'_i = \frac{p_i}{1-p_m}$, for $i = 1, \ldots, m-1$.

$$\begin{aligned} \sum_{i=1}^{m} p_i f(x_i) &= p_m f(x_m) + (1-p_m) \sum_{i=1}^{m-1} p'_i f(x_i) \\ &\geq p_m f(x_m) + (1-p_m) f\left(\sum_{i=1}^{m-1} p'_i x_i\right) \\ &\geq f\left(p_m x_m + (1-p_m) \sum_{i=1}^{m-1} p'_i x_i\right) \\ &= f\left(\sum_{i=1}^{m} p_i x_i\right). \end{aligned}$$

If f is strictly convex, but

$$\begin{aligned}\sum_{i=1}^{m} p_i f(x_i) &= p_m f(x_m) + (1-p_m)\sum_{i=1}^{m-1} p_i' f(x_i) \\ &= p_m f(x_m) + (1-p_m)\left(\sum_{i=1}^{m-1} p_i' x_i\right) \\ &= f(p_m x_m + (1-p_m)\sum_{i=1}^{m-1} p_i' x_i) \\ &= f(\sum_{i=1}^{m} p_i x_i),\end{aligned}$$

then $x_i = \mathbf{C}$, for all $i = 1, \ldots, m-1$, unless $p_i' = p_i = 0$.

Moreover, either $p_m = 0$ or $x_m = \sum_{i=1}^{m-1} p_i' x_i = \mathbf{C}$, as well. $\square$

The function $x \log x$

Convention: $0 \log_r 0 = 0 \log_r \frac{1}{0} = 0$.

Justified by
$\lim_{x\to 0} x \log_r x = \lim_{x\to 0} -x \log_r \frac{1}{x} = \lim_{y\to\infty} -\frac{\log_r y}{y} = 0$.

Fact. For $r > 1$, the function $\mathbf{x \log_r x}$ is **strictly convex** on $[0, \infty)$ (i.e., on any $[0, M]$, $M > 0$).

Proof.

$$(x \log_r x)'' = \left(\log_r x + x \cdot \frac{1}{x} \cdot \log_r e\right)' = \frac{1}{x} \cdot \log_r e > 0.$$

□

Golden lemma

Theorem (Gibbs' inequality)

Suppose $1 = \sum_{i=1}^{m} x_i \geq \sum_{i=1}^{m} y_i$, where $x_i \geq 0$ and $y_i > 0$, for $i = 1, \ldots, m$, and let $r > 1$.
Then

$$\sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \quad \geq \quad \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i},$$

and the equality holds only if $x_i = y_i$, for $i = 1, \ldots, m$.

Golden lemma

Theorem (Gibbs' inequality)

Suppose $1 = \sum_{i=1}^{m} x_i \geq \sum_{i=1}^{m} y_i$, where $x_i \geq 0$ and $y_i > 0$, for $i = 1, \ldots, m$, and let $r > 1$.
Then

$$\sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \quad \geq \quad \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i},$$

and the equality holds only if $x_i = y_i$, for $i = 1, \ldots, m$.

Corollary. If $\ell_1, \ldots, \ell_m$ satisfy $\sum_i \frac{1}{r^{\ell_i}} \leq 1$ then

$$\sum_i p_i \cdot \ell_i \quad \geq \quad \sum_i p_i \cdot \log_r \frac{1}{p_i}.$$

Hence, the minimum is achieved if $\ell_i = \log_r \frac{1}{p_i}$, for $i = 1, \ldots, m$.

. . . . $\sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}$.

Proof. Let us first assume that $\sum_{i=1}^{m} y_i = 1$. We have

$$Left - Right = \sum_{i=1}^{m} x_i \cdot \log_r \frac{x_i}{y_i} = \sum_{i=1}^{m} y_i \cdot \left(\frac{x_i}{y_i}\right) \cdot \log_r \frac{x_i}{y_i}$$

$$\geq \log_r \underbrace{\sum_{i=1}^{m} y_i \cdot \left(\frac{x_i}{y_i}\right)}_{1} = 0.$$

Here we apply **Jensen's inequality** to the function $x \log_r x$ (strictly convex on $[0, \infty)$) and the random variable which takes the value $\left(\frac{x_i}{y_i}\right)$ with probability y_i.

.... $\sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}$.

Proof. Let us first assume that $\sum_{i=1}^{m} y_i = 1$. We have

$$Left - Right = \sum_{i=1}^{m} x_i \cdot \log_r \frac{x_i}{y_i} = \sum_{i=1}^{m} y_i \cdot \left(\frac{x_i}{y_i}\right) \cdot \log_r \frac{x_i}{y_i}$$

$$\geq \log_r \underbrace{\sum_{i=1}^{m} y_i \cdot \left(\frac{x_i}{y_i}\right)}_{1} = 0.$$

Here we apply **Jensen's inequality** to the function $x \log_r x$ (strictly convex on $[0, \infty)$) and the random variable which takes the value $\left(\frac{x_i}{y_i}\right)$ with probability y_i.

The **equality** holds if this random variable is **constant**. Remembering that $y_i > 0$, and $\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$, we then have $x_i = y_i$, for $i = 1, \ldots, m$.

$\ldots\ldots\ldots\ldots \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}$.

Proof continued, the **case** $\sum_{i=1}^{m} y_i < 1$.

$\dots\dots\dots\sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}$.

Proof continued, the **case** $\sum_{i=1}^{m} y_i < 1$.

Let $y_{m+1} = 1 - \sum_{i=1}^{m} y_i$, and $x_{m+1} = 0$.

$\dots\dots\dots\dots \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}$.

Proof continued, the **case** $\sum_{i=1}^{m} y_i < 1$.

Let $y_{m+1} = 1 - \sum_{i=1}^{m} y_i$, and $x_{m+1} = 0$.

Then, by the previous case we have

$$\sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} = \sum_{i=1}^{m+1} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m+1} x_i \cdot \log_r \frac{1}{x_i} = \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}.$$

$\ldots\ldots\ldots\ldots \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}$.

Proof continued, the **case** $\sum_{i=1}^{m} y_i < 1$.

Let $y_{m+1} = 1 - \sum_{i=1}^{m} y_i$, and $x_{m+1} = 0$.

Then, by the previous case we have

$$\sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{y_i} = \sum_{i=1}^{m+1} x_i \cdot \log_r \frac{1}{y_i} \geq \sum_{i=1}^{m+1} x_i \cdot \log_r \frac{1}{x_i} = \sum_{i=1}^{m} x_i \cdot \log_r \frac{1}{x_i}.$$

The **equality may not hold** in this case, as it would imply $x_i = y_i$, for $i = 1, \ldots, m+1$, which contradicts the choice of $y_{m+1} \neq x_{m+1}$.

$\square$

Shannon's entropy

The **entropy** of a finite probabilistic space S (with parameter $r > 1$) is

$$\begin{aligned} H_r(S) &= \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)} \\ &= -\sum_{s \in S} p(s) \cdot \log_r p(s). \end{aligned}$$

Traditionally, $H = H_2$.

Shannon's entropy

The **entropy** of a finite probabilistic space S (with parameter $r > 1$) is

$$H_r(S) = \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)} = -\sum_{s \in S} p(s) \cdot \log_r p(s).$$

Traditionally, $H = H_2$.

First occurred in: Claude Shannon, *A Mathematical Theory of Communication*, **1948**.

Shannon's entropy

$H_r(S) = \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)}$

Property.

$$0 \leq H_r(S) \leq \log_r |S|.$$

Shannon's entropy

$H_r(S) = \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)}$

Property.

$$0 \leq H_r(S) \leq \log_r |S|.$$

The equality $0 = H_r(S)$ holds iff $p(s) = 1$, for some $s \in S$.

The equality $H_r(S) = \log_r |S|$ holds iff $p(s) = \frac{1}{|S|}$, for all $s \in S$.

Shannon's entropy

$H_r(S) = \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)}$

Property.

$$0 \leq H_r(S) \leq \log_r |S|.$$

The equality $0 = H_r(S)$ holds iff $p(s) = 1$, for some $s \in S$.

The equality $H_r(S) = \log_r |S|$ holds iff $p(s) = \frac{1}{|S|}$, for all $s \in S$.

Proof. By the Golden Lemma with $x_i = p(s_i)$ and $y_i = \frac{1}{|S|}$,

$$\sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)} \leq \sum_{s \in S} p(s) \cdot \log_r |S| = \log_r |S|,$$

with the equality for $p(s) = \frac{1}{|S|}$.

Minimal code length

For a code $\varphi : S \to \Sigma^*$ (with $|\Sigma| \geq 2$), by the Kraft inequality and Golden Lemma

$$\begin{array}{rcl} H_r(S) & \leq & L(\varphi) \\ & & \| \\ & & \sum_{s \in S} p(s) \cdot |\varphi(s)| \end{array}$$

Consequently,

$$\begin{array}{rcl} H_r(S) & \leq & L_r(S) \\ & & \| \\ & & \min\{L(\varphi) : \varphi : S \to \Sigma^* \text{ is a code }\} \end{array}$$

Minimal code length

For a code $\varphi : S \to \Sigma^*$ (with $|\Sigma| \geq 2$), by the Kraft inequality and Golden Lemma

$$\begin{array}{rcl} H_r(S) & \leq & L(\varphi) \\ & & \| \\ & & \displaystyle\sum_{s \in S} p(s) \cdot |\varphi(s)| \end{array}$$

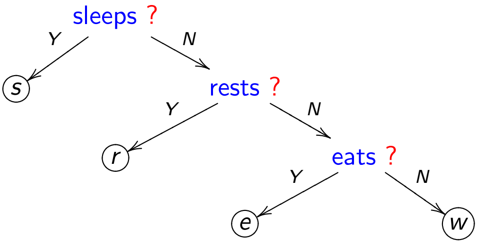
Consequently,

$$\begin{array}{rcl} H_r(S) & \leq & L_r(S) \\ & & \| \\ & & \min\{L(\varphi) : \varphi : S \to \Sigma^* \text{ is a code } \} \end{array}$$

That **min** exists is an exercise; it is realized by the **Huffman coding** ($\longrightarrow$ Tutorials).

Example — game revisited

p (sleeps) $= \frac{1}{2}$, p (rests) $= \frac{1}{4}$, p (eats) = p (works) $= \frac{1}{8}$.

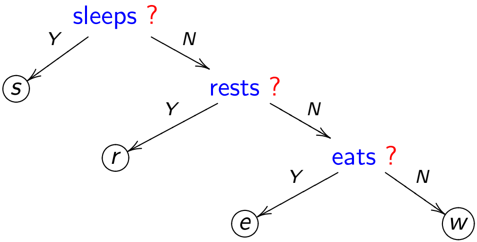


$$L(\varphi) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \left(\frac{1}{8} + \frac{1}{8}\right) = H_2(S)$$

Hence the strategy is **optimal** !

Example — game revisited

p (sleeps) $= \frac{1}{2}$, p (rests) $= \frac{1}{4}$, p (eats) = p (works) $= \frac{1}{8}$.



$$L(\varphi) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \left(\frac{1}{8} + \frac{1}{8}\right) = H_2(S)$$

Hence the strategy is **optimal** !

The number of questions for an option of probability q is $\log_2 \frac{1}{q}$.

Shannon–Fano coding

Theorem.

$$H_r(S) \le L_r(S) \le H_r(S) + 1.$$

Moreover, the equality $H_r(S) = L_r(S)$ holds if and only if $|S| \ge 2$ and all probabilities $p(s)$ are integer powers of $\frac{1}{r}$, and the equality $L_r(S) = H_r(S) + 1$ holds if and only if $H_r(S) = 0$.

Proof. If $|S| = 1$ then $0 = H_r(S) < L_r(S) = 1$. Let $|S| \ge 2$.

The inequality $H_r(S) \le L_r(S)$ already proved. The equality holds **iff** $H_r(S) = L(\varphi)$, for some code φ. The claim follows from Golden Lemma.

Proof of $L_r(S) < H_r(S) + 1$ unless $H_r(S) = 0$. Let

$$\ell(s) \quad = \quad \left\lceil \log_r \frac{1}{p(s)} \right\rceil$$

provided that $p(s) > 0$. Then

$$\sum_{s:p(s)>0} \frac{1}{r^{\ell(s)}} \leq \sum_{p(s)>0} p(s) = \sum_{s \in S} p(s) = 1.$$

If $(\forall s \in S)\, p(s) > 0$, then ℓ is defined on the whole S, and satisfies the Kraft inequality, hence there is a code with $|\varphi| = \ell$, and

$$L(\varphi) = \sum_{s \in S} p(s) \cdot \ell(s) < \sum_{s \in S} p(s) \cdot \left(\log_r \frac{1}{p(s)} + 1 \right) = H_r(S) + 1.$$

Suppose $p(s)$ is 0, for some s. If

$$\sum_{p(s)>0} \frac{1}{r^{\ell(s)}} \quad < \quad 1,$$

then we can extend ℓ to all s, preserving the Kraft inequality.

Again, there is a code with $|\varphi| = \ell$, satisfying

$$L(\varphi) = \sum_{s \in S} p(s) \cdot \ell(s) < \sum_{s \in S} p(s) \cdot \left(\log_r \frac{1}{p(s)} + 1 \right) = H_r(S) + 1.$$

Finally, suppose that

$$\sum_{p(s)>0} \frac{1}{r^{\ell(s)}} = 1. \qquad (*)$$

Finally, suppose that

$$\sum_{p(s)>0} \frac{1}{r^{\ell(s)}} \quad = \quad 1. \qquad\qquad (*)$$

We choose s' with $p(s') > 0$, and let

$$\begin{aligned} \ell'(s') &= \ell(s')+1 \\ \ell'(s) &= \ell(s), \text{ for } s \neq s'. \end{aligned}$$

Finally, suppose that

$$\sum_{p(s)>0} \frac{1}{r^{\ell(s)}} = 1. \qquad (*)$$

We choose s' with $p(s') > 0$, and let

$$\begin{aligned} \ell'(s') &= \ell(s')+1 \\ \ell'(s) &= \ell(s), \text{ for } s \neq s'. \end{aligned}$$

Again extend ℓ' so that there is a code with $|\varphi| = \ell'$.

Finally, suppose that

$$\sum_{p(s)>0} \frac{1}{r^{\ell(s)}} = 1. \qquad (*)$$

We choose s' with $p(s') > 0$, and let

$$\begin{aligned} \ell'(s') &= \ell(s') + 1 \\ \ell'(s) &= \ell(s), \text{ for } s \neq s'. \end{aligned}$$

Again extend ℓ' so that there is a code with $|\varphi| = \ell'$.

But $(*)$ implies $\ell(s) = \lceil \log_r \frac{1}{p(s)} \rceil = \log_r \frac{1}{p(s)}$. Hence

Finally, suppose that

$$\sum_{p(s)>0} \frac{1}{r^{\ell(s)}} \quad = \quad 1. \qquad\qquad (*)$$

We choose s' with $p(s') > 0$, and let

$$\begin{aligned} \ell'(s') &= \ell(s')+1 \\ \ell'(s) &= \ell(s), \text{ for } s \neq s'. \end{aligned}$$

Again extend ℓ' so that there is a code with $|\varphi| = \ell'$.

But $(*)$ implies $\ell(s) = \lceil \log_r \frac{1}{p(s)} \rceil = \log_r \frac{1}{p(s)}$. Hence

$$\begin{aligned} L(\varphi) &= \sum_{p(s)>0} p(s) \cdot \ell'(s) \\ &= p(s') + \sum_{p(s)>0} p(s) \cdot \ell(s) \\ &= p(s') + H_r(S) \\ &< H_r(S) + 1 \end{aligned}$$

unless there is no s' with $0 < p(s') < 1$.

Towards a better coding

Can we shrink the gap $[H_r(S), L_r(S)]$ further?

Towards a better coding

Can we shrink the gap $[H_r(S), L_r(S)]$ further?

Example. $S = \{s_1, s_2\}$, $p(s_1) = \frac{3}{4}$, $p(s_2) = \frac{1}{4}$.

$$H_2(S) < 1 = L_2(S).$$

Towards a better coding

Can we shrink the gap $[H_r(S), L_r(S)]$ further?

Example. $S = \{s_1, s_2\}$, $p(s_1) = \frac{3}{4}$, $p(s_2) = \frac{1}{4}$.

$$H_2(S) < 1 = L_2(S).$$

Encode 2-blocks

$s_1 s_1 \mapsto 0$ $\quad s_1 s_2 \mapsto 10$

$s_2 s_1 \mapsto 110$ $\quad s_2 s_2 \mapsto 111$

Towards a better coding

Can we shrink the gap $[H_r(S), L_r(S)]$ further?

Example. $S = \{s_1, s_2\}$, $p(s_1) = \frac{3}{4}$, $p(s_2) = \frac{1}{4}$.

$$H_2(S) < 1 = L_2(S).$$

Encode 2-blocks

$s_1s_1 \mapsto 0$ $\quad$ $s_1s_2 \mapsto 10$
$s_2s_1 \mapsto 110$ $\quad$ $s_2s_2 \mapsto 111$

With $p(s_i, s_j) = p(s_i) \cdot p(s_j)$, the average length of our encoding is

$$\left(\frac{3}{4}\right)^2 \cdot 1 + \frac{3}{4} \cdot \frac{1}{4} \cdot (2+3) + \left(\frac{1}{4}\right)^2 \cdot 3 = \frac{9}{16} + \frac{15}{16} + \frac{3}{16} = \frac{27}{16} < 2.$$

Entropy of product space

Fact. Let, for $(s.q) \in S \times Q$, $p(s,q) = p(s) \cdot p(q)$. Then

$$H_r(S \times Q) =$$

Entropy of product space

Fact. Let, for $(s.q) \in S \times Q$, $p(s,q) = p(s) \cdot p(q)$. Then

$$H_r(S \times Q) = H_r(S) + H_r(Q).$$

Entropy of product space

Fact. Let, for $(s.q) \in S \times Q$, $p(s,q) = p(s) \cdot p(q)$. Then

$$H_r(S \times Q) = H_r(S) + H_r(Q).$$

Proof.

$$\begin{aligned}
H(S \times Q) &= -\sum_{s,q} p(s,q) \cdot \log p(s,q) \\
&= -\sum_{s,q} p(s) \cdot p(q) \cdot (\log p(s) + \log p(q)) \\
&= -\sum_{s,q} p(s)\, p(q) \cdot \log p(s) \; - \; \sum_{s,q} p(s)\, p(q) \cdot \log p(q) \\
&= \sum_{q} p(q) \cdot H(S) + \sum_{s} p(s) \cdot H(Q) \\
&= H(S) + H(Q).
\end{aligned}$$

□

Shannon's coding theorem

Consequently, with $p(s_1, \ldots, s_n) = p(s_1) \cdot \ldots \cdot p(s_n)$,

$$H_r(S^n) \quad = \quad n \cdot H_r(S).$$

Shannon's coding theorem

Consequently, with $p(s_1, \dots, s_n) = p(s_1) \cdot \ldots \cdot p(s_n)$,

$$H_r(S^n) \;=\; n \cdot H_r(S).$$

Theorem. For any finite probabilistic space S and $r \geq 2$,

$$\lim_{n\to\infty} \frac{L_r(S^n)}{n} = H_r(S).$$

Shannon's coding theorem

Consequently, with $p(s_1, \ldots, s_n) = p(s_1) \cdot \ldots \cdot p(s_n)$,

$$H_r(S^n) \;=\; n \cdot H_r(S).$$

Theorem. For any finite probabilistic space S and $r \geq 2$,

$$\lim_{n \to \infty} \frac{L_r(S^n)}{n} = H_r(S).$$

Proof. Recall

$$H_r(S^n) \leq L_r(S^n) \leq H_r(S^n) + 1.$$

Since $H_r(S^n) = n \cdot H_r(S)$, this yields

$$H_r(S) \leq \frac{L_r(S^n)}{n} \leq H_r(S) + \frac{1}{n},$$

$\square$

Example — group testing

The state of a population consisting of N people is described by a vector of N bits (**1** – ill, **0** – healthy).

If the probability of being ill is $0 < p < 1$, the entropy for an individual is

$$H(p) \;=\; -p \log p - (1-p)\log(1-p),$$

and the entropy of the population is $N \cdot H(p)$ (assuming independence of events).

Group testing with 2 possible outcomes:
— someone in the group is infected,
— all people in the group are healthy,
is a **binary coding** method.

This gives us an estimation on the average number of tests T_N

$$N \cdot H(p) \;\leq\; T_N.$$

Random variables — notational conventions

For random variables $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$,

$$\begin{aligned}\sum_{s:A(s)=a} p(s) &= p(A=a) \\ &= p(a)\end{aligned}$$

$$\begin{aligned}p(A=a|B=b) &= p(a|b) \\ p\left((A=a) \wedge (B=b)\right) &= p(a \wedge b)\end{aligned}$$

etc.

Entropy of random variable

For a **random variable** $X : S \to \mathcal{T}$,

$$H_r(X) \stackrel{def}{=} \sum_{t \in \mathcal{T}} p(X = t) \cdot \log_r \frac{1}{p(X = t)}.$$

Entropy of random variable

For a **random variable** $X : S \to \mathcal{T}$,

$$H_r(X) \quad \stackrel{def}{=} \quad \sum_{t \in \mathcal{T}} p(X = t) \cdot \log_r \frac{1}{p(X = t)}.$$

Note: $H_r(X) = E\,\mathrm{LogPX}_r$, where

Entropy of random variable

For a **random variable** $X : S \to \mathcal{T}$,

$$H_r(X) \stackrel{def}{=} \sum_{t \in \mathcal{T}} p(X = t) \cdot \log_r \frac{1}{p(X = t)}.$$

Note: $H_r(X) = E\, \mathrm{LogPX}_r$, where

$$\mathrm{LogPX}_r(s) = \begin{cases} \log_r \frac{1}{p(X = X(s))} & \text{if} \quad p(s) > 0 \\ 0 & \text{if} \quad p(s) = 0. \end{cases}$$

Entropy of random variable

For a **random variable** $X : S \to \mathcal{T}$,

$$H_r(X) \stackrel{def}{=} \sum_{t \in \mathcal{T}} p(X = t) \cdot \log_r \frac{1}{p(X = t)}.$$

Note: $H_r(X) = E\,\mathrm{LogPX}_r$, where

$$\mathrm{LogPX}_r(s) = \begin{cases} \log_r \frac{1}{p(X=X(s))} & \text{if} \quad p(s) > 0 \\ 0 & \text{if} \quad p(s) = 0. \end{cases}$$

Indeed,

$$\sum_{t \in \mathcal{T}} p(X = t) \cdot \log_r \frac{1}{p(X = t)} = \sum_{t \in \mathcal{T}} \sum_{X(s)=t} p(s) \cdot \log_r \frac{1}{p(X = t)}$$

$$= \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(X = X(s))}.$$

Conditional entropy

Let $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$. For $a \in \mathcal{A}$ with $p(a) > 0$,

$$H_r(B|a) = \sum_{b \in \mathcal{B}} p(b|a) \cdot \log_r \frac{1}{p(b|a)}.$$

For $p(a) = 0$, $H_r(B|a) = 0$.

Conditional entropy

Let $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$. For $a \in \mathcal{A}$ with $p(a) > 0$,

$$H_r(B|a) \;=\; \sum_{b \in \mathcal{B}} p(b|a) \cdot \log_r \frac{1}{p(b|a)}.$$

For $p(a) = 0$, $H_r(B|a) = 0$.

$$H_r(B|A) \;\stackrel{def}{=}\; \sum_{a \in \mathcal{A}} p(a) \cdot H_r(B|a).$$

Conditional entropy

Let $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$. For $a \in \mathcal{A}$ with $p(a) > 0$,

$$H_r(B|a) \quad = \quad \sum_{b \in \mathcal{B}} p(b|a) \cdot \log_r \frac{1}{p(b|a)}.$$

For $p(a) = 0$, $H_r(B|a) = 0$.

$$H_r(B|A) \quad \stackrel{def}{=} \quad \sum_{a \in \mathcal{A}} p(a) \cdot H_r(B|a).$$

Note: if A and B are independent then $p(b|a) = p(b)$, and hence $H_r(B|A) = H_r(B)$.

Conditional entropy

Let $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$. For $a \in \mathcal{A}$ with $p(a) > 0$,

$$H_r(B|a) \quad = \quad \sum_{b \in \mathcal{B}} p(b|a) \cdot \log_r \frac{1}{p(b|a)}.$$

For $p(a) = 0$, $H_r(B|a) = 0$.

$$H_r(B|A) \quad \stackrel{def}{=} \quad \sum_{a \in \mathcal{A}} p(a) \cdot H_r(B|a).$$

Note: if A and B are independent then $p(b|a) = p(b)$, and hence $H_r(B|A) = H_r(B)$.

Similarly, $H_r(A|B) = H_r(A)$.

Conditional entropy of function

If $\varphi : \mathcal{A} \to \mathcal{B}$ then

$$H_r(\varphi(A)|A) =$$

Conditional entropy of function

If $\varphi : \mathcal{A} \to \mathcal{B}$ then

$$H_r(\varphi(A)|A) = 0.$$

Indeed, if $p(A = a) > 0$ then $p(\varphi(A) = \varphi(a)|A = a) = 1$, hence $\log_r \frac{1}{p(\varphi(A)=\varphi(a)|A=a)} = 0$.

Conditional entropy of function

If $\varphi : \mathcal{A} \to \mathcal{B}$ then

$$H_r(\varphi(A)|A) = 0.$$

Indeed, if $p(A = a) > 0$ then $p(\varphi(A) = \varphi(a)|A = a) = 1$, hence $\log_r \frac{1}{p(\varphi(A)=\varphi(a)|A=a)} = 0$.

Conversely, if

$$H_r(B|A) \quad = \quad 0$$

then, for all a, $p(a) = 0$, or

Conditional entropy of function

If $\varphi : \mathcal{A} \to \mathcal{B}$ then

$$H_r(\varphi(A)|A) = 0.$$

Indeed, if $p(A = a) > 0$ then $p(\varphi(A) = \varphi(a)|A = a) = 1$, hence $\log_r \frac{1}{p(\varphi(A)=\varphi(a)|A=a)} = 0$.

Conversely, if

$$H_r(B|A) \quad = \quad 0$$

then, for all a, $p(a) = 0$, or there is a **unique** b, such that $p(b|a) = 1$.

Conditional entropy of function

If $\varphi : \mathcal{A} \to \mathcal{B}$ then

$$H_r(\varphi(A)|A) = 0.$$

Indeed, if $p(A = a) > 0$ then $p(\varphi(A) = \varphi(a)|A = a) = 1$, hence $\log_r \frac{1}{p(\varphi(A)=\varphi(a)|A=a)} = 0$.

Conversely, if

$$H_r(B|A) \quad = \quad 0$$

then, for all a, $p(a) = 0$, or there is a **unique** b,

such that $p(b|a) = 1$.

Hence $B = \varphi(A)$, for some $\varphi : \mathcal{A} \to \mathcal{B}$.

Joint entropy

For $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$, let

$$(A, B)(s) \quad = \quad (A(s), B(s))\,.$$

Joint entropy

For $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$, let

$$(A, B)(s) \quad = \quad (A(s), B(s)) \, .$$

Note: $p\left((A, B) = (a, b)\right) = p\left((A = a) \wedge (B = b)\right)$.

Then

Joint entropy

For $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$, let

$$(A, B)(s) \quad = \quad (A(s), B(s)) \, .$$

Note: $p\,((A, B) = (a, b)) = p\,((A = a) \wedge (B = b))$.

Then

$$H_r(A, B) \quad = \quad \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(a \wedge b) \cdot \log_r \frac{1}{p(a \wedge b)} \, .$$

Joint entropy

For $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$, let

$$(A, B)(s) \quad = \quad (A(s), B(s)) \, .$$

Note: $p\left((A, B) = (a, b)\right) = p\left((A = a) \wedge (B = b)\right)$.

Then

$$H_r(A, B) \quad = \quad \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(a \wedge b) \cdot \log_r \frac{1}{p(a \wedge b)} .$$

If A and B are independent (i.e., $p(a \wedge b) = p(a) \cdot p(b)$),

Joint entropy

For $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$, let

$$(A, B)(s) \quad = \quad (A(s), B(s))\,.$$

Note: $p\,((A, B) = (a, b)) = p\,((A = a) \wedge (B = b))$.

Then

$$H_r(A, B) \quad = \quad \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(a \wedge b) \cdot \log_r \frac{1}{p(a \wedge b)}.$$

If A and B are independent (i.e., $p(a \wedge b) = p(a) \cdot p(b)$),

$$H_r(A, B) \quad = \quad H_r(A) + H_r(B).$$

Joint entropy

Theorem.

$$H_r(A, B) \quad \leq \quad H_r(A) + H_r(B),$$

and the equality holds if and only if A and B are independent.

Joint entropy

Theorem.

$$H_r(A,B) \quad \leq \quad H_r(A) + H_r(B),$$

and the equality holds if and only if A and B are independent.

Proof.

$$\begin{aligned}
H_r(A,B) &= \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a\wedge b)\cdot\log_r \frac{1}{p(a\wedge b)} \\
H_r(A)+H_r(B) &= \sum_{a\in\mathcal{A}} p(a)\log_r \frac{1}{p(a)} + \sum_{b\in\mathcal{B}} p(b)\log_r \frac{1}{p(b)} \\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a\wedge b)\log_r \frac{1}{p(a)} + \sum_{b\in\mathcal{B}}\sum_{a\in\mathcal{A}} p(a\wedge b)\log_r \frac{1}{p(b)} \\
&= \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a\wedge b)\log_r \frac{1}{p(a)p(b)}
\end{aligned}$$

Joint entropy

Theorem.

$$H_r(A,B) \quad \leq \quad H_r(A) + H_r(B),$$

and the equality holds if and only if A and B are independent.

Proof.

$$H_r(A,B) \quad = \quad \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a \wedge b) \cdot \log_r \frac{1}{p(a \wedge b)}$$

$$\begin{aligned} H_r(A) + H_r(B) \quad &= \quad \sum_{a\in\mathcal{A}} p(a) \log_r \frac{1}{p(a)} + \sum_{b\in\mathcal{B}} p(b) \log_r \frac{1}{p(b)} \\ &= \quad \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} p(a \wedge b) \log_r \frac{1}{p(a)} + \sum_{b\in\mathcal{B}} \sum_{a\in\mathcal{A}} p(a \wedge b) \log_r \frac{1}{p(b)} \\ &= \quad \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a \wedge b) \log_r \frac{1}{p(a)p(b)} \end{aligned}$$

Golden lemma!

Proof of $H_r(A, B) \leq H_r(A) + H_r(B)$,.

Let

$$\mathcal{A}^+ = \{a \in \mathcal{A} : p(a) > 0\}, \;\; \mathcal{B}^+ = \{b \in \mathcal{B} : p(b) > 0\}.$$

We have

Proof of $H_r(A,B) \le H_r(A) + H_r(B)$,.

Let

$$\mathcal{A}^+ = \{a \in \mathcal{A} : p(a) > 0\}, \;\; \mathcal{B}^+ = \{b \in \mathcal{B} : p(b) > 0\}.$$

We have

$$\begin{aligned} H_r(A) + H_r(B) &= \sum_{(a,b) \in \mathcal{A}^+ \times \mathcal{B}^+} p(a \wedge b) \log_r \frac{1}{p(a)p(b)} \\ H_r(A,B) &= \sum_{a \in \mathcal{A}^+, b \in \mathcal{B}^+} p(a \wedge b) \cdot \log_r \frac{1}{p(a \wedge b)}. \end{aligned}$$

Proof of $H_r(A,B) \leq H_r(A) + H_r(B)$,.

Let

$$\mathcal{A}^+ = \{a \in \mathcal{A} : p(a) > 0\}, \;\; \mathcal{B}^+ = \{b \in \mathcal{B} : p(b) > 0\}.$$

We have

$$\begin{aligned} H_r(A) + H_r(B) &= \sum_{(a,b)\in\mathcal{A}^+\times\mathcal{B}^+} p(a \wedge b) \log_r \frac{1}{p(a)p(b)} \\ H_r(A,B) &= \sum_{a\in\mathcal{A}^+, b\in\mathcal{B}^+} p(a \wedge b) \cdot \log_r \frac{1}{p(a \wedge b)}. \end{aligned}$$

Now the inequality follows from the Golden Lemma.

Proof of $H_r(A,B) \leq H_r(A) + H_r(B)$,.

Let

$$\mathcal{A}^+ = \{a \in \mathcal{A} : p(a) > 0\}, \;\; \mathcal{B}^+ = \{b \in \mathcal{B} : p(b) > 0\}.$$

We have

$$H_r(A) + H_r(B) = \sum_{(a,b)\in\mathcal{A}^+\times\mathcal{B}^+} p(a \wedge b) \log_r \frac{1}{p(a)p(b)}$$

$$H_r(A,B) = \sum_{a\in\mathcal{A}^+, b\in\mathcal{B}^+} p(a \wedge b) \cdot \log_r \frac{1}{p(a \wedge b)}.$$

Now the inequality follows from the Golden Lemma.

The **equality** holds if only if

$$p(a \wedge b) = p(a) \cdot p(b),$$

for **all** $(a,b) \in \mathcal{A}^{(+)} \times \mathcal{B}^{(+)}$, i.e. iff A and B are independent. $\square$

Mutual information

For $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$,

$$I_r(A;B) \quad = \quad H_r(A) + H_r(B) - H_r(A,B).$$

is the **mutual information** of variables A and B.

Mutual information

For $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$,

$$I_r(A; B) \quad = \quad H_r(A) + H_r(B) - H_r(A, B).$$

is the **mutual information** of variables A and B.

Note:

$$I(A; B) \quad = \quad \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(a \wedge b) \left(\log \frac{1}{p(a)p(b)} - \log \frac{1}{p(a \wedge b)} \right).$$

$\approx$ "distance from independence".

Chain rule

$$H_r(A, B) = H_r(A|B) + H_r(B).$$

Chain rule

$$H_r(A,B) \;=\; H_r(A|B) + H_r(B).$$

Proof.

$$\begin{aligned}
H(A,B) &= \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a\wedge b)\cdot\log\frac{1}{p(a\wedge b)} \\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}^+} p(a|b)p(b)\cdot\log\frac{1}{p(a|b)p(b)} \\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}^+} p(a|b)p(b)\cdot\left(\log\frac{1}{p(a|b)} + \log\frac{1}{p(b)}\right) \\
&= \sum_{b\in\mathcal{B}^+} p(b)\cdot\sum_{a\in\mathcal{A}} p(a|b)\cdot\log\frac{1}{p(a|b)} + \\
&\quad + \sum_{b\in\mathcal{B}^+} p(b)\log\frac{1}{p(b)}\cdot\underbrace{\sum_{a\in\mathcal{A}} p(a|b)}_{1} \\
&= H_r(A|B) + H_r(B) \qquad \square
\end{aligned}$$

Conditional entropy revisited

Joint entropy + chain rule:

$$\begin{aligned} H_r(A) + H_r(B) &\geq H_r(A, B) \\ &= H_r(A|B) + H_r(B) \end{aligned}$$

Conditional entropy revisited

Joint entropy + chain rule:

$$\begin{aligned} H_r(A) + H_r(B) &\geq H_r(A, B) \\ &= H_r(A|B) + H_r(B) \end{aligned}$$

Corollary

$$H_r(A|B) \leq H_r(A),$$

and the equality holds if and only if A and B are independent.

Conditional entropy revisited

Joint entropy + chain rule:

$$\begin{aligned} H_r(A) + H_r(B) &\geq H_r(A, B) \\ &= H_r(A|B) + H_r(B) \end{aligned}$$

Corollary

$$H_r(A|B) \leq H_r(A),$$

and the equality holds if and only if A and B are independent.

Note: It may be $H_r(A|B = b) > H_r(A)$, for some b.

Chain rule for $n \geq 2$

$$\begin{aligned}
H(A_1, \ldots, A_n) &= H(A_1|A_2, \ldots, A_n) + H(A_2, \ldots, A_n) \\
&= H(A_1|A_2, \ldots, A_n) + H(A_2|A_3, \ldots, A_n) + \\
&\quad + H(A_3, \ldots, A_n) \\
&= \ldots\ldots \\
&= \sum_{i=1}^{n} H(A_i|A_{i+1}, \ldots, A_n)
\end{aligned}$$

where $H(A_n|\emptyset) = H(A_n)$.

Chain rule for $n \geq 2$

$$\begin{aligned} H(A_1, \ldots, A_n) &= H(A_1|A_2, \ldots, A_n) + H(A_2, \ldots, A_n) \\ &= H(A_1|A_2, \ldots, A_n) + H(A_2|A_3, \ldots, A_n) + \\ &\quad + H(A_3, \ldots, A_n) \\ &= \ldots\ldots \\ &= \sum_{i=1}^{n} H(A_i|A_{i+1}, \ldots, A_n) \end{aligned}$$

where $H(A_n|\emptyset) = H(A_n)$.

Corollary.

$$H(A_1, \ldots, A_n) \leq H(A_1) + \ldots + H(A_n),$$

and the equality holds if and only if $A_1, \ldots, A_n$ are independent, i.e.

$$p(a_1 \wedge \ldots \wedge a_n) = p(a_1) \cdot \ldots \cdot p(a_n).$$

Conditional chain rule

$$H(A, B|C) \quad = \quad H(A|B, C) + H(B|C).$$

Proof.

Analogous to the unconditional case.

We use the fact that, whenever $p(a \wedge b|c) > 0$,

$$p(a \wedge b|c) = \frac{p(a \wedge b \wedge c)}{p(c)} = \frac{p(a \wedge b \wedge c)}{p(b \wedge c)} \cdot \frac{p(b \wedge c)}{p(c)} = p(a|b \wedge c) \cdot p(b|c).$$

Simple but tedious calculation.

□

Conditional joint entropy

Theorem.

$$H(A, B|C) \quad \leq \quad H(A|C) + H(B|C)$$

and the equality holds if and only if A and B are **conditionally independent given** C, i.e.,

$$p(A = a \wedge B = b|C = c) \quad = \quad p(A = a|C = c) \cdot p(B = b|C = c).$$

Proof.

Analogous to the unconditional case.

□

Conditional joint entropy

Theorem.

$$H(A, B|C) \quad \leq \quad H(A|C) + H(B|C)$$

and the equality holds if and only if A and B are **conditionally independent given** C, i.e.,

$$p(A = a \wedge B = b|C = c) \quad = \quad p(A = a|C = c) \cdot p(B = b|C = c).$$

Proof.

Analogous to the unconditional case.

□

Corollary.

$$H(A|B, C) \quad \leq \quad H(A|C),$$

and the equality holds iff A and B are conditionally independent given C.

Conditional information

Mutual information of A and B under condition C:

$$
\begin{aligned}
I(A;B|C) &= H(A|C) + H(B|C) - \underbrace{H(A,B|C)}_{H(A|B,C)+H(B|C)} \\
&= H(A|C) - H(A|B,C).
\end{aligned}
$$

Conditional information

Mutual information of A and B under condition C:

$$\begin{aligned} I(A;B|C) &= H(A|C) + H(B|C) - \underbrace{H(A,B|C)}_{H(A|B,C)+H(B|C)} \\ &= H(A|C) - H(A|B,C). \end{aligned}$$

Mutual information of A, B, and C:

$$R(A;B;C) = I(A;B) - I(A;B|C).$$

Conditional information

Mutual information of A and B under condition C:

$$\begin{aligned} I(A;B|C) &= H(A|C) + H(B|C) - \underbrace{H(A,B|C)}_{H(A|B,C)+H(B|C)} \\ &= H(A|C) - H(A|B,C). \end{aligned}$$

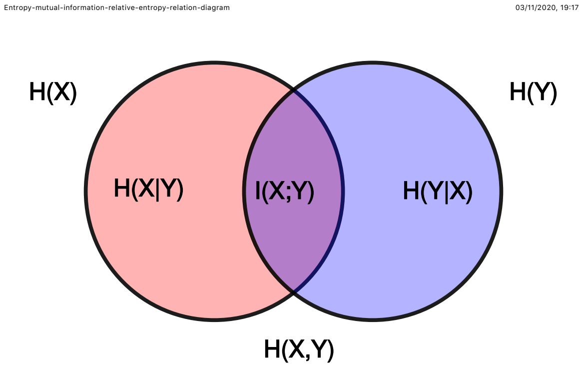
Mutual information of A, B, and C:

$$R(A;B;C) = I(A;B) - I(A;B|C).$$

Note the symmetry:

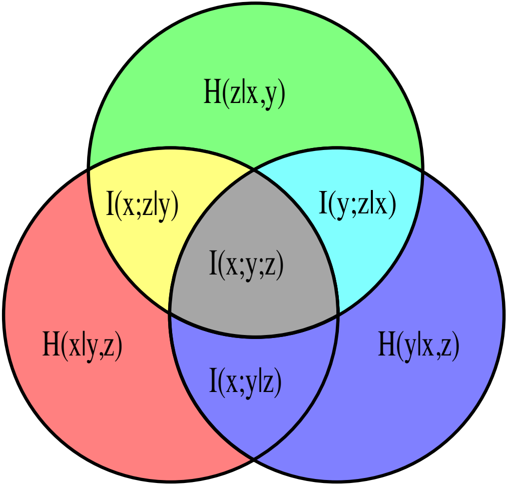
$$\begin{aligned} I(A;C) - I(A;C|B) &= H(A) - H(A|C) - (H(A|B) - H(A|B,C)) \\ &= \underbrace{H(A) - H(A|B)}_{I(A;B)} - \underbrace{(H(A|C) - H(A|B,C))}_{I(A;B|C)}. \end{aligned}$$

Venn diagram



Venn diagram

03/11/2020, 19:04



Mutual information

Note: $R(A;B;C) = I(A;B) - I(A;B|C)$ can be **negative!**

Mutual information

Note: $R(A; B; C) = I(A; B) - I(A; B|C)$ can be **negative!**

Example. Let A and B be independent random variables with values in $\{0, 1\}$, and let

$$C = A \oplus B.$$

Then $I(A; B) = 0$, while

$$I(A; B|C) = H(A|C) - \underbrace{H(A|B, C)}_{0}$$

and we can make sure that $H(A|C) > 0$, e.g.

0	0	1	1	1	1	A
0	1	0	0	1	1	B
0	1	1	1	0	0	C=A + B

Application: Perfect secrecy

A **cryptosystem** is a triple of random variables:

- M with values in $\mathcal{M}$ (messages),
- K with values in $\mathcal{K}$ (keys),
- C with values in $\mathcal{C}$ (cipher-texts),

where $\mathcal{M}, \mathcal{K}, \mathcal{C}$ are finite sets.

Application: Perfect secrecy

A **cryptosystem** is a triple of random variables:

- M with values in $\mathcal{M}$ (messages),
- K with values in $\mathcal{K}$ (keys),
- C with values in $\mathcal{C}$ (cipher-texts),

where $\mathcal{M}, \mathcal{K}, \mathcal{C}$ are finite sets.

Additionally, a function $Dec : \mathcal{C} \times \mathcal{K} \to \mathcal{M}$, such that

$$M = Dec(C, K)$$

(unique decodability).

Application: Perfect secrecy

A **cryptosystem** is a triple of random variables:

- M with values in $\mathcal{M}$ (messages),
- K with values in $\mathcal{K}$ (keys),
- C with values in $\mathcal{C}$ (cipher-texts),

where $\mathcal{M}, \mathcal{K}, \mathcal{C}$ are finite sets.

Additionally, a function $Dec : \mathcal{C} \times \mathcal{K} \to \mathcal{M}$, such that

$$M = Dec(C, K)$$

(unique decodability).

A cryptosystem is **perfectly secret** if $I(C; M) = 0$.

One time pad

Example. $\mathcal{M} = \mathcal{K} = \mathcal{C} = \{0,1\}^n$, for some $n \in \mathbb{N}$, and

$$C = M \oplus K$$

(e.g., $101101 \oplus 110110 = 011011$).

One time pad

Example. $\mathcal{M} = \mathcal{K} = \mathcal{C} = \{0,1\}^n$, for some $n \in \mathbb{N}$, and

$$C = M \oplus K$$

(e.g., $101101 \oplus 110110 = 011011$).

$$Dec(v, w) = v \oplus w.$$

K is **uniformly** distributed

$$p(K = v) = \frac{1}{2^n},$$

for $v \in \{0,1\}^n$.

K and M are **independent**.

Perfect secrecy of One time pad

$I(M; C) = 0$ iff M and C are independent, i.e.

$$p(C = w | M = u) \stackrel{?}{=} p(C = w).$$

We have

$$p(C = w) = \sum_{u \oplus v = w} p(M = u \wedge K = v) = \sum_{u} p(M = u) \cdot \frac{1}{2^n} = \frac{1}{2^n},$$

$$\begin{aligned} p(C = w | M = u) &= \frac{p(C = w \wedge M = u)}{p(M = u)} \\ &= \frac{p(K = u \oplus w \wedge M = u)}{p(M = u)} \\ &= \frac{p(K = u \oplus w) \cdot p(M = u)}{p(M = u)} \\ &= \frac{1}{2^n}. \end{aligned}$$

Why one time ?

Because C and K may be **dependent**!.

0	0	1	1	1	1	M
0	1	0	0	1	1	K
0	1	1	1	0	0	C=M+K

Why **one time** ?

Because C and K may be **dependent**!.

0	0	1	1	1	1	M
0	1	0	0	1	1	K
0	1	1	1	0	0	C=M+K

$p(K=1|C=0) = p(K=0|C=1) = \frac{2}{3}$, hence $K \approx 1 - C$.

Why one time ?

Because C and K may be **dependent**!.

0	0	1	1	1	1	M
0	1	0	0	1	1	K
0	1	1	1	0	0	C=M+K

$p(K = 1|C = 0) = p(K = 0|C = 1) = \frac{2}{3}$, hence $K \approx 1 - C$.

C.f. the American *VENONA* project (1943–1980).

Shannon's Pessimistic Theorem

Theorem. Any perfectly secret cryptosystem satisfies

$$H(K) \;\geq\; H(M).$$

Consequently

$$L_r(K) \geq H_r(K) \geq H_r(M) \geq L_r(M) - 1,$$

i.e., keys must be as long as messages (almost).

Shannon's Pessimistic Theorem

Theorem. Any perfectly secret cryptosystem satisfies

$$H(K) \quad \geq \quad H(M).$$

Shannon's Pessimistic Theorem

Theorem. Any perfectly secret cryptosystem satisfies

$$H(K) \quad \geq \quad H(M).$$

Proof.

$$H(M) \quad = \quad H(M|C,K) + \underbrace{I(M;C)}_{H(M)-H(M|C)} + \underbrace{I(M;K|C)}_{H(M|C)-H(M|K,C)} .$$

But $H(M|C;K) = 0$, since $M = Dec(C,K)$, and $I(M;C) = 0$, by assumption, hence

$$H(M) \quad = \quad I(M;K|C).$$

By symmetry, we have

$$H(K) \quad = \quad H(K|M,C) + I(K;C) + \underbrace{I(K;M|C)}_{H(M)}.$$

Can functional processing increase information ?

Maybe $I(K; C) > 0$.

Can functional processing increase information ?

Maybe $I(K; C) > 0$.

Can we increase this information, e.g., by a computation, i.e.

$$I(K; f(C)) \quad > \quad I(K; C),$$

for some f **?**

Can functional processing increase information ?

Lemma. If A and C are conditionally independent given B, then

$$I(A; C) \quad \leq \quad I(A; B).$$

Can functional processing increase information ?

Lemma. If A and C are conditionally independent given B, then

$$I(A;C) \quad \leq \quad I(A;B).$$

Proof.

$$\begin{array}{ccc} \underbrace{I(A;(B,C))}_{H(A)-H(A|B,C)} & = & \underbrace{I(A;C)}_{H(A)-H(A|C)} + \underbrace{I(A;B|C)}_{H(A|C)-H(A|B,C)} \\ \| & & \| \\ I(A;(B,C)) & = & I(A;B) + \underbrace{I(A;C|B)}_{0}. \end{array}$$

□

Can functional processing increase information ?

Lemma. If A and C are conditionally independent given B, then

$$I(A;C) \quad \leq \quad I(A;B).$$

Can functional processing increase information ?

Lemma. If A and C are conditionally independent given B, then

$$I(A;C) \quad \leq \quad I(A;B).$$

Corollary. For any function f,

$$I(A;f(B)) \quad \leq \quad I(A;B).$$

Can functional processing increase information ?

Lemma. If A and C are conditionally independent given B, then

$$I(A;C) \quad \leq \quad I(A;B).$$

Corollary. For any function f,

$$I(A;f(B)) \quad \leq \quad I(A;B).$$

Proof. Follows from the Lemma, since

$$I(A;f(B)|B) = \underbrace{H(f(B)|B)}_{0} - \underbrace{H(f(B)|A,B)}_{0} = 0.$$

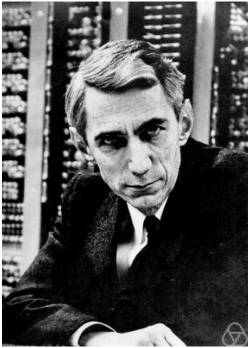
$\square$

The birth of modern information theory

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

. . . .

Claude Shannon, *A Mathematical Theory of Communication*, **1948**.



Seldom do more than a few of nature's secrets give way at one time.

Claude E. Shannon, *The Bandwagon*, 1956

Information channels

A **communication channel** Γ is given by

- a finite set $\mathcal{A}$ of **input** objects,
- a finite set $\mathcal{B}$ of **output** objects,
- a mapping $\mathcal{A} \times \mathcal{B} \ni (a, b) \mapsto P(a \to b) \in [0, 1]$, such that, for all $a \in \mathcal{A}$,

$$\sum_{b \in \mathcal{B}} P(a \to b) \quad = \quad 1.$$

Information channels

A **communication channel** Γ is given by

- a finite set $\mathcal{A}$ of **input** objects,
- a finite set $\mathcal{B}$ of **output** objects,
- a mapping $\mathcal{A} \times \mathcal{B} \ni (a, b) \mapsto P(a \to b) \in [0, 1]$, such that, for all $a \in \mathcal{A}$,

$$\sum_{b \in \mathcal{B}} P(a \to b) \quad = \quad 1.$$

Random variables A and B form an **input-output pair** for the channel Γ if, for all $a \in \mathcal{A}, b \in \mathcal{B}$,

$$p(B = b | A = a) \quad = \quad P(a \to b).$$

Information channels

$$A \to \boxed{\ \Gamma\ } \to B.$$

Recall: A and B form an **input-output pair** for Γ if $\forall a, b$,

$$p(B = b | A = a) \quad = \quad P(a \to b).$$

If it is the case then

$$p(A = a \,\wedge\, B = b) \quad = \quad P(a \to b) \cdot p(A = a).$$

Information channels

$$A \to \boxed{\ \Gamma\ } \to B.$$

Recall: A and B form an **input-output pair** for Γ if $\forall a, b$,

$$p(B = b | A = a) \quad = \quad P(a \to b).$$

If it is the case then

$$p(A = a \wedge B = b) \quad = \quad P(a \to b) \cdot p(A = a).$$

Therefore the distribution of (A, B) is uniquely determined by A and Γ, and B satisfies

Information channels

$$A \rightarrow \boxed{\ \Gamma\ } \rightarrow B.$$

Recall: A and B form an **input-output pair** for Γ if $\forall a, b$,

$$p(B = b | A = a) \quad = \quad P(a \rightarrow b).$$

If it is the case then

$$p(A = a \ \wedge \ B = b) \quad = \quad P(a \rightarrow b) \cdot p(A = a).$$

Therefore the distribution of (A, B) is uniquely determined by A and Γ, and B satisfies

$$p(B = b) \quad = \quad \sum_{a \in \mathcal{A}} P(a \rightarrow b) \cdot p(A = a).$$

Channel capacity

The **capacity** of a channel Γ is

$$C_\Gamma \quad = \quad \max_A I_2(A;B),$$

where, (A, B) ranges over all input-output pair for Γ.

Channel capacity

The **capacity** of a channel Γ is

$$C_\Gamma \quad = \quad \max_A I_2(A; B),$$

where, (A, B) ranges over all input-output pair for Γ.

The maximum exists because $I(A; B)$ is a continuous mapping from a compact set

$$\left\{ p \in [0,1]^{\mathcal{A}} : \sum_{a \in \mathcal{A}} p(a) = 1 \right\} \to \mathbb{R},$$

which is bounded since $I(A; B) \leq H(A) \leq \log |\mathcal{A}|$.

Matrix representation

$$\Gamma = \begin{pmatrix} P_{11} & \dots & P_{1n} \\ \dots & \dots & \dots \\ P_{m1} & \dots & P_{mn}, \end{pmatrix}$$

where $P_{ij} = P(a_i \to b_j)$.

Matrix representation

$$\Gamma = \begin{pmatrix} P_{11} & \dots & P_{1n} \\ \dots & \dots & \dots \\ P_{m1} & \dots & P_{mn}, \end{pmatrix}$$

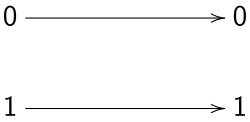
where $P_{ij} = P(a_i \to b_j)$.

Computing distribution of B from distribution of A

$$(p(a_1), \dots, p(a_m)) \cdot \begin{pmatrix} P_{11} & \dots & P_{1n} \\ \dots & \dots & \dots \\ P_{m1} & \dots & P_{mn}, \end{pmatrix} \quad = \quad (p(b_1), \dots, p(b_n))\,.$$

Examples

Faithful (noiseless) channel



Examples

Faithful (noiseless) channel

$$0 \longrightarrow 0$$

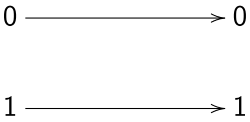
$$1 \longrightarrow 1$$

The matrix representation

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Examples

Faithful (noiseless) channel



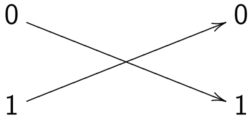
The matrix representation

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$C_\Gamma = \max_A \underbrace{I(A;B)}_{H(A)} = \log_2 |\mathcal{A}| = 1,$$

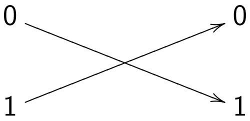
since A is a function of B.

Inverse faithful channel



$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Inverse faithful channel

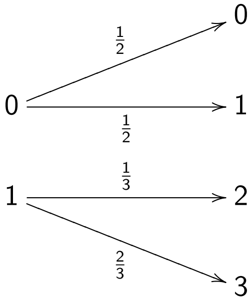


$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$C_\Gamma = \max_A \underbrace{I(A;B)}_{H(A)} = 1,$$

Noisy channel without overlap

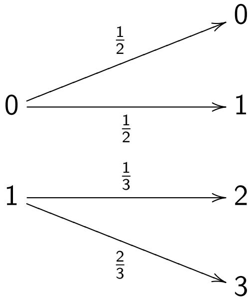
$\mathcal{A} = \{0, 1\}$, $\mathcal{B} = \{0, 1, 2, 3\}$.



$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

Noisy channel without overlap

$\mathcal{A} = \{0, 1\}$, $\mathcal{B} = \{0, 1, 2, 3\}$.



$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

$$C_\Gamma = \max_A \underbrace{I(A; B)}_{H(A)} = 1,$$

Noisy typewriter

$\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters)

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = 0.5$$

where $next(a) = b$, $next(b) = c$, $\ldots$, $next(y) = z$, $next(z) = a$.

Noisy typewriter

$\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters)

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = 0.5$$

where $next(a) = b$, $next(b) = c$, $\ldots$, $next(y) = z$, $next(z) = a$.

$$\begin{pmatrix} 0.5 & 0 & 0 & \ldots & 0.5 \\ 0.5 & 0.5 & 0 & \ldots & 0 \\ 0 & 0.5 & 0.5 & \ldots & 0 \\ 0 & 0 & 0.5 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0.5 \end{pmatrix}$$

Noisy typewriter

$\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters)

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = 0.5$$

where $next(a) = b$, $next(b) = c$, $\ldots$, $next(y) = z$, $next(z) = a$.

$$\begin{pmatrix} 0.5 & 0 & 0 & \ldots & 0.5 \\ 0.5 & 0.5 & 0 & \ldots & 0 \\ 0 & 0.5 & 0.5 & \ldots & 0 \\ 0 & 0 & 0.5 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0.5 \end{pmatrix}$$

$$C_\Gamma = \max_A I(A;B) = \max_A H(B) - \underbrace{H(B|A)}_{1} = \log 26 - 1 = \log 13,$$

the maximum for A uniform, which causes B uniform as well, because the columns sum up to 1.

Bad channels

$C_\Gamma = 0$ iff $I(A;B) = 0$, for all input-output pairs, i.e.,

$$\underbrace{p(B = b|A = a)}_{P(a \to b)} \quad = \quad p(B = b),$$

for all $a \in \mathcal{A}$, $b \in \mathcal{B}$ (unless $p(A = a) = 0$).

Bad channels

$C_\Gamma = 0$ iff $I(A;B) = 0$, for all input-output pairs, i.e.,

$$\underbrace{p(B=b|A=a)}_{P(a\to b)} = p(B=b),$$

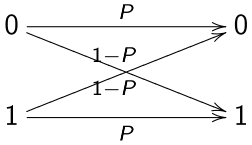
for all $a \in \mathcal{A}$, $b \in \mathcal{B}$ (unless $p(A=a)=0$).

That is, the values within a column must be equal.

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{6} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{6} & \frac{1}{3} \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Binary symmetric channel (BSC)

$\mathcal{A} = \mathcal{B} = \{0, 1\}$.



Letting $\bar{P} = 1 - P$,

$$\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$$

Fact. Any input-output pair (A, B) satisfies

$$H(B) \geq H(A),$$

with the equality if $P \in \{0, 1\}$ or if $H(A) = 1$.

For $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$, $H(B) \geq H(A)$. **Proof.**

Let	$p(A=0)=q$	$p(A=1)=\bar{q}$,
compute	$p(B=0)=r$	$p(B=1)=\bar{r}$.

For $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$, $H(B) \geq H(A)$. **Proof.**

Let $p(A=0) = q \qquad p(A=1) = \bar{q}$,

compute $p(B=0) = r \qquad p(B=1) = \bar{r}$.

$$(q, \bar{q}) \cdot \begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix} = (\underbrace{qP + \bar{q}\bar{P}}_{r}, \underbrace{q\bar{P} + \bar{q}P}_{\bar{r}})$$

For $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$, $H(B) \geq H(A)$. **Proof.**

Let compute

$$p(A=0) = q \qquad p(A=1) = \bar{q},$$
$$p(B=0) = r \qquad p(B=1) = \bar{r}.$$

$$(q, \bar{q}) \cdot \begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix} = (\underbrace{qP + \bar{q}\bar{P}}_{r}, \underbrace{q\bar{P} + \bar{q}P}_{\bar{r}})$$

Then

$$\begin{aligned} H(A) &= -q \log q - \bar{q} \log \bar{q} \\ H(B) &= -r \log r - \bar{r} \log \bar{r} \end{aligned}$$

For $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$, $H(B) \geq H(A)$. **Proof.**

Let $p(A=0) = q$ $\quad p(A=1) = \bar{q}$,

compute $p(B=0) = r$ $\quad p(B=1) = \bar{r}$.

$$(q, \bar{q}) \cdot \begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix} = (\underbrace{qP + \bar{q}\bar{P}}_{r}, \underbrace{q\bar{P} + \bar{q}P}_{\bar{r}})$$

Then
$$\begin{aligned} H(A) &= -q \log q - \bar{q} \log \bar{q} \\ H(B) &= -r \log r - \bar{r} \log \bar{r} \end{aligned}$$

The function $x \log_2 x + (1-x)\log_2(1-x)$ is strictly convex.

Taking $x_1 = q$, $x_2 = \bar{q}$, $r = Px_1 + \bar{P}x_2$,

$$\begin{aligned} P \cdot (q \log q + \bar{q} \log \bar{q}) + \bar{P} \cdot (q \log q + \bar{q} \log \bar{q}) &\geq r \log r + \bar{r} \log \bar{r} \\ \text{i.e., } H(A) &\leq H(B), \end{aligned}$$

with the equality if $P \in \{0, 1\}$ or $q = \bar{q}$. □

Binary symmetric channel $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$

Computing the capacity.

Binary symmetric channel $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$

Computing the capacity.

$$\begin{aligned} H(B|A) &= (p(A=0)+p(A=1)) \cdot \\ &\quad \cdot \left(p(s|s) \cdot \log \frac{1}{p(s|s)} + p(\bar{s}|s) \cdot \log \frac{1}{p(\bar{s}|s)} \right) \\ &= P \cdot \log \frac{1}{P} + \bar{P} \cdot \log \frac{1}{\bar{P}}. \end{aligned}$$

Binary symmetric channel $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$

Computing the capacity.

$$\begin{aligned} H(B|A) &= (p(A=0)+p(A=1)) \cdot \\ &\quad \cdot \left(p(s|s) \cdot \log \frac{1}{p(s|s)} + p(\bar{s}|s) \cdot \log \frac{1}{p(\bar{s}|s)} \right) \\ &= P \cdot \log \frac{1}{P} + \bar{P} \cdot \log \frac{1}{\bar{P}}. \end{aligned}$$

Letting $H(s) = -s \log_2 s - (1-s) \log_2 (1-s)$,

$$C_\Gamma = \max_A H(B) - H(B|A) = 1 - H(P),$$

achieved for A with uniform distribution.

Binary symmetric channel $\begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$

Computing the capacity.

$$\begin{aligned} H(B|A) &= (p(A=0)+p(A=1)) \cdot \\ &\quad \cdot \left(p(s|s) \cdot \log \frac{1}{p(s|s)} + p(\bar{s}|s) \cdot \log \frac{1}{p(\bar{s}|s)} \right) \\ &= P \cdot \log \frac{1}{P} + \bar{P} \cdot \log \frac{1}{\bar{P}}. \end{aligned}$$

Letting $H(s) = -s \log_2 s - (1-s) \log_2 (1-s)$,

$$C_\Gamma = \max_A H(B) - H(B|A) = 1 - H(P),$$

achieved for A with uniform distribution.

Note: $0 \le C_\Gamma \le 1$ (bounds achieved for $P \in \{0, \frac{1}{2}, 1\}$).

Shannon's scheme

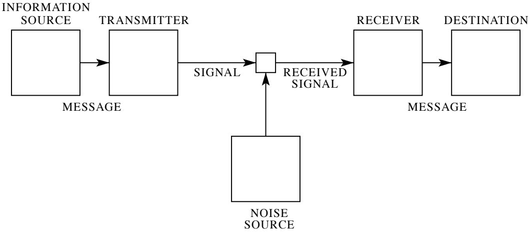


Fig. 1 — Schematic diagram of a general communication system.

a decimal digit is about $3\frac{1}{3}$ bits. A digit wheel on a desk computing machine has ten stable positions and therefore has a storage capacity of one decimal digit. In analytical work where integration and differentiation are involved the base e is sometimes useful. The resulting units of information will be called natural units. Change from the base a to base b merely requires multiplication by $\log_b a$.

By a communication system we will mean a system of the type indicated schematically in Fig. 1. It consists of essentially five parts:

Decision rules

A mapping $\Delta : \mathcal{B} \to \mathcal{A}$ chosen to maximise $p(A = \Delta(b)|B = b)$.

Decision rules

A mapping $\Delta : \mathcal{B} \to \mathcal{A}$ chosen to maximise $p(A = \Delta(b)|B = b)$.

The **quality of the rule** is measured by

$$Pr_C(\Delta, A) \stackrel{def}{=} p(\Delta \circ B = A).$$

Decision rules

A mapping $\Delta : \mathcal{B} \to \mathcal{A}$ chosen to maximise $p(A = \Delta(b) | B = b)$.

The **quality of the rule** is measured by

$$Pr_C(\Delta, A) \stackrel{def}{=} p(\Delta \circ B = A).$$

$$\begin{aligned}
&= \sum_{b \in \mathcal{B}} p(B = b \wedge A = \Delta(b)) \\
&= \sum_{b \in \mathcal{B}} p(B = b) \cdot p(A = \Delta(b) | B = b) \\
&= \sum_{b \in \mathcal{B}} p(A = \Delta(b)) \cdot p(B = b | A = \Delta(b)) \\
&= \sum_{a \in \mathcal{A}} p(A = a) \cdot p(\Delta(B) = a | A = a).
\end{aligned}$$

Decision rules

Dually, the **error probability** of the rule Δ is

$$\begin{aligned} Pr_E(\Delta, A) &= 1 - Pr_C(\Delta, A) \\ &= \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(A = a \wedge B = b \wedge \Delta(b) \neq a) \\ &= \sum_{a \in \mathcal{A}} p(A = a) \cdot p(\Delta \circ B \neq a | A = a) \end{aligned}$$

Ideal observer rule

Dedicated to A,

$\mathcal{B} \ni b \mapsto \Delta_o(b) = a \in \mathcal{A}$, maximising

$$p(a|b) = \frac{p(a \wedge b)}{p(b)} = \frac{P(a \to b) \cdot p(a)}{\sum_{a' \in \mathcal{A}} P(a' \to b) \cdot p(a')}.$$

Maximal likelihood rule

If we don't know A,

$\mathcal{B} \ni b \mapsto \Delta_{\mathrm{max}}(b) = a \in \mathcal{A}$, maximising

$$p(b|a) \quad = \quad P(a \to b).$$

Maximal likelihood rule

If we don't know A,

$\mathcal{B} \ni b \mapsto \Delta_{\max}(b) = a \in \mathcal{A}$, maximising

$$p(b|a) = P(a \to b).$$

Note: If A has uniform distribution then

$$Pr_C(\Delta_{\max}, A) = Pr_C(\Delta_o, A)$$

Maximal likelihood rule

If we don't know A,

$\mathcal{B} \ni b \mapsto \Delta_{\max}(b) = a \in \mathcal{A}$, maximising

$$p(b|a) \quad = \quad P(a \to b).$$

Note: If A has uniform distribution then

$$Pr_C(\Delta_{\max}, A) \quad = \quad Pr_C(\Delta_o, A)$$

($\Delta_{\max} = \Delta_o$ if they agree on multiple choices).

Maximal likelihood rule

If we don't know A,

$\mathcal{B} \ni b \mapsto \Delta_{\max}(b) = a \in \mathcal{A}$, maximising

$$p(b|a) \quad = \quad P(a \to b).$$

Note: If A has uniform distribution then

$$Pr_C(\Delta_{\max}, A) \quad = \quad Pr_C(\Delta_o, A)$$

($\Delta_{\max} = \Delta_o$ if they agree on multiple choices).

Indeed, for $b \in \mathcal{B}$, both rules maximise

$$p(a|b) \cdot p(b) = p(a \wedge b) = P(a \to b) \cdot \frac{1}{|\mathcal{A}|}.$$

Maximal likelihood rule

Global optimality. Let

$$\begin{aligned} \mathcal{P} &= \left\{ \mathbf{p} : \sum_{a \in \mathcal{A}} \mathbf{p}(a) = 1 \right\} \\ \mathbf{p}(a) &= p(A = a). \end{aligned}$$

Maximal likelihood rule

Global optimality. Let

$$\begin{aligned} \mathcal{P} &= \left\{ \mathbf{p} : \sum_{a \in \mathcal{A}} \mathbf{p}(a) = 1 \right\} \\ \mathbf{p}(a) &= p(A = a). \end{aligned}$$

Then

$$\begin{aligned} \int_{\mathbf{p} \in \mathcal{P}} Pr_C(\Delta, \mathbf{p})\, d\mathbf{p} &= \int_{\mathbf{p} \in \mathcal{P}} \sum_{b \in \mathcal{B}} \mathbf{p}(\Delta(b)) \cdot P(\Delta(b) \to b)\, d\mathbf{p} \\ &= \sum_{b \in \mathcal{B}} P(\Delta(b) \to b) \cdot \int_{\mathbf{p} \in \mathcal{P}} \mathbf{p}(\Delta(b))\, d\mathbf{p} \end{aligned}$$

Maximal likelihood rule

Global optimality. Let

$$\begin{aligned} \mathcal{P} &= \left\{ \mathbf{p} : \sum_{a \in \mathcal{A}} \mathbf{p}(a) = 1 \right\} \\ \mathbf{p}(a) &= p(A = a). \end{aligned}$$

Then

$$\begin{aligned} \int_{\mathbf{p} \in \mathcal{P}} Pr_C(\Delta, \mathbf{p}) \, d\mathbf{p} &= \int_{\mathbf{p} \in \mathcal{P}} \sum_{b \in \mathcal{B}} \mathbf{p}(\Delta(b)) \cdot P(\Delta(b) \to b) \, d\mathbf{p} \\ &= \sum_{b \in \mathcal{B}} P(\Delta(b) \to b) \cdot \int_{\mathbf{p} \in \mathcal{P}} \mathbf{p}(\Delta(b)) \, d\mathbf{p} \end{aligned}$$

Maximal for $\Delta = \Delta_{\text{max}}$.

Multiple use of channel

$$A_1, A_2, \dots A_k \rightarrow \boxed{\Gamma} \rightarrow B_1, B_2, \dots B_k$$

Multiple use of channel

$$A_1, A_2, \dots A_k \rightarrow \boxed{\Gamma} \rightarrow B_1, B_2, \dots B_k$$

$$p\left(b_1, b_2, \dots b_k \mid a_1, a_2 \dots a_k\right) \quad = \quad ?$$

Multiple use of channel

$$A_1, A_2, \ldots A_k \rightarrow \boxed{\Gamma} \rightarrow B_1, B_2, \ldots B_k$$

$$p\left(b_1, b_2, \ldots b_k \mid a_1, a_2 \ldots a_k\right) \quad = \quad ?$$

$$? \quad = \quad p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \ldots \cdot p(b_k \mid a_k)$$

Multiple use of channel

$$A_1, A_2, \dots A_k \to \boxed{\Gamma} \to B_1, B_2, \dots B_k$$

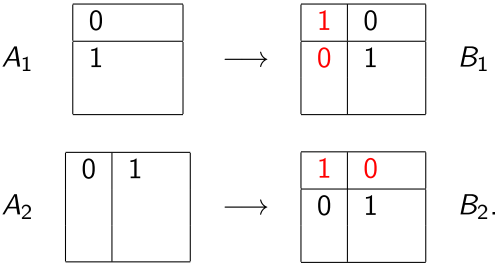
$$p\,(b_1, b_2, \dots b_k \mid a_1, a_2 \dots a_k) \quad = \quad ?$$

$$? \quad = \quad p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \ldots \cdot p(b_k \mid a_k)$$

Is it enough that $A_1, \dots, A_k$ are independent?

Multiple use of channel $\begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$.

$$p\,(b_1, b_2 \mid a_1, a_2) \stackrel{?}{=} p(b_1 \mid a_1) \cdot p(b_2 \mid a_2)$$

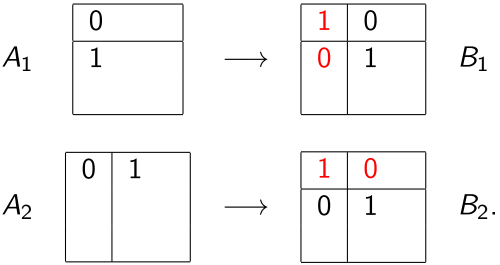


A_1 and A_2 are independent, with $A_i(0) = \frac{1}{3}$, $A_i(1) = \frac{2}{3}$.

B_1 and B_2 are identical.

Multiple use of channel $\begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$.

$$p\left(b_1, b_2 \mid a_1, a_2\right) \stackrel{?}{=} p(b_1 \mid a_1) \cdot p(b_2 \mid a_2)$$



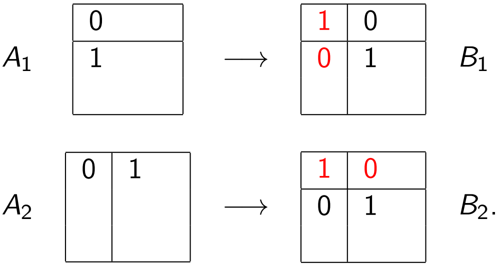
A_1 and A_2 are independent, with $A_i(0) = \frac{1}{3}$, $A_i(1) = \frac{2}{3}$.

B_1 and B_2 are identical.

$p(11|00) = p(00|01) = p(00|10) = p(11|11) = 1$ (!)

Multiple use of channel $\begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$.

$$p\left(b_1, b_2 \mid a_1, a_2\right) \neq p(b_1 \mid a_1) \cdot p(b_2 \mid a_2)$$

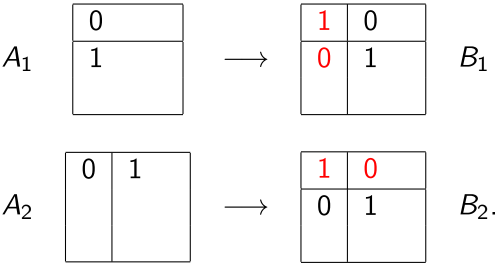


A_1 and A_2 are independent, with $A_i(0) = \frac{1}{3}$, $A_i(1) = \frac{2}{3}$.

B_1 and B_2 are identical.

Multiple use of channel $\begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$.

$$p\left(b_1, b_2 \mid a_1, a_2\right) \neq p(b_1 \mid a_1) \cdot p(b_2 \mid a_2)$$



A_1 and A_2 are independent, with $A_i(0) = \frac{1}{3}$, $A_i(1) = \frac{2}{3}$.

B_1 and B_2 are identical.

$p(11|00) = p(00|01) = p(00|10) = p(11|11) = 1$.

Multiple use of channel $\begin{pmatrix} 1/2 & 1/2 \\ 1/5 & 4/5 \end{pmatrix}$.

$$p\left(b_1, b_2 \mid a_1, a_2\right) \stackrel{?}{=} p(b_1 \mid a_1) \cdot p(b_2 \mid a_2)$$

The independence of $B_1, B_2, \ldots$ does not suffice either.

A_1

1	0
0	1

$\longrightarrow$

0	0
1	1

B_1

A_2

1	0
0	1

$\longrightarrow$

0	1
0	1

B_2

Multiple use of channel $\begin{pmatrix} 1/2 & 1/2 \\ 1/5 & 4/5 \end{pmatrix}$.

$$p\left(b_1, b_2 \mid a_1, a_2\right) \overset{?}{=} p(b_1 \mid a_1) \cdot p(b_2 \mid a_2)$$

The independence of $B_1, B_2, \ldots$ does not suffice either.

A_1

1	0
0	1

$\longrightarrow$

0	0
1	1

B_1

A_2

1	0
0	1

$\longrightarrow$

0	1
0	1

B_2

Here A_1 and A_2 are identical, hence obviously $p(x^n \mid y^n) = p(x|y)$, for any pair of symbols x, y. In particular

$p(00|11) = \frac{1}{9} : \frac{5}{9} = \frac{1}{5}$, whereas
$p(0|1) \cdot p(0|1) = \frac{1}{5} \cdot \frac{1}{5} = \frac{1}{25}$.

Multiple use of channel

$$A_1, A_2, \ldots A_k \rightarrow \boxed{\Gamma} \rightarrow B_1, B_2, \ldots B_k$$

independence of symbols

$$p\left(b_1, b_2, \ldots b_k \mid a_1, a_2 \ldots a_k\right) \quad = \quad p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \ldots \cdot p(b_k \mid a_k)$$

Multiple use of channel

$$A_1, A_2, \ldots A_k \rightarrow \boxed{\Gamma} \rightarrow B_1, B_2, \ldots B_k$$

independence of symbols

$$p\left(b_1, b_2, \ldots b_k \mid a_1, a_2 \ldots a_k\right) = p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \ldots \cdot p(b_k \mid a_k)$$

no memory

$$p(b_k | a_1 \ldots a_k, b_1 \ldots b_{k-1}) = p(b_k | a_k)$$

Multiple use of channel

$$A_1, A_2, \dots A_k \to \boxed{\Gamma} \to B_1, B_2, \dots B_k$$

independence of symbols

$$p\left(b_1, b_2, \dots b_k \mid a_1, a_2 \dots a_k\right) \;=\; p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \ldots \cdot p(b_k \mid a_k)$$

no memory

$$p(b_k | a_1 \dots a_k, b_1 \dots b_{k-1}) \;=\; p(b_k | a_k)$$

no feedback

$$p(a_k | a_1 \dots a_{k-1}, b_1 \dots b_{k-1}) \;=\; p(a_k | a_1 \dots a_{k-1})$$

Multiple use of channel

$$A_1, A_2, \dots A_k \to \boxed{\Gamma} \to B_1, B_2, \dots B_k$$

independence of symbols

$$p\left(b_1, b_2, \dots b_k \mid a_1, a_2 \dots a_k\right) = p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \dots \cdot p(b_k \mid a_k)$$

no memory

$$p(b_k | a_1 \dots a_k, b_1 \dots b_{k-1}) = p(b_k | a_k)$$

no feedback

$$p(a_k | a_1 \dots a_{k-1}, b_1 \dots b_{k-1}) = p(a_k | a_1 \dots a_{k-1})$$

Hold if $(A_1, B_1), \dots, (A_k, B_k)$ are independent.

Multiple use of channel

Theorem.

Independence of symbols $\iff$ no memory and no feedback.

Multiple use of channel

Theorem.

Independence of symbols $\iff$ **no memory and no feedback.**

Note. The conditions are indeed weaker than the independence of $(A_1, B_1), \ldots, (A_k, B_k)$.

Multiple use of channel

Theorem.

Independence of symbols $\iff$ **no memory and no feedback.**

Note. The conditions are indeed weaker than the independence of $(A_1, B_1), \ldots, (A_k, B_k)$.

For example, they hold for the faithfull channel, for any sequence $A_1, \ldots, A_k$.

Proof

$$\left.\begin{array}{rcl} p(b_k \mid a_1 \ldots a_k, b_1 \ldots b_{k-1}) & = & p(b_k|a_k) \\ p(a_k \mid a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) & = & p(a_k|a_1 \ldots a_{k-1}) \end{array}\right\} \Longrightarrow$$

$$p(a_1 \ldots a_k, b_1 \ldots b_k) = p(b_1|a_1) \cdot \ldots \cdot p(b_k|a_k) \cdot \underbrace{p(a_1 \ldots a_k)}_{>0},$$

For the induction step,

$$p(a_1 \ldots a_k, b_1 \ldots b_k) \quad = \quad \underbrace{p(b_k|a_k)}_{\textbf{no mem.}} \cdot \underbrace{p(a_1 \ldots a_k, b_1 \ldots b_{k-1})}_{\|},$$

Proof

$$\left.\begin{aligned} p(b_k \mid a_1 \dots a_k, b_1 \dots b_{k-1}) &= p(b_k|a_k) \\ p(a_k \mid a_1 \dots a_{k-1}, b_1 \dots b_{k-1}) &= p(a_k|a_1 \dots a_{k-1}) \end{aligned}\right\} \Longrightarrow$$

$$p(a_1 \dots a_k, b_1 \dots b_k) = p(b_1|a_1) \cdot \dots \cdot p(b_k|a_k) \cdot \underbrace{p(a_1 \dots a_k)}_{>0},$$

For the induction step,

$$p(a_1 \dots a_k, b_1 \dots b_k) = \underbrace{p(b_k|a_k)}_{\textbf{no mem.}} \cdot \underbrace{p(a_1 \dots a_k, b_1 \dots b_{k-1})}_{\|},$$

$$\underbrace{p(a_1 \dots a_{k-1}, b_1 \dots b_{k-1})}_{\|\textbf{ind.}} \cdot \left.\frac{p(a_1 \dots a_k)}{p(a_1 \dots a_{k-1})}\right\}_{\textbf{no feed.}}$$

Proof

$$\left.\begin{array}{rcl} p(b_k \mid a_1 \dots a_k, b_1 \dots b_{k-1}) & = & p(b_k|a_k) \\ p(a_k \mid a_1 \dots a_{k-1}, b_1 \dots b_{k-1}) & = & p(a_k|a_1 \dots a_{k-1}) \end{array}\right\} \Longrightarrow$$

$$p(a_1 \dots a_k, b_1 \dots b_k) = p(b_1|a_1) \cdot \dots \cdot p(b_k|a_k) \cdot \underbrace{p(a_1 \dots a_k)}_{>0},$$

For the induction step,

$$p(a_1 \dots a_k, b_1 \dots {\color{red}b_k}) = \underbrace{p({\color{red}b_k}|a_k)}_{\textbf{no mem.}} \cdot \underbrace{p(a_1 \dots {\color{purple}a_k}, b_1 \dots b_{k-1})}_{\|},$$

$$\underbrace{p(a_1 \dots a_{k-1}, b_1 \dots b_{k-1})}_{\|\textbf{ind.}} \cdot \left.\frac{p(a_1 \dots {\color{purple}a_k})}{p(a_1 \dots a_{k-1})}\right\}_{\textbf{no feed.}}$$

$$p(b_1|a_1) \cdot \dots \cdot p(b_{k-1}|a_{k-1}) \cdot p(a_1 \dots a_{k-1}),$$

if $p(a_1 \dots a_k, b_1 \dots b_{k-1}) > 0$.

Remaining case of $\boxed{p(a_1 \ldots a_{k-1}, a_k, b_1 \ldots b_{k-1}) = 0}$.

(By assumption, $p(a_1 \ldots a_k) \neq 0$.)

Remaining case of $\boxed{p(a_1 \ldots a_{k-1}, a_k, b_1 \ldots b_{k-1}) = 0}$.

(By assumption, $p(a_1 \ldots a_k) \neq 0$.)

If $p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) = 0$, we have, by induction hypothesis,

$$\underbrace{p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1})}_{=0} = \underbrace{p(b_1|a_1) \cdot \ldots \cdot p(b_{k-1}|a_{k-1})}_{=0} \cdot p(a_1 \ldots a_{k-1})$$

$$\|$$

Remaining case of $\boxed{p(a_1 \ldots a_{k-1}, a_k, b_1 \ldots b_{k-1}) = 0}$.

(By assumption, $p(a_1 \ldots a_k) \neq 0$.)

If $p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) = 0$, we have, by induction hypothesis,

$$\begin{array}{ccc}\underbrace{p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1})}_{=0} & = & \underbrace{p(b_1|a_1) \cdot \ldots \cdot p(b_{k-1}|a_{k-1})}_{=0} \cdot p(a_1 \ldots a_{k-1}) \\ \| & & \\ p(a_1 \ldots a_k, b_1 \ldots b_k) & = & p(b_1|a_1) \cdot \ldots \cdot p(b_k|a_k) \cdot p(a_1 \ldots a_k).\end{array}$$

Remaining case of $\boxed{p(a_1 \ldots a_{k-1}, a_k, b_1 \ldots b_{k-1}) = 0}$.

(By assumption, $p(a_1 \ldots a_k) \neq 0$.)

If $p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) = 0$, we have, by induction hypothesis,

$$\begin{array}{ccc} \underbrace{p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1})}_{=0} & = & \underbrace{p(b_1|a_1) \cdot \ldots \cdot p(b_{k-1}|a_{k-1})}_{=0} \cdot p(a_1 \ldots a_{k-1}) \\ \| & & \\ p(a_1 \ldots a_k, b_1 \ldots b_k) & = & p(b_1|a_1) \cdot \ldots \cdot p(b_k|a_k) \cdot p(a_1 \ldots a_k). \end{array}$$

If $p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) > 0$, we have

$$0 = \underbrace{p(a_k|a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1})}_{\textbf{well defined}} \overset{\textbf{no feed.}}{=} p(a_k|a_1 \ldots a_{k-1}),$$

which contradicts the assumption that $p(a_1 \ldots a_k) > 0$.

Remaining case of $\boxed{p(a_1 \ldots a_{k-1}, a_k, b_1 \ldots b_{k-1}) = 0}$.

(By assumption, $p(a_1 \ldots a_k) \neq 0$.)

If $p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) = 0$, we have, by induction hypothesis,

$$\underbrace{p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1})}_{=0} = \underbrace{p(b_1|a_1) \cdot \ldots \cdot p(b_{k-1}|a_{k-1})}_{=0} \cdot p(a_1 \ldots a_{k-1})$$

$$\|$$

$$p(a_1 \ldots a_k, b_1 \ldots b_k) = p(b_1|a_1) \cdot \ldots \cdot p(b_k|a_k) \cdot p(a_1 \ldots a_k).$$

If $p(a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) > 0$, we have

$$0 = \underbrace{p(a_k|a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1})}_{\textbf{well defined}} \overset{\textbf{no feed.}}{=} p(a_k|a_1 \ldots a_{k-1}),$$

which contradicts the assumption that $p(a_1 \ldots a_k) > 0$.

For the proof of "$\Longleftarrow$" see Lecture notes. □

Multiple use of channel

Proviso.

If not stated otherwise, we assume that the independence of symbols property

$$p\left(b_1, b_2, \dots b_k \mid a_1, a_2 \dots a_k\right) = p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \dots \cdot p(b_k \mid a_k)$$

always holds.

BSC revisited

Let $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

Then $\Delta_{\max}(i) =$

BSC revisited

Let $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

Then $\Delta_{\max}(i) = i$, for $i = 0, 1$, and, for any A,

$$\begin{aligned} Pr_C(\Delta_{\max}, A) &= \sum_{b \in \{0,1\}} p(\Delta_{\max}(b)) \cdot p(\Delta_{\max}(b) \to b) \\ &= p(A = 0) \cdot P + p(A = 1) \cdot P \\ &= P, \end{aligned}$$

hence

BSC revisited

Let $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

Then $\Delta_{\max}(i) = i$, for $i = 0, 1$, and, for any A,

$$\begin{aligned} Pr_C(\Delta_{\max}, A) &= \sum_{b \in \{0,1\}} p(\Delta_{\max}(b)) \cdot p(\Delta_{\max}(b) \to b) \\ &= p(A = 0) \cdot P + p(A = 1) \cdot P \\ &= P, \end{aligned}$$

hence

$$Pr_E(\Delta_{\max}, A) = Q$$

BSC revisited

Let $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

Then $\Delta_{\max}(i) = i$, for $i = 0, 1$, and, for any A,

$$\begin{aligned} Pr_C(\Delta_{\max}, A) &= \sum_{b \in \{0,1\}} p(\Delta_{\max}(b)) \cdot p(\Delta_{\max}(b) \to b) \\ &= p(A=0) \cdot P + p(A=1) \cdot P \\ &= P, \end{aligned}$$

hence

$$\begin{aligned} Pr_E(\Delta_{\max}, A) &= Q \\ &\overset{\textbf{short}.}{=} Pr_E(\Delta_{\max}). \end{aligned}$$

Improving reliability – redundancy

Improving reliability – redundancy

I LOVE YOU.

Improving reliability – redundancy

I LOVE YOU.

↓

Improving reliability – redundancy

I LOVE YOU.

$\downarrow$

III LLLOOOOOOOOVVVVEEE YYYYOOOOOOOOUUUU.

Improving reliability

For $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

Improving reliability

For $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

$$\begin{array}{ccccccccccc} 0 & \mapsto & 000 & \rightarrow & \boxed{\Gamma} & \rightarrow & 000 & 001 & 010 & 100 & \mapsto & 0 \\ 1 & \mapsto & 111 & \rightarrow & & \rightarrow & 011 & 101 & 110 & 111 & \mapsto & 1 \end{array}$$

Improving reliability

For $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

$$\begin{array}{ccccccccccccc}
0 & \mapsto & 000 & \rightarrow & \boxed{\Gamma} & \rightarrow & 000 & 001 & 010 & 100 & \mapsto & 0 \\
1 & \mapsto & 111 & \rightarrow & & \rightarrow & 011 & 101 & 110 & 111 & \mapsto & 1
\end{array}$$

$$\begin{array}{ccc}
\rightarrow & \boxed{\Gamma'} & \rightarrow \\
\rightarrow & & \rightarrow
\end{array}$$

Improving reliability

For $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

$$\begin{array}{ccccccccccccc} 0 & \mapsto & 000 & \rightarrow & \boxed{\Gamma} & \rightarrow & 000 & 001 & 010 & 100 & \mapsto & 0 \\ 1 & \mapsto & 111 & \rightarrow & & \rightarrow & 011 & 101 & 110 & 111 & \mapsto & 1 \end{array}$$

$$\begin{array}{ccc} \rightarrow & \boxed{\Gamma'} & \rightarrow \\ \rightarrow & & \rightarrow \end{array}$$

where

$$\Gamma' = \begin{pmatrix} P^3 + 3P^2Q & Q^3 + 3Q^2P \\ Q^3 + 3Q^2P & P^3 + 3P^2Q \end{pmatrix}.$$

Improving reliability

For $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, with $P > Q$.

$$\begin{array}{ccccccccccc} 0 & \mapsto & 000 & \to & \boxed{\Gamma} & \to & 000 & 001 & 010 & 100 & \mapsto & 0 \\ 1 & \mapsto & 111 & \to & & \to & 011 & 101 & 110 & 111 & \mapsto & 1 \end{array}$$

$$\begin{array}{ccc} \to & \boxed{\Gamma'} & \to \\ \to & & \to \end{array}$$

where

$$\Gamma' = \begin{pmatrix} P^3 + 3P^2Q & Q^3 + 3Q^2P \\ Q^3 + 3Q^2P & P^3 + 3P^2Q \end{pmatrix}.$$

$$Pr_E(\Delta_{\max}) = Q^3 + 3Q^2P.$$

Improving reliability

$$\begin{array}{ccccccccc} 0 & \mapsto & 0^n & \rightarrow & \boxed{\begin{array}{c}\Gamma \\ \end{array}} & \rightarrow & \text{majority is} & 0 & \mapsto & 0 \\ 1 & \mapsto & 1^n & \rightarrow & & \rightarrow & \ldots\ldots & 1 & \mapsto & 1 \end{array}$$

Improving reliability

$$\begin{array}{ccccccc} 0 & \mapsto & 0^n & \to & \boxed{\Gamma} & \to & \text{majority is } 0 & \mapsto & 0 \\ 1 & \mapsto & 1^n & \to & & \to & \ldots\ldots\ 1 & \mapsto & 1 \end{array}$$

$$\begin{pmatrix} \sum_{i=\lceil \frac{n}{2} \rceil}^{n} \binom{n}{i} P^i \cdot Q^{n-i} & \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i} P^i \cdot Q^{n-i} \\ \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i} P^i \cdot Q^{n-i} & \sum_{i=\lceil \frac{n}{2} \rceil}^{n} \binom{n}{i} P^i \cdot Q^{n-i} \end{pmatrix}$$

Improving reliability

The probability of error

$$Pr_E(\Delta_{\max}) = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i} P^i \cdot Q^{n-i} \leq \underbrace{\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i}}_{2^{n-1}} P^{\lfloor \frac{n}{2} \rfloor} \cdot Q^{\lfloor \frac{n}{2} \rfloor}$$

Since $\frac{1}{4} > P \cdot Q$, we have $PQ = \frac{\delta}{4}$, for some $\delta < 1$. Hence

$$Pr_E(\Delta_{\max}) \leq 2^{n-1} \cdot (PQ)^{\lfloor \frac{n}{2} \rfloor} = 2^{n-1} \cdot \frac{\delta^{\lfloor \frac{n}{2} \rfloor}}{2^{2 \cdot \lfloor \frac{n}{2} \rfloor}} = \delta^{\lfloor \frac{n}{2} \rfloor}$$

Therefore

$$\boxed{Pr_E(\Delta_{\max}) \to 0 \text{ if } n \to \infty.}$$

Improving reliability

The probability of error

$$Pr_E(\Delta_{\max}) = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i} P^i \cdot Q^{n-i} \leq \underbrace{\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i}}_{2^{n-1}} P^{\lfloor \frac{n}{2} \rfloor} \cdot Q^{\lfloor \frac{n}{2} \rfloor}$$

Since $\frac{1}{4} > P \cdot Q$, we have $PQ = \frac{\delta}{4}$, for some $\delta < 1$. Hence

$$Pr_E(\Delta_{\max}) \leq 2^{n-1} \cdot (PQ)^{\lfloor \frac{n}{2} \rfloor} = 2^{n-1} \cdot \frac{\delta^{\lfloor \frac{n}{2} \rfloor}}{2^{2 \cdot \lfloor \frac{n}{2} \rfloor}} = \delta^{\lfloor \frac{n}{2} \rfloor}$$

Therefore

$$\boxed{Pr_E(\Delta_{\max}) \to 0 \text{ if } n \to \infty.}$$

But can we avoid stretching of the message to ∞ **?**

Hamming distance

For $u, v \in \mathcal{A}^n$,

$$d(u, v) \quad = \quad |\{i : u_i \neq v_i\}|$$

Hamming distance

For $u, v \in \mathcal{A}^n$,

$$d(u,v) \quad = \quad |\{i : u_i \neq v_i\}|$$

$$\begin{aligned} \text{positivity} \qquad & d(u,v) = 0 \iff u = v, \\ \text{symmetry} \qquad & d(u,v) = d(v,u), \\ \text{triangle inequality} \qquad & d(u,w) \ \leq \ d(u,v) + d(v,w) \end{aligned}$$

Hamming distance

For $u, v \in \mathcal{A}^n$,

$$d(u, v) \quad = \quad |\{i : u_i \neq v_i\}|$$

$$\begin{array}{rl} \text{positivity} & d(u,v) = 0 \iff u = v, \\ \text{symmetry} & d(u,v) = d(v,u), \\ \text{triangle inequality} & d(u,w) \leq d(u,v) + d(v,w) \\ & (\{i : u_i \neq w_i\} \subseteq \{i : u_i \neq v_i\} \cup \{i : v_i \neq w_i\}). \end{array}$$

Hamming distance

For $u, v \in \mathcal{A}^n$,

$$d(u,v) \quad = \quad |\{i : u_i \neq v_i\}|$$

$$\begin{array}{rl} \text{positivity} & d(u,v) = 0 \iff u = v, \\ \text{symmetry} & d(u,v) = d(v,u), \\ \text{triangle inequality} & d(u,w) \ \leq \ d(u,v) + d(v,w) \\ & (\{i : u_i \neq w_i\} \subseteq \{i : u_i \neq v_i\} \cup \{i : v_i \neq w_i\}). \end{array}$$

For a BSC $\Gamma =, \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, and an input-output pair (A, B),

$$p(b_1 \ldots b_k | a_1 \ldots a_k) \quad = \quad Q^{d(\vec{a},\vec{b})} \cdot P^{k-d(\vec{a},\vec{b})}.$$

Transmission error

For a BSC $\Gamma =, \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, and an input-output pair (A, B), let

$$E = A \oplus B.$$

Transmission error

For a BSC $\Gamma =, \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, and an input-output pair (A, B), let

$$E = A \oplus B.$$

Note:

$$p(b|a) = p(E = a \oplus b)$$

Indeed,

$$p(b|a) = \begin{cases} P & a = b \quad (E = a \oplus b = 0) \\ Q & a \neq b \quad (E = a \oplus b = 1) \end{cases}$$

On the other hand,

$$p(E = 0) = p(A = 0) \cdot p(0 \to 0) + p(A = 1) \cdot p(1 \to 1) = P$$

and

$$p(E = 1) = p(A = 0) \cdot p(0 \to 1) + p(A = 1) \cdot p(1 \to q) = Q.$$

Transmission error in the multiple use of channels

Let $E_i = A_i \oplus B_i$, for $i = 1, \ldots, k$.
Assuming the independence of symbols

$$p\,(b_1, b_2, \ldots b_k \mid a_1, a_2 \ldots a_k) \quad = \quad p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \ldots \cdot p(b_k \mid a_k),$$

the variables $E_1, \ldots, E_k$ are **independent**.

Transmission error in the multiple use of channels

Let $E_i = A_i \oplus B_i$, for $i = 1, \ldots, k$.
Assuming the independence of symbols

$$p\left(b_1, b_2, \ldots b_k \mid a_1, a_2 \ldots a_k\right) \quad = \quad p(b_1 \mid a_1) \cdot p(b_2 \mid a_2) \cdot \ldots \cdot p(b_k \mid a_k),$$

the variables $E_1, \ldots, E_k$ are **independent**.

$$p(e_1 \ldots e_k) = \sum_{\vec{a}} p(\vec{A} = \vec{a} \wedge \vec{B} = \vec{a} \oplus \vec{e}) = \sum_{p(\vec{a})>0} p(\vec{A} = \vec{a}) \cdot p\left(\vec{B} = \vec{a} \oplus \vec{e} | \vec{A} = \vec{a}\right),$$

$$\begin{aligned} p(\vec{B} = \vec{a} \oplus \vec{e} | \vec{A} = \vec{a}) \quad &= \quad p(B_1 = a_1 \oplus e_1 | A_1 = a_1) \ldots p(B_k = a_k \oplus e_k | A_k = a_k) \\ &= \quad p(E_1 = e_1) \cdot \ldots \cdot p(E_k = e_k) \end{aligned}$$

for any $\vec{a}$, hence

$$p(e_1 \ldots e_k) \quad = \quad p(e_1) \cdot \ldots \cdot p(e_k).$$

Transmission algorithm – outline

Given: a random $X \in \mathcal{X}$, $|\mathcal{X}| = m$, $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, $P > Q$.

1. Choose $n \in \mathbb{N}$, and $C \subseteq \{0,1\}^n$ with $|C| = m$.
2. Choose $\varphi : \mathcal{X} \overset{1:1}{\to} C$. Let $\vec{A} = \varphi \circ X$.
3. Send

$$\underbrace{a_1, a_2, \ldots a_k}_{\vec{A}} \to \boxed{\Gamma} \to \underbrace{b_1, b_2, \ldots b_k}_{\vec{B}}$$

$$p(b_1 \ldots b_n | a_1 \ldots a_n) \quad = \quad Q^{d(\vec{a},\vec{b})} \cdot P^{n-d(\vec{a},\vec{b})}.$$

4. To decode, given $\vec{B} = b_1 \ldots b_n$, choose

$$\Delta(b_1 \ldots b_n) = a_1 \ldots a_n \in C$$

maximising $p(b_1 \ldots b_n | a_1 \ldots a_n)$ (minimising $d(\vec{a}, \vec{b})$).

Goal: minimise the probability of error

$$Pr_E(\Delta, \vec{A}) \quad = p(\Delta \circ \vec{B} \neq \vec{A}).$$

keeping the ratio $\frac{n}{\log m}$ as small as possible $< \infty$.

Worst case distribution

Fact. Let $\vec{A}, \vec{U} \in C \subseteq \{0,1\}^n$, with $\vec{U}$ uniform and $\vec{A}$ arbitrary. Then there is a permutation $\sigma : C \stackrel{1:1}{\rightarrow} C$ such that

$$Pr_E(\Delta, \sigma \circ \vec{A}) \quad \leq \quad Pr_E(\Delta, \vec{U}).$$

Worst case distribution

Fact. Let $\vec{A}, \vec{U} \in C \subseteq \{0,1\}^n$, with $\vec{U}$ uniform and $\vec{A}$ arbitrary. Then there is a permutation $\sigma : C \stackrel{1:1}{\to} C$ such that

$$Pr_E(\Delta, \sigma \circ \vec{A}) \quad \leq \quad Pr_E(\Delta, \vec{U}).$$

Lemma. Let $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, and $p_1, \ldots, p_m \in [0,1]$ with $p_1 + \cdots + p_m = 1$.

Worst case distribution

Fact. Let $\vec{A}, \vec{U} \in C \subseteq \{0,1\}^n$, with $\vec{U}$ uniform and $\vec{A}$ arbitrary. Then there is a permutation $\sigma : C \stackrel{1:1}{\rightarrow} C$ such that

$$Pr_E(\Delta, \sigma \circ \vec{A}) \quad \leq \quad Pr_E(\Delta, \vec{U}).$$

Lemma. Let $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, and $p_1, \ldots, p_m \in [0,1]$ with $p_1 + \cdots + p_m = 1$.
If $\alpha_1 \leq \cdots \leq \alpha_m$ and $p_1 \geq \cdots \geq p_m$, then

$$\sum_{i=1}^{m} p_i \alpha_i \leq \frac{1}{m} \sum_{i=1}^{m} \alpha_i.$$

Lemma. $\alpha_1 \le \cdots \le \alpha_m$, $1 \ge p_1 \ge \cdots \ge p_m \ge 0$, $p_1 + \cdots + p_m = 1$, then $\sum_{i=1}^{m} p_i \alpha_i \le \frac{1}{m} \sum_{i=1}^{m} \alpha_i$.

Proof by induction on m.

$p_m = \frac{1}{m} - h$, for some $h \ge 0$, $\frac{1}{m-1} \sum_{i=1}^{m-1} \alpha_i \le \alpha_m$. By induction hypo.

$$\frac{p_1}{p_1 + \cdots + p_{m-1}} \alpha_1 + \cdots + \frac{p_{m-1}}{p_1 + \cdots + p_{m-1}} \alpha_{m-1} \le \frac{1}{m-1} \sum_{i=1}^{m-1} \alpha_i.$$

Lemma. $\alpha_1 \le \cdots \le \alpha_m$, $1 \ge p_1 \ge \cdots \ge p_m \ge 0$, $p_1 + \cdots + p_m = 1$, then $\sum_{i=1}^m p_i \alpha_i \le \frac{1}{m} \sum_{i=1}^m \alpha_i$.

Proof by induction on m.

$p_m = \frac{1}{m} - h$, for some $h \ge 0$, $\frac{1}{m-1} \sum_{i=1}^{m-1} \alpha_i \le \alpha_m$. By induction hypo.

$$\frac{p_1}{p_1 + \cdots + p_{m-1}} \alpha_1 + \cdots + \frac{p_{m-1}}{p_1 + \cdots + p_{m-1}} \alpha_{m-1} \le \frac{1}{m-1} \sum_{i=1}^{m-1} \alpha_i.$$

$$p_1\alpha_1 + \cdots + p_{m-1}\alpha_{m-1} + p_m\alpha_m \le \underbrace{(p_1 + \cdots + p_{m-1})}_{1-p_m} \cdot \frac{1}{m-1} \cdot \sum_{i=1}^{m-1} \alpha_i + p_m\alpha_m =$$

$$\left(\frac{m-1}{m} + h\right) \frac{1}{m-1} \sum_{i=1}^{m-1} \alpha_i + \left(\frac{1}{m} - h\right)\alpha_m = \frac{1}{m} \sum_{i=1}^{m} \alpha_i + h \cdot \underbrace{\left(\frac{1}{m-1} \sum_{i=1}^{m-1} \alpha_i - \alpha_m\right)}_{\le 0}$$

$$\le \frac{1}{m} \cdot \sum_{i=1}^{m} \alpha_i.$$

Proof of the Fact $\ldots Pr_E(\Delta, \sigma \circ \vec{A}) \leq Pr_E(\Delta, \vec{U})$, for some σ.

Recall: $p(\vec{B} = \vec{b} | \vec{A} = \vec{a}) = p(\vec{E} = \vec{a} \oplus \vec{b})$ (for any in-out A, B).

Proof of the Fact ... $Pr_E(\Delta, \sigma \circ \vec{A}) \leq Pr_E(\Delta, \vec{U})$, for some σ.

Recall: $p(\vec{B} = \vec{b} | \vec{A} = \vec{a}) = p(\vec{E} = \vec{a} \oplus \vec{b})$ (for any in-out A, B).

$$\begin{aligned} Pr_E(\Delta, \vec{A}) &= \sum_{\vec{a} \in C} p(\vec{A} = \vec{a}) p(\Delta \circ \vec{B} \neq \vec{a} | \vec{A} = \vec{a}) \\ &= \sum_{\vec{a} \in C} p(\vec{A} = \vec{a}) p(\Delta(\vec{a} \oplus \vec{E}) \neq \vec{a}) \\ Pr_E(\Delta, \vec{U}) &= \frac{1}{|C|} \sum_{\vec{a} \in C} p(\Delta(\vec{a} \oplus \vec{E}) \neq \vec{a}) \end{aligned}$$

Use the Lemma for numbers:

$$\begin{aligned} p(\vec{A} = \vec{a}), &\qquad \vec{a} \in C, \\ p(\Delta(\vec{a} \oplus \vec{E}) \neq \vec{a}), &\qquad \vec{a} \in C. \end{aligned}$$

□

Transmission rate

For an alphabet with $|\mathcal{A}| = r \geq 2$,
the **transmission rate** of a code $C \subseteq \mathcal{A}^n$ is

$$R_r(C) \quad = \quad \frac{\log_r |C|}{n}.$$

As usual, $R = R_2$.

Transmission rate

For an alphabet with $|\mathcal{A}| = r \geq 2$,
the **transmission rate** of a code $C \subseteq \mathcal{A}^n$ is

$$R_r(C) \quad = \quad \frac{\log_r |C|}{n}.$$

As usual, $R = R_2$.

Example. If $C = \{000, 111\}^m \subseteq \{0,1\}^{3m}$ then

$$R(C) = \frac{m}{3m} = \frac{1}{3}.$$

No error

Theorem If $Pr_E(\Delta, \vec{A}) = 0$ (with A uniform) then $R_r(C) \leq \log_r 2 \cdot C_\Gamma$.

In particular,

$$R(C) \quad \leq \quad C_\Gamma.$$

No error

Theorem If $Pr_E(\Delta, \vec{A}) = 0$ (with A uniform) then $R_r(C) \leq \log_r 2 \cdot C_\Gamma$.

In particular,

$$R(C) \leq C_\Gamma.$$

Proof. The independence of symbols implies

$$H(\vec{B}|\vec{A}) = H(B_1|A_1) + \ldots + H(B_n|A_n).$$

No error

Theorem If $Pr_E(\Delta, \vec{A}) = 0$ (with A uniform) then $R_r(C) \leq \log_r 2 \cdot C_\Gamma$.

In particular,

$$R(C) \leq C_\Gamma.$$

Proof. The independence of symbols implies

$$H(\vec{B}|\vec{A}) = H(B_1|A_1) + \ldots + H(B_n|A_n).$$

Further

$$\begin{aligned} I(\vec{A}, \vec{B}) &= H(\vec{B}) - H(\vec{B}|\vec{A}) \\ &\leq \sum_{i=1}^{n} H(B_i) - \sum_{i=1}^{n} H(B_i|A_i) \\ &= \sum_{i=1}^{n} \underbrace{(H(B_i) - H(B_i|A_i))}_{I(A_i,B_i)} \\ &\leq n \cdot C_\Gamma. \end{aligned}$$

Proof of $R_r(C) \leq \log_r 2 \cdot C_\Gamma$ cont'd.

We got $I(\vec{A}, \vec{B}) \leq n \cdot C_\Gamma$, hence

Proof of $R_r(C) \leq \log_r 2 \cdot C_\Gamma$ cont'd.

We got $I(\vec{A}, \vec{B}) \leq n \cdot C_\Gamma$, hence

$$I_r(\vec{A}, \vec{B}) \quad \leq \quad \log_r 2 \cdot n \cdot C_\Gamma.$$

But

$$\begin{aligned} I_r(\vec{A}, \vec{B}) &= H_r(\vec{A}) - \underbrace{H_r(\vec{A}|\vec{B})}_{0} \\ &= \log_r m \end{aligned}$$

where $m = |C|$.

Proof of $R_r(C) \leq \log_r 2 \cdot C_\Gamma$ cont'd.

We got $I(\vec{A}, \vec{B}) \leq n \cdot C_\Gamma$, hence

$$I_r(\vec{A}, \vec{B}) \quad \leq \quad \log_r 2 \cdot n \cdot C_\Gamma .$$

But

$$\begin{aligned} I_r(\vec{A}, \vec{B}) &= H_r(\vec{A}) - \underbrace{H_r(\vec{A}|\vec{B})}_{0} \\ &= \log_r m \end{aligned}$$

where $m = |C|$. Hence

$$R_r(C) = \frac{\log_r m}{n} \leq \log_r 2 \cdot C_\Gamma .$$

□

Example: noisy typewriter revisited

$\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters)

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = 0.5$$

where $next(a) = b$, $next(b) = c$, ..., $next(y) = z$, $next(z) = a$.

$C_\Gamma = \max_A I(A; B) = \max_A H(B) - \underbrace{H(B|A)}_{1} = \log 26 - 1 = \log 13.$

Example: noisy typewriter revisited

$\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters)

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = 0.5$$

where $next(a) = b$, $next(b) = c$, $\ldots$, $next(y) = z$, $next(z) = a$.
$C_\Gamma = \max_A I(A;B) = \max_A H(B) - \underbrace{H(B|A)}_{1} = \log 26 - 1 = \log 13.$

If $|C| = 26^k$ then

$$\frac{\log_{26}|C|}{m} = \frac{k}{m}$$

Example: noisy typewriter revisited

$\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters)

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = 0.5$$

where $next(a) = b$, $next(b) = c$, ..., $next(y) = z$, $next(z) = a$.
$C_\Gamma = \max_A I(A; B) = \max_A H(B) - \underbrace{H(B|A)}_{1} = \log 26 - 1 = \log 13.$

If $|C| = 26^k$ then

$$\frac{\log_{26} |C|}{m} = \frac{k}{m} \leq \log_{26} 2 \cdot \log_2 13 = \frac{\log_2 13}{\log_2 13 + 1}.$$

Example: noisy typewriter revisited

$\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters)

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = 0.5$$

where $next(a) = b$, $next(b) = c$, ..., $next(y) = z$, $next(z) = a$.
$C_\Gamma = \max_A I(A; B) = \max_A H(B) - \underbrace{H(B|A)}_{1} = \log 26 - 1 = \log 13.$

If $|C| = 26^k$ then

$$\frac{\log_{26}|C|}{m} = \frac{k}{m} \leq \log_{26} 2 \cdot \log_2 13 = \frac{\log_2 13}{\log_2 13 + 1}.$$

Note: this bound also follows from the inequality $26^k \leq \frac{26^m}{2^m}$ (a word of length m can give 2^m results.)

Example: noisy typewriter cont'd

$$C = \left\{ \begin{array}{ccccccc} aa & cc & ee & \ldots & \ldots & ww & yy \\ ac & ce & eg & \ldots & \ldots & wy & ya \end{array} \right\}, \ |C| = 26, m = 2.$$

$$\frac{\log_{26} |C|}{m} = \frac{1}{2} \lll \frac{\log_2 13}{\log_2 13 + 1}.$$

Example: noisy typewriter cont'd

$$C = \left\{ \begin{array}{ccccccc} aa & cc & ee & \dots & \dots & ww & yy \\ ac & ce & eg & \dots & \dots & wy & ya \end{array} \right\}, \ |C| = 26, m = 2.$$

$$\frac{\log_{26}|C|}{m} = \frac{1}{2} \lll \frac{\log_2 13}{\log_2 13 + 1}.$$

$$C = \left\{\dots, \dots, \boxed{\mathtt{x\ y\ z\quad t}}, \dots, \dots\right\}, \ |C| = 26^3, m = 4,$$

where t is on the list $a, c, e, \dots, w, y$ on the position $(x \bmod 2) \cdot 4 + (y \bmod 2) \cdot 2 + (z \bmod 2) \cdot 1$.

$$\frac{\log_{26}|C|}{m} = \frac{3}{4} \lessapprox \frac{\log_2 13}{\log_2 13 + 1}.$$

Example: noisy typewriter cont'd

$$C = \{\ldots, \ldots, \boxed{w}, \ldots, \ldots\}, \ |C| = 26^k,$$

where w encodes a number $1 \cdot 26^k + a_{k-1} \cdot 26^{k-1} + \cdots + a_0 \cdot 26^0$ using m of the 13 digits $a, c, e, \ldots, w, y$, where

$$m = k + \log_{13} 2 \cdot (k+1)$$

hence

$$\frac{\log_{26}|C|}{m} = \frac{k}{k + \log_{13} 2 \cdot (k+1)} = \frac{\log_2 13}{1 + \log_2 13 + \frac{1}{k}} \approx \frac{\log_2 13}{\log_2 13 + 1}.$$

Shannon channel coding theorem

Theorem. $\Gamma = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, $P > Q$. Then $\forall \varepsilon, \delta > 0 \;\; \exists n_0 \;\; \forall n \geq n_0$ $\exists C \subseteq \{0,1\}^n$

$$C_\Gamma - \varepsilon \leq R(C) \leq C_\Gamma$$

$$Pr_E(\Delta, C) \leq \delta$$

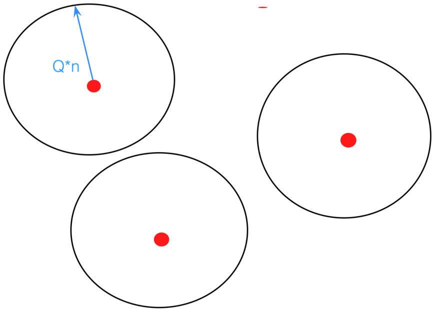
We assume $\Delta = \Delta_{\max}$ and C is uniform.

Shannon channel coding theorem

Idea. The expected distance between A and B is $\mathbf{Q}\cdot\mathbf{n}$. Try to pack in $\{0,1\}^n$ as many disjoint balls of radius $\mathbf{Q}\cdot\mathbf{n}$ as possible.

Shannon channel coding theorem

Idea. The expected distance between A and B is $\mathbf{Q}\cdot\mathbf{n}$. Try to pack in $\{0,1\}^n$ as many disjoint balls of radius $\mathbf{Q}\cdot\mathbf{n}$ as possible.



The centers of the **m** balls will be the code words.

Proof of the Shannon channel coding theorem

$\vec{a} \in C, \quad \vec{e} \in \{0,1\}^n, \quad \rho > 0.$

$$\left(d(\vec{a}, \vec{a} \oplus \vec{e}) \leq \rho\right) \land \left(\forall \vec{b} \in C - \{\vec{a}\}, d(\vec{b}, \vec{a} \oplus \vec{e}) > \rho\right) \implies$$

Proof of the Shannon channel coding theorem

$\vec{a} \in C, \quad \vec{e} \in \{0,1\}^n, \quad \rho > 0.$

$$\left(d(\vec{a}, \vec{a} \oplus \vec{e}) \leq \rho\right) \wedge \left(\forall \vec{b} \in C - \{\vec{a}\}, d(\vec{b}, \vec{a} \oplus \vec{e}) > \rho\right) \implies$$

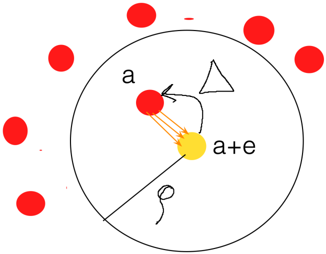
$$\implies \Delta(\vec{a} \oplus \vec{e}) \; = \; \vec{a}.$$

Proof of the Shannon channel coding theorem

$\vec{a} \in C, \quad \vec{e} \in \{0,1\}^n, \quad \rho > 0.$

$$\left(d(\vec{a}, \vec{a} \oplus \vec{e}) \leq \rho\right) \wedge \left(\forall \vec{b} \in C - \{\vec{a}\}, d(\vec{b}, \vec{a} \oplus \vec{e}) > \rho\right) \implies$$

$$\implies \Delta(\vec{a} \oplus \vec{e}) \; = \; \vec{a}.$$



$$p(\Delta(\vec{a} \oplus \vec{E}) \neq \vec{a}) \;\leq\; p(d(\vec{a}, \vec{a} \oplus \vec{E}) > \rho) + \sum_{\vec{b} \in C - \{\vec{a}\}} p(d(\vec{b}, \vec{a} \oplus \vec{E}) \leq \rho)$$

Weak Law of Large Numbers

$X_1, X_2, \ldots, X_n$ independent with the same distribution, $\mu = E(X_i)$, then, for $\eta > 0$,

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \eta\right) \to 0 \text{ if } n \to \infty.$$

Weak Law of Large Numbers

$X_1, X_2, \ldots, X_n$ independent with the same distribution, $\mu = E(X_i)$, then, for $\eta > 0$,

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \eta\right) \rightarrow 0 \text{ if } n \rightarrow \infty.$$

Hence

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} E_i - Q\right| > \eta\right) \rightarrow 0 \text{ if } n \rightarrow \infty,$$

since $E(E_i) = 0 \cdot P + \cdot Q = Q$.

Weak Law of Large Numbers

$X_1, X_2, \ldots, X_n$ independent with the same distribution, $\mu = E(X_i)$, then, for $\eta > 0$,

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \eta\right) \rightarrow 0 \text{ if } n \rightarrow \infty.$$

Hence

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} E_i - Q\right| > \eta\right) \rightarrow 0 \text{ if } n \rightarrow \infty,$$

since $E(E_i) = 0 \cdot P + \cdot Q = Q$. Therefore, with $\rho = n \cdot (Q + \eta)$,

$$p(d(\vec{a}, \vec{a} \oplus \vec{E}) > \rho) \leq p\left(\frac{1}{n} \cdot \sum_{i=1}^{n} E_i > Q + \eta\right) \leq$$

$$p\left(\left|\frac{1}{n} \cdot \sum_{i=1}^{n} E_i - Q\right| > \eta\right) \leq \frac{\delta}{2},$$

for n sufficiently large.

Proof of the Shannon channel coding theorem cont'd

Recall, with $\delta, \eta > 0$, $\rho = n \cdot (Q + \eta)$,

$$
\begin{aligned}
Pr_E(\Delta, C) &= \frac{1}{m} \sum_{\vec{a} \in C} p(\Delta(\vec{a} \oplus \vec{E}) \neq \vec{a}) \\
&\leq \frac{1}{m} \sum_{\vec{a} \in C} \left(p(d(\vec{a}, \vec{a} \oplus \vec{E}) > \rho) + \sum_{\vec{b} \in C - \{\vec{a}\}} p(d(\vec{b}, \vec{a} \oplus \vec{E}) \leq \rho) \right) \\
&\leq \frac{\delta}{2} + \frac{1}{m} \sum_{\vec{a} \in C} \sum_{\vec{b} \in C - \{\vec{a}\}} p(d(\vec{b}, \vec{a} \oplus \vec{E}) \leq \rho),
\end{aligned}
$$

The size of a ball

Lemma. For $\lambda \leq \frac{1}{2}$,

$$\sum_{i \leq \lambda \cdot n} \binom{n}{i} \leq 2^{n \cdot H(\lambda)},$$

where $H(x) = -x \log x - (1-x) \cdot \log(1-x)$.

The size of a ball

Lemma. For $\lambda \leq \frac{1}{2}$,

$$\sum_{i \leq \lambda \cdot n} \binom{n}{i} \leq 2^{n \cdot H(\lambda)},$$

where $H(x) = -x \log x - (1-x) \cdot \log(1-x)$.

Proof. Let $\kappa = 1 - \lambda$, then $\kappa \geq \lambda$.

The size of a ball

Lemma. For $\lambda \leq \frac{1}{2}$,

$$\sum_{i \leq \lambda \cdot n} \binom{n}{i} \leq 2^{n \cdot H(\lambda)},$$

where $H(x) = -x \log x - (1 - x) \cdot \log(1 - x)$.

Proof. Let $\kappa = 1 - \lambda$, then $\kappa \geq \lambda$. We first show that, for all $i \leq \lambda n$,

$$\lambda^i \kappa^{n-i} \geq \lambda^{\lambda n} \cdot \kappa^{\kappa n}.$$

The size of a ball

Lemma. For $\lambda \leq \frac{1}{2}$,

$$\sum_{i \leq \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \quad \leq \quad 2^{n \cdot H(\lambda)},$$

where $H(x) = -x \log x - (1-x) \cdot \log(1-x)$.

Proof. Let $\kappa = 1 - \lambda$, then $\kappa \geq \lambda$. We first show that, for all $i \leq \lambda n$,

$$\lambda^i \kappa^{n-i} \quad \geq \quad \lambda^{\lambda n} \cdot \kappa^{\kappa n}.$$

For λn integer just replace bigger by smaller, otherwise $\lambda n = \lfloor \lambda n \rfloor + \Delta\lambda$, $\kappa n = \lfloor \kappa n \rfloor + \Delta\kappa$, $\lfloor \lambda n \rfloor + \lfloor \kappa n \rfloor = n - 1$, and $\Delta\lambda + \Delta\kappa = 1$. For $i \leq \lambda n$,

$$\lambda^i \kappa^{n-i} \geq \lambda^{\lfloor \lambda n \rfloor} \cdot \kappa^{\lfloor \kappa n \rfloor + 1} = \lambda^{\lfloor \lambda n \rfloor} \cdot \kappa^{\lfloor \kappa n \rfloor} \underbrace{\kappa^{\Delta\lambda + \Delta\kappa}}_{\geq \lambda^{\Delta\lambda} \cdot \kappa^{\Delta\kappa}} \geq \lambda^{\lambda n} \cdot \kappa^{\kappa n}.$$

Proof

$\sum_{i \le \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \le 2^{n \cdot H(\lambda)}$, for $\lambda \le \frac{1}{2}$.

We have shown $\lambda^i \kappa^{n-i} \ge \lambda^{\lambda n} \cdot \kappa^{\kappa n}$.

Note

$$\begin{aligned} -\log_2 \lambda^{\lambda n} \cdot \kappa^{\kappa n} &= -n \cdot (\lambda \cdot \log_2 \lambda + \kappa \cdot \log_2 \kappa) \\ &= n \cdot H(\lambda). \end{aligned}$$

Hence

$$1 \ge \sum_{i \le \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \lambda^i \kappa^{n-i} \ge \sum_{i \le \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \lambda^{\lambda n} \cdot \kappa^{\kappa n}$$

and consequently

$$\sum_{i \le \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \le \frac{1}{\lambda^{\lambda n} \cdot \kappa^{\kappa n}} = 2^{n \cdot H(\lambda)},$$

$\square$

Proof of the Shannon channel coding theorem cont'd

Recall, with $\delta, \eta > 0$, $\rho = n \cdot (Q + \eta)$,

$$
\begin{aligned}
Pr_E(\Delta, C) &= \frac{1}{m} \sum_{\vec{a} \in C} p(\Delta(\vec{a} \oplus \vec{E}) \neq \vec{a}) \\
&\leq \frac{1}{m} \sum_{\vec{a} \in C} \left(p(d(\vec{a}, \vec{a} \oplus \vec{E}) > \rho) + \sum_{\vec{b} \in C - \{\vec{a}\}} p(d(\vec{b}, \vec{a} \oplus \vec{E}) \leq \rho) \right) \\
&\leq \frac{\delta}{2} + \underbrace{\frac{1}{m} \sum_{\vec{a} \in C} \sum_{\vec{b} \in C - \{\vec{a}\}} p(d(\vec{b}, \vec{a} \oplus \vec{E}) \leq \rho)}_{???},
\end{aligned}
$$

Probabilistic argument

Let $\mathcal{C}$ be the set of all sequences of **different** $c_1, \dots, c_m \in \{0,1\}^n$.

Let $N = |\mathcal{C}|$.

For $\bar{C} = (c_1, \dots, c_m)$, let $C = \{c_1, \dots . c_m\}$.

If

$$\frac{1}{N} \sum_{\bar{C}} \text{something}(C) \quad \leq \quad \delta$$

then there exists a code C, such that

$$\text{something}(C) \quad \leq \quad \delta$$

Probabilistic argument

Proof of the Shannon channel coding theorem cont'd

We will estimate

$$\frac{1}{N}\sum_{\bar{C}}\frac{1}{m}\sum_{\vec{a}\in C}\sum_{\vec{b}\in C-\{\vec{a}\}} p(d(\vec{b},\vec{a}\oplus\vec{E})\leq\rho)$$

$$= \frac{1}{N}\sum_{\bar{C}}\frac{1}{m}\sum_{i=1}^{m}\sum_{j\neq i} p(d(c_j,c_i\oplus\vec{E})\leq\rho)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\sum_{j\neq i}\underbrace{\frac{1}{N}\sum_{\bar{C}} p(d(c_j,c_i\oplus\vec{E})\leq\rho)}_{(*)}$$

We then estimate (*****), for a *fixed* pair of indices $i\neq j$.

Estimation

Let

$$S_{\rho}(\vec{e}) = \{\vec{b} \in \{0,1\}^n : d(\vec{b}, \vec{e}) \leq \rho\}.$$

Estimation

Let

$$S_{\rho}(\vec{e}) \quad = \quad \{\vec{b} \in \{0,1\}^n : d(\vec{b}, \vec{e}) \leq \rho\}.$$

Clearly $d(\vec{x}, \vec{y} \oplus \vec{e}) = d(\vec{x} \oplus \vec{y}, \vec{e})$, hence

$$\frac{1}{N} \sum_{\bar{C}} p(d(c_j, c_i \oplus \vec{E}) \leq \rho) = \frac{1}{N} \sum_{\bar{C}} p\left(c_i \oplus c_j \in S_{\rho}(\vec{E})\right)$$

$$= \sum_{\vec{e} \in \{0,1\}^n} p(\vec{E} = \vec{e}) \cdot \underbrace{\frac{1}{N} \sum_{\bar{C}} \overbrace{c_i \oplus c_j \in S_{\rho}(\vec{e})}^{boole}}_{(**)}$$

We now estimate the value of (**), for a fixed $\vec{e}$.

Estimation

$$\frac{1}{N}\sum_{\bar{C}} \overbrace{c_i \oplus c_j \in S_\rho(\vec{e})}^{boole}$$

Clearly, for any $\vec{a}, \vec{b} \in \{0,1\}^n - \{0^n\}$,

$$|\{\bar{C} : \vec{a} = c_i \oplus c_j\}| = |\{\bar{C} : \vec{b} = c_i \oplus c_j\}| = \frac{N}{2^n - 1}.$$

Hence

$$\underbrace{\frac{1}{N}\sum_{\bar{C}} \overbrace{c_i \oplus c_j \in S_\rho(\vec{e})}^{boole}}_{(**)} = \frac{1}{N} \cdot \frac{N}{2^n - 1}|S_\rho(\vec{e}) - \{0^n\}|,$$

Estimation

$$\frac{1}{N}\sum_{\bar{C}} \overbrace{c_i \oplus c_j \in S_\rho(\vec{e})}^{boole}$$

Clearly, for any $\vec{a}, \vec{b} \in \{0,1\}^n - \{0^n\}$,

$$|\{\bar{C} : \vec{a} = c_i \oplus c_j\}| = |\{\bar{C} : \vec{b} = c_i \oplus c_j\}| = \frac{N}{2^n - 1}.$$

Hence

$$\underbrace{\frac{1}{N}\sum_{\bar{C}} \overbrace{c_i \oplus c_j \in S_\rho(\vec{e})}^{boole}}_{(**)} \quad = \quad \frac{1}{N} \cdot \frac{N}{2^n - 1}|S_\rho(\vec{e}) - \{0^n\}|,$$

$$\sum_{\vec{e} \in \{0,1\}^n} p(\vec{E} = \vec{e}) \cdot \frac{1}{2^n - 1}|S_\rho(\vec{e}) - \{0^n\}| \quad = \quad \frac{1}{2^n - 1}|S_\rho(\vec{e}) - \{0^n\}|.$$

Proof of the Shannon channel coding theorem cont'd

But

$$|S_{\rho}(\vec{e}) - \{0^n\}| \quad \leq \quad 2^{n \cdot H(Q+\eta)}$$

(recall that $\rho = n(Q + \eta)$).

Proof of the Shannon channel coding theorem cont'd

But

$$|S_\rho(\vec{e}) - \{0^n\}| \leq 2^{n \cdot H(Q+\eta)}$$

(recall that $\rho = n(Q + \eta)$).

Hence

$$\frac{1}{N}\sum_{\bar{C}} \frac{1}{m}\sum_{\vec{a}\in C} \sum_{\vec{b}\in C-\{\vec{a}\}} p(d(\vec{b}, \vec{a}\oplus\vec{E}) \leq \rho) \leq \frac{1}{m}\sum_{i=1}^{m}\sum_{j\neq i} \frac{1}{2^n - 1} \cdot 2^{n\cdot H(Q+\eta)}$$

Proof of the Shannon channel coding theorem cont'd

But

$$|S_{\rho}(\vec{e}) - \{0^n\}| \quad \leq \quad 2^{n \cdot H(Q+\eta)}$$

(recall that $\rho = n(Q + \eta)$).

Hence

$$\frac{1}{N}\sum_{\bar{C}} \frac{1}{m}\sum_{\vec{a}\in C}\sum_{\vec{b}\in C-\{\vec{a}\}} p(d(\vec{b}, \vec{a}\oplus\vec{E}) \leq \rho) \;\leq\; \frac{1}{m}\sum_{i=1}^{m}\sum_{j\neq i} \frac{1}{2^n - 1} \cdot 2^{n \cdot H(Q+\eta)}$$

$$= \quad \frac{1}{m} \cdot m \cdot \underbrace{(m-1) \cdot \frac{1}{2^n - 1}}_{\leq \frac{m}{2^n}} \cdot 2^{n \cdot H(Q+\eta)}$$

$$\leq \quad m \cdot 2^{n(H(Q+\eta)-1)}$$

Proof of the Shannon channel coding theorem cont'd

Summarize

$$
\begin{aligned}
\frac{1}{N}\sum_{\bar{C}} Pr_E(\Delta, C) &\leq \frac{\delta}{2} + \frac{1}{m}\sum_{\vec{a}\in C}\sum_{\vec{b}\in C-\{\vec{a}\}} p(d(\vec{b}, \vec{a}\oplus\vec{E}) \leq \rho) \\
&\leq \frac{\delta}{2} + m\cdot 2^{n(H(Q+\eta)-1)} \\
&= \frac{\delta}{2} + 2^{n\cdot\left(\frac{\log m}{n}+H(Q+\eta)-1\right)}
\end{aligned}
$$

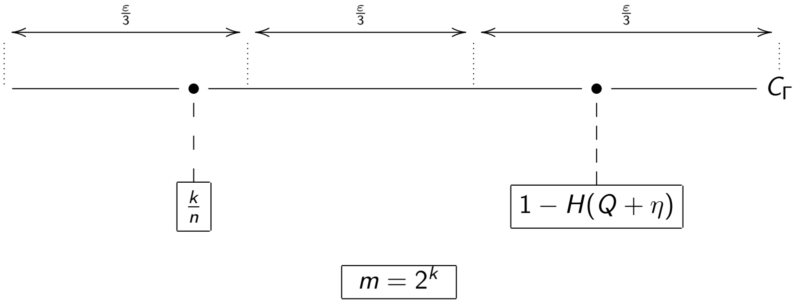
Note $\left(\frac{\log m}{n} + H(Q+\eta) - 1\right) \approx R(C) - C_\Gamma$.

Proof of the Shannon channel coding theorem cont'd

We can choose n_0, η, such that $\forall n \geq n_0,\ \exists m$,

$$C_\Gamma - \varepsilon \leq \frac{\log m}{n} \leq C_\Gamma$$

$$\frac{\log_2 m}{n} + H(Q+\eta) - 1 \leq -\frac{\varepsilon}{3}.$$



Proof of the Shannon channel coding theorem cont'd

We can choose n_0, η, such that $\forall n \geq n_0, \ \exists m$,

$$C_\Gamma - \varepsilon \ \leq \ \frac{\log m}{n} \ \leq \ C_\Gamma$$

$$\frac{\log_2 m}{n} + H(Q + \eta) - 1 \ \leq \ -\frac{\varepsilon}{3}.$$

Hence

$$\frac{1}{N} \sum_{\bar{C}} Pr_E(\Delta, C) \ \leq \ \frac{\delta}{2} + \underbrace{2^{n \cdot \left(\frac{\log m}{n} + H(Q+\eta) - 1 \right)}}_{\leq \frac{1}{2^{n \cdot \frac{\varepsilon}{3}}}}$$

$$\leq \ \frac{\delta}{2} + \frac{\delta}{2}.$$

By probabilistic argument, a desired code C exists (with $R(C) = \frac{\log m}{n}$). $\square$

The Shannon channel coding theorem generally

For any channel Γ, and $\varepsilon, \delta > 0$, for sufficiently large n, there exists a code $C \subseteq \{0,1\}^n$, along with some decision rule Δ_n satisfying

$$C_\Gamma - \varepsilon \leq \quad \frac{\log |C|}{n} \quad \leq C_\Gamma$$

$$Pr_E(\Delta, C) \quad \leq \delta.$$

In other words, there is a sequence of codes $C_\ell \subseteq \{0,1\}^{n_\ell}$, $\ell \to \infty$, along with decision rules Δ_ℓ such that

$$\frac{\log |C_\ell|}{n_\ell} \to C_\Gamma \quad \text{and} \quad Pr_E(Delta_\ell, C_\ell) \to 0.$$

Error correcting codes

Trading optimality for efficiency. Let $C \subseteq \{0,1\}^n$.

$$C \ni a_1, \dots a_n \to \boxed{\Gamma} \to b_1, \dots b_n \to \Delta(b_1 \dots, b_n) \in C$$

C **corrects** **k** errors if, for any $\vec{a} \in C$, $\vec{b} \in \{0,1\}^n$,

$$\text{if } \; d(\vec{a}, \vec{b}) \leq k \; \text{ then } \Delta(\vec{b}) = \vec{a}.$$

Error correcting codes

Trading optimality for efficiency. Let $C \subseteq \{0,1\}^n$.

$$C \ni a_1, \ldots a_n \rightarrow \boxed{\Gamma} \rightarrow b_1, \ldots b_n \rightarrow \Delta(b_1 \ldots, b_n) \in C$$

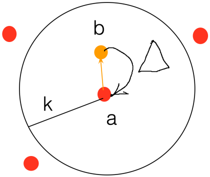
C **corrects k** errors if, for any $\vec{a} \in C$, $\vec{b} \in \{0,1\}^n$,

$$\text{if } \; d(\vec{a}, \vec{b}) \leq k \; \text{ then } \Delta(\vec{b}) = \vec{a}.$$

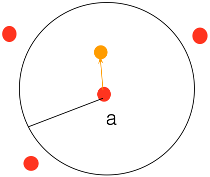
C **detects k** errors if, for any $\vec{a} \in C$, $\vec{b} \in \{0,1\}^n$,

$$\text{if } \; 0 < d(\vec{a}, \vec{b}) \leq k \; \text{ then } \vec{b} \notin C.$$

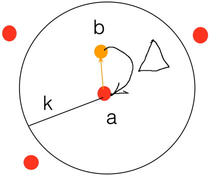
corrects

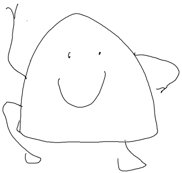


detects

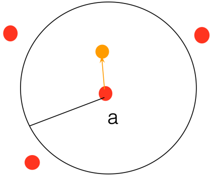


corrects

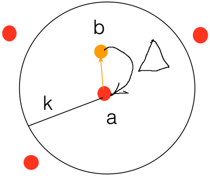


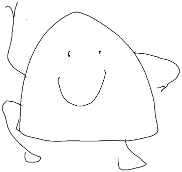


detects

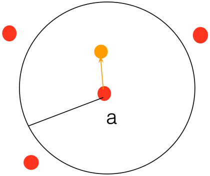


corrects





detects



Error correcting codes

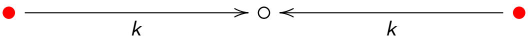
Let

$$d(C) \quad = \quad \min\{d(v,w) : v, w \in C,\ v \neq w\}.$$

Fact.

A code C corrects k errors if, and only if, $2k+1 \leq d(C)$.

A code C detects k errors if, and only if, $k < d(C)$.



Error correcting codes

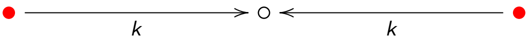
Let

$$d(C) \;=\; \min\{d(v,w) : v, w \in C,\ v \neq w\}.$$

Fact.

A code C corrects k errors if, and only if, $2k+1 \leq d(C)$.

A code C detects k errors if, and only if, $k < d(C)$.



Example. $\{0^n, 1^n : n \in \mathbb{N}\}$ corrects $\lfloor \frac{n-1}{2} \rfloor$ errors.

$\{w_1 w_2 \ldots w_n \in \{0,1\}^n : \sum_i w_i = 0 \bmod 2\}$ detects one error, but does not correct it.

One error

Problem. Find $C \subseteq \{0,1\}^{n+k}$ with $|C| = 2^n$ that corrects a **single** error.

One error

Problem. Find $C \subseteq \{0,1\}^{n+k}$ with $|C| = 2^n$ that corrects a **single** error.

To **detect**, $k = 1$ suffices.

One error

Problem. Find $C \subseteq \{0,1\}^{n+k}$ with $|C| = 2^n$ that corrects a **single** error.

To **detect**, $k = 1$ suffices. Prolongate $w = w_1 \dots w_k$ by

$$\textbf{check-bit}\ (w) \quad = \quad \sum_i w_i \bmod 2.$$

One error

Problem. Find $C \subseteq \{0,1\}^{n+k}$ with $|C| = 2^n$ that corrects a **single** error.

To **detect**, $k = 1$ suffices. Prolongate $w = w_1 \dots w_k$ by

$$\textbf{check-bit}\ (w) \quad = \quad \sum_i w_i \bmod 2.$$

Heuristics.

n original bits	**k** check bits

An error can appear on $n + k$ positions, hence

$$n + k + 1 \quad \leq \quad 2^k.$$

It is possible with $n + k + 1 = 2^k$ (for $k \geq 2$).

Hamming $(2^k - 1, k)$ code

Let $a_1 \dots a_n$ with $n = 2^k - k - 1$.

Add the **check bits** on the positions 2^i, for $i = 0, 1, \dots, k - 1$.

$\square$	$\square$	a_1	$\square$	a_2	a_3	a_4
x_1	x_2	x_3	x_4	x_5	x_6	x_7

Hamming $(2^k - 1, k)$ code

Let $a_1 \dots a_n$ with $n = 2^k - k - 1$.

Add the **check bits** on the positions 2^i, for $i = 0, 1, \dots, k - 1$.

$$\begin{array}{ccccccc} \square & \square & a_1 & \square & a_2 & a_3 & a_4 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{array}$$

They are computed by solving **k** equations over $\mathbb{Z}_2$ (i.e., mod2)

$$\begin{array}{lrcl} (\mathbf{0}) & x_1 + x_3 + x_5 + x_7 & = & 0 \\ (\mathbf{1}) & x_2 + x_3 + x_6 + x_7 & = & 0 \\ (\mathbf{2}) & x_4 + x_5 + x_6 + x_7 & = & 0, \end{array}$$

Hamming $(2^k - 1, k)$ code

Let $a_1 \dots a_n$ with $n = 2^k - k - 1$.
Add the **check bits** on the positions 2^i, for $i = 0, 1, \dots, k - 1$.

$\square$	$\square$	a_1	$\square$	a_2	a_3	a_4
x_1	x_2	x_3	x_4	x_5	x_6	x_7

They are computed by solving **k** equations over $\mathbb{Z}_2$ (i.e., mod2)

$$\begin{array}{lrcl} (\mathbf{0}) & x_1 + x_3 + x_5 + x_7 & = & 0 \\ (\mathbf{1}) & x_2 + x_3 + x_6 + x_7 & = & 0 \\ (\mathbf{2}) & x_4 + x_5 + x_6 + x_7 & = & 0, \end{array}$$

where in the equation (**i**), we sum up those x_t,

$$t = b_0 + b_1 2 + \dots + b_{k-1} 2^{k-1},$$

where the bit **i** is **one**.

$\square$	$\square$	a_1	$\square$	a_2	a_3	a_4
x_1	x_2	x_3	x_4	x_5	x_6	x_7

$$\begin{array}{lrcl} (\mathbf{0}) & x_1 + x_3 + x_5 + x_7 & = & 0 \\ (\mathbf{1}) & x_2 + x_3 + x_6 + x_7 & = & 0 \\ (\mathbf{2}) & x_4 + x_5 + x_6 + x_7 & = & 0 \end{array}$$

The **unknown** are x_{2^i}, where $i = 0, 1, \ldots, k - 1$.

$$x_1\, x_2, \ldots x_{n+k} \rightarrow \boxed{\Gamma} \rightarrow x'_1\, x'_2, \ldots x'_{n+k}$$

For example

$$\begin{array}{lrcl} (\mathbf{0}) & x'_1 + x'_3 + x'_5 + x'_7 & = & 0 \\ (\mathbf{1}) & x'_2 + x'_3 + x'_6 + x'_7 & = & 1 \\ (\mathbf{2}) & x'_4 + x'_5 + x'_6 + x'_7 & = & 1. \end{array}$$

Then an error has occurred on the position

$$\begin{array}{ccccccc} \square & \square & a_1 & \square & a_2 & a_3 & a_4 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{array}$$

$$\begin{array}{lrcl} \mathbf{(0)} & x_1 + x_3 + x_5 + x_7 & = & 0 \\ \mathbf{(1)} & x_2 + x_3 + x_6 + x_7 & = & 0 \\ \mathbf{(2)} & x_4 + x_5 + x_6 + x_7 & = & 0 \end{array}$$

The **unknown** are x_{2^i}, where $i = 0, 1, \ldots, k - 1$.

$$x_1\, x_2, \ldots x_{n+k} \rightarrow \boxed{\Gamma} \rightarrow x'_1\, x'_2, \ldots x'_{n+k}$$

For example

$$\begin{array}{lrcl} \mathbf{(0)} & x'_1 + x'_3 + x'_5 + x'_7 & = & 0 \\ \mathbf{(1)} & x'_2 + x'_3 + x'_6 + x'_7 & = & 1 \\ \mathbf{(2)} & x'_4 + x'_5 + x'_6 + x'_7 & = & 1. \end{array}$$

Then an error has occurred on the position

$$6 \quad =$$

$$\begin{array}{ccccccc} \square & \square & a_1 & \square & a_2 & a_3 & a_4 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{array}$$

$$\begin{array}{lrcl} \mathbf{(0)} & x_1 + x_3 + x_5 + x_7 & = & 0 \\ \mathbf{(1)} & x_2 + x_3 + x_6 + x_7 & = & 0 \\ \mathbf{(2)} & x_4 + x_5 + x_6 + x_7 & = & 0 \end{array}$$

The **unknown** are x_{2^i}, where $i = 0, 1, \ldots, k - 1$.

$$x_1\, x_2, \ldots x_{n+k} \to \boxed{\Gamma} \to x'_1\, x'_2, \ldots x'_{n+k}$$

For example

$$\begin{array}{lrcl} \mathbf{(0)} & x'_1 + x'_3 + x'_5 + x'_7 & = & 0 \\ \mathbf{(1)} & x'_2 + x'_3 + x'_6 + x'_7 & = & 1 \\ \mathbf{(2)} & x'_4 + x'_5 + x'_6 + x'_7 & = & 1. \end{array}$$

Then an error has occurred on the position

$$6 = 0 \cdot 2^0 + 1 \cdot 2^1 + 1 \cdot 2^2.$$

Hamming $(2^k - 1, k)$ code cont'd

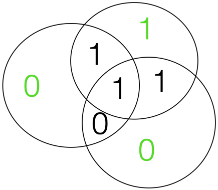
$$
\begin{array}{lrcl}
(\mathbf{0}) & x'_1 + x'_3 + x'_5 + x'_7 & = & 0 \\
(\mathbf{1}) & x'_2 + x'_3 + x'_6 + x'_7 & = & 1 \\
(\mathbf{2}) & x'_4 + x'_5 + x'_6 + x'_7 & = & 1.
\end{array}
$$

A single error (if any) has occurred on the position

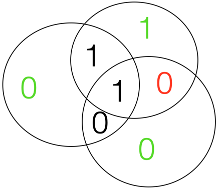
$$t = b_0 + b_1 2 + \ldots + b_{k-1} 2^{k-1}.$$

where b_i is the value of the equation $(\mathbf{i})$ after substitution.

Hamming (7,4) code



The sum in each circle should be **even**.



Then a “guilty” bit can be easily found.

Hamming's bound

If $C \subseteq \{0,1\}^m$ corrects t errors then

$$|C| \cdot \left(1 + m + \binom{m}{2} + \ldots + \binom{m}{t}\right) \leq 2^m,$$

Hamming's bound

If $C \subseteq \{0,1\}^m$ corrects t errors then

$$|C| \cdot \left(1 + m + \binom{m}{2} + \ldots + \binom{m}{t}\right) \leq 2^m,$$

Example. For $C = \{0^{2n+2}, 1^{2n+2}\}$, we have

$$\{0,1\}^{2n+2} = B\left(0^{2n+2}, n\right) \dot\cup B\left(1^{2n+2}, n\right) \dot\cup \{w \in \{0,1\}^{2n+2} : \sharp_0(w) = \sharp_1(w)\}.$$

Hamming's bound

If $C \subseteq \{0,1\}^m$ corrects t errors then

$$|C| \cdot \left(1 + m + \begin{pmatrix} m \\ 2 \end{pmatrix} + \ldots + \begin{pmatrix} m \\ t \end{pmatrix}\right) \quad \leq \quad 2^m,$$

Example. For $C = \{0^{2n+2}, 1^{2n+2}\}$, we have

$$\{0,1\}^{2n+2} \quad = \quad B\left(0^{2n+2}, n\right) \dot{\cup} B\left(1^{2n+2}, n\right) \dot{\cup} \{w \in \{0,1\}^{2n+2} : \sharp_0(w) = \sharp_1(w)\}.$$

But for the Hamming $\left(2^k - 1, k\right)$ code we have

$$\underbrace{2^{2^k-k-1}}_{|C|} \cdot \left(1 + \underbrace{(2^k - 1)}_{m}\right) \quad = \quad 2^{2^k-1}.$$

In this sense the Hamming code is **optimal**.

Hamming code

Recall

$$2^{2^k-k-1} \cdot \left(\underbrace{1 + (2^k - 1)}_{|ball|} \right) \;=\; 2^{2^k-1}.$$

Thus

$$d\left(Hamming(2^k - 1, k)\right) \;=$$

Hamming code

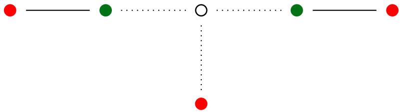
Recall

$$2^{2^k-k-1} \cdot \left(\underbrace{1 + (2^k - 1)}_{|ball|} \right) \;=\; 2^{2^k-1}.$$

Thus

$$d\left(Hamming(2^k - 1, k) \right) \;=\; 3.$$

Indeed, assumption that $d(v, w) \geq 4$, for the **closest** words v, w, leads to contradiction.



Hadamard code

Hadamard matrices. Values ± 1, any two distinct rows are orthogonal.

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Hadamard code

Hadamard matrices. Values ± 1, any two distinct rows are orthogonal.

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Note

$$\begin{aligned} H \cdot H^T &= n \cdot I_n \\ (\det H)^2 &= n^n \\ \det H &= n^{\frac{n}{2}}, \end{aligned}$$

which is maximal over $[-1, 1]$ (Hadamard).

Hadamard code

A Hadamard matrix H of order n induces a binary code $C \subseteq \{0, 1\}^n$.

Hadamard code

A Hadamard matrix H of order n induces a binary code $C \subseteq \{0,1\}^n$.

For the rows r_i of H, form $\pm r_1, \ldots, \pm r_n$, and replace -1 by 0. Then $|C| = 2n$ and

$$\forall v, w \in C,\ v \neq w \Rightarrow d(v,w) = n \ \vee \ d(v.w) = \frac{n}{2}$$

hence $d(C) = n$.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \mapsto \begin{matrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{matrix}$$

Linear codes

Recall

□	□	a_1	□	a_2	a_3	a_4
x_1	x_2	x_3	x_4	x_5	x_6	x_7

$$\begin{aligned} x_1 + x_3 + x_5 + x_7 &= 0 \\ x_2 + x_3 + x_6 + x_7 &= 0 \\ x_4 + x_5 + x_6 + x_7 &= 0 \end{aligned}$$

Note: the Hamming $(2^k - 1, k)$ code is closed under vector $\oplus$: if x and y are in the code, then so is $z = x \oplus y$

$$\begin{array}{cccccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ \oplus & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 \\ \hline & z_1 & z_2 & z_3 & z_4 & z_5 & z_6 & z_7 \end{array}$$

Linear codes

Recall

□	□	a_1	□	a_2	a_3	a_4
x_1	x_2	x_3	x_4	x_5	x_6	x_7

$$\begin{aligned} x_1 + x_3 + x_5 + x_7 &= 0 \\ x_2 + x_3 + x_6 + x_7 &= 0 \\ x_4 + x_5 + x_6 + x_7 &= 0 \end{aligned}$$

Note: the Hamming $(2^k - 1, k)$ code is closed under vector $\oplus$: if x and y are in the code, then so is $z = x \oplus y$

$$\begin{array}{cccccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ \oplus & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 \\ \hline & z_1 & z_2 & z_3 & z_4 & z_5 & z_6 & z_7 \end{array}$$

Thus it forms a **linear space** over the field $\mathbb{Z}_2$.

Linear codes

Similarly,

$$\{w_1 w_2 \dots w_n \in \{0,1\}^n : \sum_i w_i = 0 \bmod 2\}$$

which is the **maximal** (of size

Linear codes

Similarly,

$$\{w_1 w_2 \ldots w_n \in \{0,1\}^n : \sum_i w_i = 0 \bmod 2\}$$

which is the **maximal** (of size 2^{n-1}) code that **detects** one error, is a linear code.

Linear codes

Similarly,

$$\{w_1 w_2 \dots w_n \in \{0,1\}^n : \sum_i w_i = 0 \bmod 2\}$$

which is the **maximal** (of size 2^{n-1}) code that **detects** one error, is a linear code.

In general, for a finite field $\mathbb{F}_q$ ($q = |\mathbb{F}_q|$, $q = p^\alpha$, p prime), $C \subseteq \mathbb{F}_q^n$ is a **linear code** if it is a linear subspace of $\mathbb{F}_q^n$ over the field $\mathbb{F}_q$.

Linear codes

Let

$$\begin{aligned} weight(\mathbf{w}) &= |\{i : w_i \neq 0\}| \\ &= d(\mathbf{w}, \mathbf{0}). \end{aligned}$$

Fact. For a linear code $C \subseteq \mathbb{F}_q^n$,

$$d(C) = \min\{weight(\mathbf{w}) : \mathbf{w} \in C,\ \mathbf{w} \neq \mathbf{0}\}.$$

$\boxed{\leq}$ because $\mathbf{0} \in C$.

$\boxed{\geq}$ because $\forall \mathbf{v}, \mathbf{w} \in C,\ d(\mathbf{v}, \mathbf{w}) = weight(\mathbf{v} - \mathbf{w})$.

Linear codes

Let

$$\begin{aligned} weight(\mathbf{w}) &= |\{i : w_i \neq 0\}| \\ &= d(\mathbf{w}, \mathbf{0}). \end{aligned}$$

Fact. For a linear code $C \subseteq \mathbb{F}_q^n$,

$$d(C) = \min\{weight(\mathbf{w}) : \mathbf{w} \in C,\ \mathbf{w} \neq \mathbf{0}\}.$$

$\boxed{\leq}$ because $\mathbf{0} \in C$.

$\boxed{\geq}$ because $\forall \mathbf{v}, \mathbf{w} \in C,\ d(\mathbf{v}, \mathbf{w}) = weight(\mathbf{v} - \mathbf{w})$.

Example. In any Hamming $(2^k - 1, k)$ code there is an element with exactly **three** 1's, e.g., from

$$\begin{matrix} \square & \square & 1 & \square & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{matrix}$$