Information Theory On-Line. Synopsis of Lecture

> *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

21.02.2006.

***Notation*, what is it?**

An experiment: guess a *word* which somebody has thought of. Should it work as well with a *number*?

Note that integers written in a positional system are "densely packed", unlike words of natural language. That is, all strings over $\{0, 1, \ldots, 9\}$ denote some numbers (up to leading 0's), while only few strings over $\{a, b, \ldots, z\}$ are (meaningful) words. One explanation of this dissimilarity is that we dispose of efficient algorithms to operate on (short) encodings of numbers, while our "algorithms" to communicate with words require more redundancy.

Everyday life examples: writing the amount on cheque both by digits and by words, or spelling a flight number by phone.

Information theory tries to reconcile two antagonistic objectives:

- to make the message as short as possible,
- to prevent errors while the message is sent by an uncertain channel.

Is there any message that we could not make shorter? We are warned by Berry's paradox:

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

(Thus we have described it.) The concept of notation should be understood properly. Notation is not a part of an object, but it is given "from outside" to a set of objects, in order to distinguish between them.

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

Fact If $|S| = m > 0$ and $|\Sigma| = r \geq 2$ then, for some $s \in S$,

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

Proof The number of string shorter than k is

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

Letting $k = \lfloor \log_r m \rfloor$, we see that there is not enough words shorter than k to denote all elements of S. QED

Corollary If $\alpha : \mathbb{N} \to \Sigma^*$ is notation for natural numbers then, for infinitely many n's, $\alpha(n) > \lfloor \log_r n \rfloor$.

Proof Choose m such that $\lfloor \log_r m \rfloor > |\alpha(0)|$. By Fact above, some $i_0 \in \{0, 1, \ldots, m-1\}$ must satisfy $|\alpha(i_0)| \geq \lfloor \log_r m \rfloor > \lfloor \log_r i_0 \rfloor$. (By assumption, $i_0 > 0$.)

Now choose m' such that $\lfloor \log_r m' \rfloor > |\alpha(i_0)|$. Again, some $i_1 \in \{0, 1, \ldots, m'-1\}$ satisfies $|\alpha(i_1)| \geq \lfloor \log_r m' \rfloor > \lfloor \log_r i_1 \rfloor$, and, by assumption, $i_1 > i_0$. And so on. QED

As an application, we can see an "information-theoretical" proof of

Fact (Euclid) There are infinitely many prime numbers.

Proof Suppose to the contrary, that there are only $p_1, \ldots, p_M$. This would induce a notation $\alpha : \mathbb{N} \to \{0, 1, \#\}$, for $n = p_1^{\beta_1} \ldots p_M^{\beta_M}$,

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

where $bin(\beta)$ is the usual binary notation for β ($|bin(\beta)| \leq 1 + \log_2 \beta$). Since $2^{\beta_i} \leq p_i^{\beta_i} \leq n$, we have $\beta_i \leq \log_2 n$, for all i. Consequently

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

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

Codes

Any mapping $\varphi : S \to \Sigma^*$ can be naturally extended to the morphism $\hat{\varphi} : S^* \to \Sigma^*$,

$$\hat{\varphi}(s_1 \ldots s_\ell) = \varphi(s_1) \ldots \varphi(s_\ell)$$

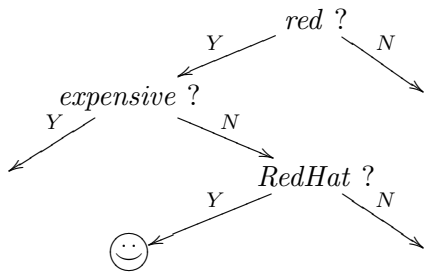
Definition A notation $\varphi : S \to \Sigma^*$ for a finite non-empty set S is a *code* if $\hat{\varphi}$ is 1:1. A code is *instantaneous* (prefix–free) if moreover $\neg\hat{\varphi}(s) \leq \hat{\varphi}(s')$, for $s \neq s'$.

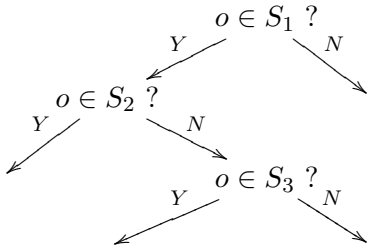
Note that the property of being an (instantaneous) code depends only on the set $\hat{\varphi}(S)$. Notice that $\varepsilon \notin \hat{\varphi}(S)$ (why ?). Any prefix-free set is a code, the set $\{aa, baa, ba\}$ is example of a non-instantaneous code, while $\{a, ab, ba\}$ is not a code at all.

In the sequel we will usually omit "hat" and identify $\hat{\varphi}$ with φ.

Clearly, in order to encode a set S of m elements with an alphabet Σ of r letters (with $m, r \geq 2$, say), it is enough to use strings of length $\lceil \log_r m \rceil$, so that $|\varphi(w)| \leq |w| \cdot \lceil \log_r m \rceil$, for $w \in S^*$. However, in order to make the coding more efficient, i.e., to keep $|\varphi(w)|$ as short as possible, it is useful to use shorter strings for those elements of S which occur more frequently.

There is an analogy between efficient codes and strategies in a so-called *20 question game*. In this game one person invents an object o (presumably, from some large set S), and the remaining players try to guess it by asking questions (normally, up to 20), the answers to which can be only *yes* or *no*. So the questions are generally of the form $o \in S'$ *?*, where $S' \subseteq S$.

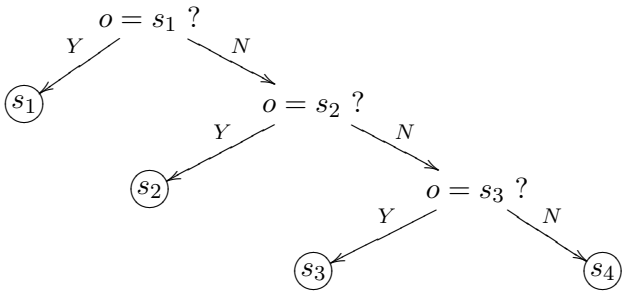




Clearly, $\lceil \log_2 |S| \rceil$ questions suffice to identify any object in S. Can we do better?

In general of course not, since a tree with 2^k leaves must have depth at least k. However, if some objects are more *probable* than others, we can improve the *expected* number of questions. (Besides, this feature makes the real game interesting.)

Suppose the elements of a set $S = \{s_1, s_2, s_3, s_4\}$ are given with probabilities $p(s_1) = \frac{1}{2}$, $p(s_2) = \frac{1}{4}$, $p(s_3) = p(s_4) = \frac{1}{8}$. Then the strategy



guarantees the expected number of questions

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

which is less than $\lceil \log_2 4 \rceil = 2$.

In general, any binary tree with leaves labeled by elements of a finite set S represents some strategy for the game over S (if we neglect the 20 threshold). All questions can be reconstructed bottom-up from the leaves, so we need not bother about them. Identifying directions *left* and *right* with 0 and 1, respectively, we have a mapping $S \to \{0,1\}^*$, which sends each s to the corresponding leaf. In the example above, this would be

$$s_1 \mapsto 0,\ s_2 \mapsto 10,\ s_3 \mapsto 110,\ s_4 \mapsto 111.$$

Clearly, this mapping is an instantaneous code, in which the maximal (expected) length of a code word equals the maximal (expected) number of questions.

The situation can be extended to the case of $|\Sigma| = r \geq 2$. We do not develop a corresponding game, but will often explore the correspondence between instantaneous codes and r-ary trees.

Generally, a *tree over a set* X (or X-tree, for short) is any non-empty set $T \subseteq X^*$ closed under prefix relation (denoted $\leq$). In this context, an element w of T is a *node* of level $|w|$, ε is the *root*, $\leq$-maximal nodes are *leaves*, a node wv (with $w, v \in X^*$) is *below* w, and wx (with $x \in X$) is an immediate *successor* (or child) of w. A *subtree* of T induced by $w \in T$ is $T_w = \{v : wv \in T\}$.

Now, any instantaneous code $\varphi : S \to \Sigma^*$ induces a tree over Σ, $T_\varphi = \{w : \text{for some } s,\ w \leq \varphi(s)\}$. Conversely, any tree $T \subseteq \Sigma^*$ with $|S|$ leaves induces an instantaneous code; in fact many ($|S|!$) codes, depending on permutation of S.

As mentioned above, our goal is to optimize the code length, keeping the resistance for transmission errors. The following is the first step toward the first objective.
Given a code $\varphi : S \to \Sigma^*$, let $|\varphi| : S \to \mathbb{N}$ denote the *length function*, given by $|\varphi|(s) = |\varphi(s)|$.

Theorem (Kraft inequality) Let $2 \leq |S| < \infty$ and $|\Sigma| = r$. A function $\ell : S \to \mathbb{N}$ is the length function, i.e., $\ell = |\varphi|$, for some instantaneous code $\varphi : S \to \Sigma^*$, if and only if

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

Proof ($\Rightarrow$) If all words $\varphi(s)$ have the same length k then, considering that φ is 1:1, we clearly have

$$\sum_{s \in S} \frac{1}{r^{|\varphi(s)|}} \leq \frac{r^k}{r^k} = 1. \quad (*)$$

More generally, let k be the maximal length of all $\varphi(s)$'s. For any s with $|\varphi(s)| = i$, let

$$P_s = \{\varphi(s)v : v \in \Sigma^{k-i}\}$$

(in other words, this is the set of nodes of level k below $\varphi(s)$ in the full Σ-tree). Clearly

$$\sum_{w \in P_s} \frac{1}{r^{|w|}} = \frac{r^{k-i}}{r^k} = \frac{1}{r^i}$$

and the sets P_s, $P_{s'}$ are disjoint for $s \neq s'$. Hence again

$$\sum_{s \in S} \frac{1}{r^{|\varphi(s)|}} = \sum_{s \in S} \sum_{w \in P_s} \frac{1}{r^{|w|}} \leq \frac{r^k}{r^k} = 1.$$

($\Leftarrow$) Let us enumerate $S = \{s_1, \ldots, s_m\}$ in such a way that $\ell(s_1) \leq \ldots \leq \ell(s_m)$. For $i = 0, 1, \ldots, m-1$, we inductively define $\varphi(s_{i+1})$ to be the first *lexicographically* element w of $\Sigma^{\ell(i+1)}$ which is not comparable to any of $\varphi(s_1), \ldots, \varphi(s_i)$ w.r.t. the prefix ordering $\leq$. It remains to show that there is always such w. Like in the previous case, let P_{s_j} be the set of nodes of level $\ell(s_{i+1})$ below $\varphi(s_j)$, we have $|P_{s_j}| = r^{\ell(i+1)-\ell(j)}$. We need to verify that

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

which amounts to

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

This follows directly from the hypothesis; we may assume that the inequality is strict since $i < m$. QED

28.02.2006.

If a code is not instantaneous, the Kraft inequality still holds, but the argument is more subtle.

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

Proof The case of $|S| = 1$ is trivial, and if $|S| \geq 2$ then $r = |\Sigma| \geq 2$ as well. It is then enough to show that φ satisfies the Kraft inequality. Let $K = \sum_{s \in S} \frac{1}{r^{|\varphi(s)|}}$. Suppose to the contrary that $K > 1$. 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 $i = |\varphi(q_1)| + \ldots + |\varphi(q_n)| = |\varphi(q_1 \ldots q_n)|$. Since φ is a code, at most one such sequence can be encoded by a word in Σ^i, hence

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

This follows

$$K^n \leq (Max - Min) \cdot n + 1$$

which clearly fails for sufficiently large n. The contradiction proves that $K \leq 1$. QED

Properties of convex functions

Before proceeding with further investigation of codes, we need to recall some concepts from the calculus.

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

$$\lambda f(x_1) + (1 - \lambda) f(x_2) \quad \geq \quad f(\lambda x_1 + (1 - \lambda) x_2). \tag{2}$$

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

Geometrically, it means that any chord linking two points of the function graph lies (strictly) above the graph.

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).

Proof Assume $f'' \geq 0$. Then by the Mean value theorem, f' is weakly increasing on (a, b) (for $a < t_1 < t_2 < b$, $f'(t_2) - f'(t_1) = f''(\tilde{t})(t_2 - t_1) \geq 0$).

Let $x_\lambda = \lambda x_1 + (1 - \lambda) x_2$. Rearranging our formula a bit, we have to show

$$\lambda(f(x_\lambda) - f(x_1)) \stackrel{?}{\leq} (1 - \lambda)(f(x_2) - f(x_\lambda)).$$

Using the Mean value theorem, this time for f, it reduces to

$$\begin{aligned} \lambda f'(\tilde{x_1})(x_\lambda - x_1) \quad &\stackrel{?}{\leq} \quad (1 - \lambda) f'(\tilde{x_2})(x_2 - x_\lambda) \\ \lambda(1 - \lambda) f'(\tilde{x_1})(x_2 - x_1) \quad &\stackrel{?}{\leq} \quad \lambda(1 - \lambda) f'(\tilde{x_2})(x_2 - x_1), \end{aligned}$$

which holds since f' is weakly increasing. If $f'' > 0$ the argument is similar. QED

In this course, unless stated otherwise, we consider only *finite* probabilistic spaces. If we say that X is a *random variable* on S, we tacitly assume that S is given with *probability* mapping $p : S \to [0,1]$ (i.e., $\sum_{s\in S} p(s) = 1$), and $X : S \to \mathbb{R}$. Recall that the *expected value* of X is

$$EX = \sum_{s\in S} p(s) \cdot X(s).$$

If $S = \{s_1, \ldots, s_m\}$, we adopt the notation $p(s_i) = p_i$, $X(s) = x_i$. In this writing $EX = p_1x_1 + \ldots + p_mx_m$.
Note that EX does not depend on those x_i's for which $p_i = 0$. We say that X is *constant* if there are no $x_i \neq x_j$ with $p_i, p_j > 0$.

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

$$Ef(X) \quad \geq f(EX). \tag{3}$$

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

Proof By induction on $|S|$. The case of $|S| = 1$ is trivial, and if $|S| = 2$, the inequality amounts to

$$p_1f(x_1) + p_2f(x_2) \quad \geq (>) \quad f(p_1x_1 + p_2x_2)$$

which is just the definition of (strict) convexity. (Note that X is constant iff $p_1 \in \{0,1\}$ or $x_1 = x_2$.)

Let $S = \{s_1, \ldots, s_m\}$, and suppose the claim holds for any random variables over S', $|S'| \leq m-1$.

Without loss of generality we may assume that $p_m < 1$. Let $p'_i = \frac{p_i}{1-p_m}$, for $i = 1, \ldots, m-1$. We have

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

Note that we have used the induction hypothesis twice: for the random variable given by probabilities $p'_1, \ldots, p'_{m-1}$ and values $x_1, \ldots, x_{m-1}$, and for the random variable given by probabilities $p_m, 1-p_m$, and values x_m and $\sum_{i=1}^{m-1} p'_i x_i$, respectively.

Now suppose f is strictly convex and in the above the equalities hold. Then the first auxiliary random variable is constant, i.e., $x_i = C$, for all $i = 1, \ldots, m-1$, unless $p'_i = p_i = 0$. Since the second auxiliary random variable must also be constant, we have, whenever $p_m > 0$, $x_m = \sum_{i=1}^{m-1} p'_i x_i = C$, as well. QED

Convention We let

$$0 \log_r 0 = 0 \log_r \frac{1}{0} = 0$$

This is justified by the fact that $\lim_{x\to 0} x \log_r x = \lim_{x\to 0} -x \log_r \frac{1}{x} = \lim_{|y|\to\infty} -\frac{\log_r y}{y} = 0$.
From the above lemma, we deduce that, if $r > 1$ then the function $x \log_r x$ is strictly convex on $[0, \infty)$ (i.e., on any $[0, M]$, $M > 0$). Indeed,

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

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

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

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

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

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

Applying Jensen's inequality to function $x \log_r x$ (on $[0, \infty)$), we get

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

Here we consider the random variable which takes the value $\left(\frac{x_i}{y_i} \right)$ with probability y_i. As the function $x \log_r x$ is even strictly convex on $[0, \infty)$ (c.f. page 7), the equality implies that this random variable is constant. Remembering that $y_i > 0$, and $\sum_{i=1}^{q} x_i = \sum_{i=1}^{q} y_i$, we then have $x_i = y_i$, for $i = 1, \ldots, q$.

Now suppose $\sum_{i=1}^{q} y_i < 1$. Let $y_{q+1} = 1 - \sum_{i=1}^{q} y_i$, and $x_{q+1} = 0$. Then, by the previous case we have

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

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

Entropy

We come back to the strategy presented on page 3. The number of questions it needs to identify an object s_i is precisely $\log_2 \frac{1}{p(s_i)}$. So, the expected number of questions is $\sum_{i=1}^{m} p(s_i) \cdot \log_2 \frac{1}{p(s_i)}$. It is possibly since probabilities in that game are powers of $\frac{1}{2}$.
Using the Golden Lemma, we can see that this number of questions is optimal. For, consider any strategy, with the number of questions to identify s_i equal $\ell(s_i)$. By the Kraft inequality $\sum_{i=1}^{m} \frac{1}{2^{\ell(s_i)}} \leq 1$.

Taking in the Golden Lemma $x_i = p(s_i)$ and $y_i = \frac{1}{2^{\ell(s_i)}}$, we obtain

$$\sum_{i=1}^{m} p(s_i) \cdot \ell(s_i) \quad \geq \quad \sum_{i=1}^{m} p(s_i) \cdot \log_2 \frac{1}{p(s_i)}.$$

7.03.2006.

The right-hand side of the inequality above makes sense even if the $p(s)$'s are not powers of $\frac{1}{2}$. We thus arrive to the central concept of Information Theory.

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

$$H_r(S) = \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)} \tag{4}$$

$$= -\sum_{s \in S} p(s) \cdot \log_r p(s). \tag{5}$$

In other words, $H_r(S)$ is the expected value of a random variable defined on S by $s \mapsto \log_r \frac{1}{p(s)}$.

Traditionally, we abbreviate $H = H_2$.

Remark We may note that the concept of entropy combines two ideas:

- computing the mean value of some function composed with probability, $\sum_{s \in S} p(s) \cdot f \circ p(s)$,
- choosing $f = \log$, which is perhaps most important.

Indeed, function log plays a crucial role in the so-called *Weber-Fechner law* of cognitive science, stating that the human *perception* (P) of the growth of a physical *stimuli* (S), is proportional to the *relative* growth of the stimuli rather than to its absolute growth,

$$\partial P \quad \approx \quad \frac{\partial S}{S}$$

which, after integration, gives us

$$P \quad \approx \quad \log S.$$

This has been observed in perception of weight, brightness, sound (both intensity and height), and even one's economic status. In this context, we might view entropy as our "perception of probability".

What values entropy can take, depending on $|S|$ and p ? From definition we readily have $H_r(S) \geq 0$, and this value is indeed attained if the whole probability is concentrated in one point. On the other hand, we have

Fact

$$H_r(S) \quad \leq \quad \log_r |S| \tag{6}$$

and the equality holds if and only if $p(s) = \frac{1}{|S|}$, for all $s \in S$.

Proof Indeed, taking in the Golden Lemma $x_i = p(s_i)$ and $y_i = \frac{1}{|S|}$, we obtain

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

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

As we have seen, if all probabilities are powers of $\frac{1}{2}$ then the entropy equals to the (average) length of an optimal code. We will see that it is always a lower bound.

Definition (minimal code length) For a code φ, let

$$L(\varphi) = \sum_{s \in S} p(s) \cdot |\varphi(s)|.$$

Given S and integer $r \geq 2$, let $L_r(S)$ be the minimum of all $L(\varphi)$'s, where φ ranges over all codes $\varphi : S \to \Sigma^*$, with $|\Sigma| = r$.

Note that, because of the McMillan Theorem (page 4), the value of $L_r(S)$ would not change if φ have ranged over instantaneous codes.

Theorem For any finite probabilistic space S

$$H_r(S) \leq L_r(S) \tag{7}$$

and the equality holds if and only if all probabilities $p(s)$ are powers of $\frac{1}{r}$.

Proof For the first half of the claim, it is enough to show that

$$H_r(S) \leq L(\varphi)$$

holds for any code $\varphi : S \to \Sigma^*$, with $|\Sigma| = r$. We obtain this readily taking in the Golden Lemma $x_i = p(s_i)$ and $y_i = \frac{1}{r^{|\varphi(s)|}}$.

Now, if the equality $H_r(S) = L_r(S)$ holds then we have also $H_r(S) = L(\varphi)$, for some code φ. Again from Golden Lemma, we obtain $p(s) = \frac{1}{r^{|\varphi(s)|}}$, for all $s \in S$.

On the other hand, if each probability $p(s)$ is of the form $\frac{1}{r^{\ell(s)}}$, then by the Kraft inequality, there exists a code φ with $|\varphi(s)| = \ell(s)$, and for this code $L(\varphi) = H_r(S)$. Hence $L_r(S) \leq H_r(S)$, but by the previous inequality, the equality must hold. QED

The second part of the above theorem may appear pessimistic, as it infers that in most cases our coding is "imperfect" ($H_r(S) < L_r(S)$). Note that probabilities usually are not chosen by us, but rather come from Nature.

However, it turns out that, even with a *fixed* S and p we can, in a sense, bring the average code length closer and closer to $H_r(S)$. This is achieved by some relaxation of the concept of a code.

Example Let $S = \{s_1, s_2\}$ with $p(s_1) = \frac{3}{4}$, $p(s_2) = \frac{1}{4}$. Then clearly $L_2(S) = 1$. (The reader may convince herself or himself by elementary calculation that $H_2(S) < 1$, in accordance with our Theorem.)

This means that we are unable to make the encoding of a message $\alpha \in S^*$ shorter than α itself, even on average. Now, consider the following mapping:

$s_1 s_1 \mapsto 0 \qquad s_1 s_2 \mapsto 10$

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

Of course, this is not a code of S, but apparently we could use this mapping to encode sequences over S of even length. Indeed, it *is* a code for the set S^2. Consider $S^2 = S \times S$ as the product (probabilistic) space with

$$p(s_i, s_j) \quad = \quad p(s_i) \cdot p(s_j).$$

Then the average length of our encoding of the *two*-symbols blocks is

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

As the reader may expect, if we proceed in this vein for $n = 2, 3, \ldots$, we can obtain more and more efficient encodings. But can we overcome the entropy bound, i.e., to get

$$\frac{L_r(S^n)}{n} \quad \stackrel{?}{<} \quad H_r(S)$$

for some n ?

We will see that this is *not* the case, but the Shannon First Theorem (next lecture) will tell us that the entropy bound can be approached arbitrarily close, as $n \to \infty$.

To this end, we have first to find the entropy $H(S^n)$ of S^n viewed as the product space. This could be done by a tedious elementary calculation, but we prefer to deduce the formula from general properties of random variables.

Recall that the expected value of a random variable $X : S \to \mathbb{R}$ can be presented in two ways, readily equivalent to each other:

$$EX = \sum_{s \in S} p(s) \cdot X(s) = \sum_{t \in X(S) \subseteq \mathbb{R}} t \cdot p(X = t).$$

The last term is often written simply as

$$\sum_{t \in \mathbb{R}} t \cdot p(X = t),$$

where we assume that the sum of arbitrarily many 0's is 0.

The notation $p(X = t)$ used above is a particular case of $p(\psi(X))$, for some formula ψ, which denotes the *probability that* $\psi(X)$ *holds*, i.e., the sum of $p(s)$'s, for those s, for which $\psi(X(s))$ is satisfied.

We recall a basic fact from Probability Theory.[1]

[1]The reader may verify it by elementary calculation, using conditional probabilities.

Linearity of expectation If X and Y are arbitrary random variables (defined on the same probabilistic space) then, for any $\alpha, \beta \in \mathbb{R}$,

$$E(\alpha X + \beta Y) \quad = \quad \alpha EX + \beta EY. \tag{8}$$

Now consider two probabilistic spaces S and Q. (According to the tradition, if confusion does not arise, we use the same letter p for the probability functions on all spaces.)

Let $S \times Q$ be the product space, with the probability given by

$$p(s,q) \quad = \quad p(s) \cdot p(q).$$

Given random variables $X : S \to \mathbb{R}$ and $Y : Q \to \mathbb{R}$, we define the random variables $\hat{X}$, $\hat{Y}$, over $S \times Q$, by

$$\begin{aligned} \hat{X}(s,q) &= X(s) \\ \hat{Y}(s,q) &= Y(q). \end{aligned}$$

Note[2] that

$$p(\hat{X} = t) = \sum_{\hat{X}(s,q)=t} p(s,q) = \sum_{X(s)=t} \sum_{q \in Q} p(s) \cdot p(q) = \sum_{X(s)=t} p(s) \; = \; P(X = t).$$

Similarly, $p(\hat{Y} = t) = p(Y = t)$.
Therefore, $E\hat{X} = EX$ and $E\hat{Y} = EY$. By linearity of expectation,

$$E(\hat{X} + \hat{Y}) = E\hat{X} + E\hat{Y} = EX + EY.$$

Let in the above $X : s \mapsto \log_r \frac{1}{p(s)}$, and $Y : q \mapsto \log_r \frac{1}{p(q)}$. Then

$$(\hat{X} + \hat{Y})(s,q) = \log_r \frac{1}{p(s)} + \log_r \frac{1}{p(q)} = \log_r \frac{1}{p(s)} \cdot \frac{1}{p(q)} = \log_r \frac{1}{p(s,q)}.$$

But, by the remark after definition of entropy (c.f. 8), this is precisely the random variable on $S \times Q$ whose expected value amounts to the entropy of $S \times Q$, i.e., $H_r(S \times Q) = E(\hat{X} + \hat{Y})$. Hence, the equation above gives us

$$H_r(S \times Q) \quad = \quad H_r S + H_r Q. \tag{9}$$

Consequently,

$$H_r S^n \quad = \quad n \cdot H_r S. \tag{10}$$

14. 03. 2006

In order to estimate $\frac{L_r(S^n)}{n} - H_r(S)$, we first complete the inequality of the Theorem from page 9 by the upper bound.

[2]Throughout these notes, we generally use notation $\sum_{\psi(a_1,\ldots,a_k)} t(a_1,\ldots,a_k)$, for the sum of terms $t(a_1,\ldots,a_k)$, where $(a_1,\ldots,a_k)$ ranges over all tuples satisfying $\psi(a_1,\ldots,a_k)$.

Theorem (Shannon-Fano coding) For any finite probabilistic space S and $r \geq 2$, there is a code $\varphi : S \to \Sigma^*$ (with $|\Sigma| = r$), satisfying

$$L(\varphi) \quad \leq \quad H_r(S) + 1.$$

Consequently

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

Moreover, the strict inequality $L_r(S) < H_r(S) + 1$ holds unless $p(s) = 1$, for some $s \in S$ (hence $H_r(S) = 0$).

Proof For $|S| = 1$, we have trivially $H_r(S) = 0$ and $L_r(S) = 1$. Assume $|S| \geq 2$. We let

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

for those $s \in S$ for which $p(s) > 0$. Then

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

We consider several cases. If $(\forall s \in S)\, p(s) > 0$, then the above coincides with the Kraft inequality, and hence there is a code φ satisfying $|\varphi(s)| = \ell(s)$, for $s \in S$. But as $\ell(s) < \log_r \frac{1}{p(s)} + 1$, we obtain

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

Now suppose that $p(s)$ may be 0, for some s. If

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

then we can readily extend the definition of ℓ to all s, such that the Kraft inequality $\sum_{s \in S} \frac{1}{r^{\ell(s)}} \leq 1$ is satisfied. Again, there is a code with length ℓ, satisfying $\ell(s) < \log_r \frac{1}{p(s)} + 1$, whenever $p(s) > 0$, and hence

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

(Remember our convention that $0 \cdot \log \frac{1}{0} = 0$.)

Finally, suppose that

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

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

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

Now again we can extend ℓ' to all s in such a way that the Kraft inequality holds. In order to evaluate the average length of this code, let us first observe that our assumptions yield that $\ell(s) = \log_r \frac{1}{p(s)}$, whenever $p(s) > 0$. (Indeed, we have $\frac{1}{r^{\ell(s)}} \leq p(s)$ by definition of ℓ, and $1 = \sum_{p(s)>0} \frac{1}{r^{\ell(s)}} = \sum_{p(s)>0} p(s)$, hence $p(s) = \frac{1}{r^{\ell(s)}}$, whenever $p(s) > 0$.) Then the code with length ℓ' satisfies

$$\sum_{s\in S} p(s)\cdot \ell'(s) = \sum_{p(s)>0} p(s)\cdot \ell'(s) = p(s') + \sum_{p(s)>0} p(s)\cdot \ell(s) = p(s') + H_r(S).$$

Hence we get $L_r(S) \leq H_r(S) + 1$ and the inequality is strict unless we cannot find s' with $0 < p(s') < 1$. QED

We are ready to state

Shannon's First 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 We have from the previous theorem

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

but since $H_r(S^n) = n \cdot H_r(S)$,

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

which yields the claim. QED

Relative entropy and mutual information

Entropy of random variable If $X : S \to \mathcal{X}$ is a random variable, we let

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

Thus, $H_r(X)$ amounts to the expected value

$$H_r(X) \quad = \quad E\left(\log_r \frac{1}{p(X)}\right),$$

where $p(X)$ is the random variable on S, given by $p(X) : s \mapsto p(X = X(s))$. Indeed,

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

Notational conventions: If the actual random variables are known from the context, we often abbreviate the event $X = a$ by just a; so we may write, e.g., $p(x|y)$ instead of $p(X = x|Y = y)$, $p(x \wedge y)$ instead of $p((X = x) \wedge (Y = y))$, etc.

Conditional entropy Let $A : S \to \mathcal{A}$, $B : S \to \mathcal{B}$, be two random variables. For $b \in \mathcal{B}$, let

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

Now let

$$H_r(A|B) \quad = \quad \sum_{b \in \mathcal{B}} p(b) H_r(A|b).$$

Note that if A and B are independent then, in the above formula $p(a|b) = a$, and hence $H_r(A|B) = A$. On the other hand, $H_r(A|A) = 0$; more generally, if $\varphi : \mathcal{A} \to \mathcal{B}$ is any function then

$$H_r(\varphi(A)|A) \quad = \quad 0. \tag{11}$$

Indeed, if $p(A = a) > 0$ then $p(\varphi(A) = \varphi(a)|A = a) = 1$, and hence $\log_r \frac{1}{p(\varphi(A)=\varphi(a)|A=a)} = 0$.
We will see more properties of the the conditional entropy in the sequel.

Joint entropy We also consider the couple (A, B) as a random variable $(A, B) : S \to \mathcal{A} \times \mathcal{B}$,

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

Note that the probability that this variable takes value (a, b) is $p\left((A, B) = (a, b)\right) = p\left((A = a) \wedge (B = b)\right)$, which we abbreviate by $p(a \wedge b)$. This probability is, in general, different from $p(a) \cdot p(b)$. In the case if, for all $a \in \mathcal{A}, b \in \mathcal{B}$,

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

(i.e., the events $A = a$ and $B = b$ are independent), the variables A and B are called *independent*.

Now $H_r(A, B)$ is well defined by

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

Note that if A and B are independent then

$$\log_r \frac{1}{p(A, B)} \quad = \quad \log_r \frac{1}{p(A)} + \log_r \frac{1}{p(B)},$$

Remembering the characterization $H_r(X) = E\left(\log_r \frac{1}{p(X)}\right)$, we have, by linearity of expectation (8),

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

In general case we have the following.

Theorem

$$H_r(A,B) \quad \leq \quad H_r(A) + H_r(B). \tag{12}$$

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

Proof We rewrite the right-hand side a bit, in order to apply the Golden Lemma. We use the obvious equalities $p(a) = \sum_{b\in\mathcal{B}} p(a \wedge b)$, and $p(b) = \sum_{a\in\mathcal{A}} p(a \wedge b)$.

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

Note that the above expression is well defined, because if $p(a) = 0$ or $p(b) = 0$ then $p(a \wedge b) = 0$, as well.

Let us momentarily denote

$$(\mathcal{A}\times\mathcal{B})^+ \quad = \quad \{(a,b) : p(a) > 0 \text{ and } p(b) > 0\}.$$

We have clearly

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

Then, applying the Golden Lemma (page 7) to $x = p(a \wedge b)$, $y = p(a) \cdot p(b)$, where (a, b) ranges over $(\mathcal{A} \times \mathcal{B})^+$, we obtain

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

Moreover, the equality holds only if $p(a \wedge b) = p(a) \cdot p(b)$, for all $(a, b) \in (\mathcal{A} \times \mathcal{B})^+$, and consequently, for all $a \in \mathcal{A}$, $b \in \mathcal{B}$. On the other hand, we have already seen that independence of A and B implies this equality. QED

Definition (information) The value

$$I(A;B) \quad = \quad H_r(A) + H_r(B) - H_r(A,B). \tag{13}$$

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

Remark The above concepts and properties have some interpretation in terms of *20 questions game* (page 2). Suppose an object to be identified is actually a couple (a, b), where a and b are values of random variables A and B, respectively. Now, if A and B are independent, we can do nothing better than identify a and b separately. Thus our series of questions splits into "questions about a" and "questions about b", which is reflected by the equality $H_r(A, B) = H_r(A)+H_r(B)$. However, if A and B are dependent, we can take advantage of mutual information and decrease the number of questions.

21.03.2006.

To increase readability, since now on we will omit subscript r, writing H, I, ..., instead of H_r, I_r, ... Unless stated otherwise, all our results apply to any $r > 1$. Without loss of generality, the reader may assume $r = 2$.

Remark From the transformations used in the proof of the theorem above, we easily deduce

$$I(A;B) = \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a\wedge b)\left(\log\frac{1}{p(a)p(b)} - \log\frac{1}{p(a\wedge b)}\right). \tag{14}$$

Hence $I(A;B)$ can be viewed as a measure of the distance between the actual distribution of the joint variable $(A;B)$ and its distribution if A and B were independent.

Note that the above sum is nonnegative, although some summands $\left(\log\frac{1}{p(a)p(b)} - \log\frac{1}{p(a\wedge b)}\right)$ can be negative.

The following property generalizes the equality $H(A, B) = H(A)+H(B)$ to the case of dependent variables.

Fact (Chain rule)

$$H(A,B) = H(A|B) + H(B). \tag{15}$$

Proof Just calculate:

$$\begin{aligned}
H(A,B) &= \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a\wedge b)\cdot\log\frac{1}{p(a\wedge b)}\\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a|b)p(b)\cdot\log\frac{1}{p(a|b)p(b)}\\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p(a|b)p(b)\cdot\left(\log\frac{1}{p(a|b)} + \log\frac{1}{p(b)}\right)\\
&= \sum_{b\in\mathcal{B}} p(b)\cdot\sum_{a\in\mathcal{A}} p(a|b)\cdot\log\frac{1}{p(a|b)} + \sum_{b\in\mathcal{B}} p(b)\log\frac{1}{p(b)}\cdot\sum_{a\in\mathcal{A}} p(a|b)\\
&= H(A|B) + H(B).
\end{aligned}$$

QED

Applying the chain rule, we get alternative formulas for information:

$$I(A;B) = H(A) - H(A|B) \tag{16}$$
$$= H(B) - H(B|A). \tag{17}$$

This also implies that $I(A;B) \leq \min\{H(A), H(B)\}$.
The Chain rule generalizes easily to the case of $n \geq 2$ variables $A_1, A_2, \ldots, A_n$.

$$\begin{aligned} H(A_1, \ldots, A_n) &= H(A_1|A_2, \ldots, A_n) + H(A_2, \ldots, A_n) \\ &= H(A_1|A_2, \ldots, A_n) + H(A_2|A_3, \ldots, A_n) + H(A_3, \ldots, A_n) \\ &= \sum_{i=1}^{n} H(A_i|A_{i+1}, \ldots, A_n) \end{aligned} \tag{18}$$

if we adopt convention $H(A|\emptyset) = A$.
A more subtle generalization follows from relativization.

Conditional chain rule

$$H(A, B|C) = H(A|B, C) + H(B|C). \tag{19}$$

Proof We have

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

In the above, a and b range over those values for which the respective conditional probabilities are defined[3]. We use the fact that, whenever $p(a \wedge b|c) > 0$,

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

By taking the average over $p(c)$, we further have

$$\begin{aligned} H(A, B|C) &= \sum_{c \in \mathcal{C}} p(c) \cdot H(A, B|c) \\ &= \sum_{c} p(c) \cdot \sum_{b} p(b|c) \cdot \sum_{a} p(a|b \wedge c) \cdot \log \frac{1}{p(a|b \wedge c)} + \sum_{c} p(c) \cdot \sum_{b} p(b|c) \cdot \log \frac{1}{p(b|c)} \\ &= \underbrace{\sum_{b,c} p(b \wedge c) \cdot \sum_{a} p(a|b \wedge c) \cdot \log \frac{1}{p(a|b \wedge c)}}_{H(A|B,C)} + \underbrace{\sum_{c} p(c) \cdot \sum_{b} p(b|c) \cdot \log \frac{1}{p(b|c)}}_{H(B|C)}, \end{aligned}$$

as required. QED
We leave to the reader to show that

$$H(A, B|C) \leq H(A|C) + H(B|C) \tag{20}$$

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

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

The proof can go along the same lines as on the page 15.

[3]Recall that $p(x|y)$ is undefined if $p(y) = 0$.

Conditional information We let the *mutual information of A and B given C* be defined by

$$\begin{aligned} I(A;B|C) &= H(A|C) + H(B|C) - \underbrace{H(A,B|C)}_{H(A|B,C)+H(B|C)} \quad (21) \\ &= H(A|C) - H(A|B,C). \quad (22) \end{aligned}$$

Finally, let *mutual information of A, B, and C* be defined by

$$R(A;B;C) = I(A;B) - I(A;B|C). \quad (23)$$

Let us see that this definition is indeed symmetric, i.e., does not depend on the particular ordering of A, B, C:

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

Note however, that in contrast to $I(A;B)$ and $I(A;B|C)$, $R(A;B;C)$ can be *negative*.

The set of equations relating the quantities $H(X), H(Y), H(Z), H(X,Y), H(X,Y|Z), I(X;Y), I(X;Y|Z)$, $R((X;Y;Z), \ldots,$ can be pictorially represented by the so-called *Venn diagram*. (See the Internet; note however that this is only a helpful representation without *extra* meaning.)

Application: Perfect secrecy

A *cryptosystem* is a triple of random variables:

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

Moreover, there must be 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 $I(C;M) = 0$.

Example: One time pad Here $\mathcal{M} = \mathcal{K} = \mathcal{C} = \{0,1\}^n$, for some $n \in \mathbb{N}$, and

$$C = M \oplus K$$

where $\oplus$ is the component-wise *xor* (e.g., $101101 \oplus 110110 = 011011$). Hence $Dec(v,w) = v \oplus w$, as well. Moreover we assume that K has uniform distribution over $\{0,1\}^n$, i.e., $p(K = v) = \frac{1}{2^n}$, for $v \in \{0,1\}^n$, and that K and M are independent.

In order to show perfect secrecy, it is enough to prove that M and C are independent (see Theorem on page 15), i.e.

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

We have

$$\begin{aligned} p(C = w) &= \sum_{u \in \mathcal{M}} p(K = u \oplus w | M = u) \cdot p(M = u) \\ &= \sum_{u \in \mathcal{M}} p(K = u \oplus w) \cdot p(M = u) \text{ (by independence of } M \text{ and } K) \\ &= \sum_{u \in \mathcal{M}} \frac{1}{2^n} \cdot p(M = u) \\ &= \frac{1}{2^n}. \end{aligned}$$

On the other hand, we have, by definition of C and independence of M and K,

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

which gives the desired equality. QED

Exercise Show that the independence of M and K is really necessary to achieve perfect secrecy of one-time pad.

Shannon's Pessimistic Theorem Any perfectly secret cryptosystem satisfies

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

Consequently (c.f. the Shannon-Fano coding, page 12)

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

Roughly speaking, to guarantee perfect secrecy, the keys must be (almost) as long as messages, which is highly impractical.

Proof We have

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

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

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

By symmetry, we have

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

which gives the desired inequality. QED

28.03.2006.

We observe a property of information which at first sight may appear a bit surprising. Let A and B be random variables; we may think that A represents some experimental data, and B our knowledge about them. Can we increase the information about A by processing B (say, by analysis, computation, etc.)? It turns out that we cannot.

Lemma Suppose A and C are conditionally independent, given B (see page 17). Then

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

Proof First note the following *chain rule for information*:

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

By symmetry, and from the conditional independence of A and C

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

which yields the desired inequality. QED

Note that the equality holds iff, additionally, A and B are conditionally independent given C.

Corollary If f is a function then

$$I(A;f(B)) \quad \leq \quad I(A;B). \tag{24}$$

Proof Follows from the Lemma, since

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

QED

Channels

Definition 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 $\mathcal{A} \times \mathcal{B} \to [0,1]$, sending (a,b) to $P(a \to b)$, 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 with values in $\mathcal{A}$ and $\mathcal{B}$, respectively, form an *input-output* pair for the channel Γ if, for all $a \in \mathcal{A}, b \in \mathcal{B}$,

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

We visualize it by

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

Note that if A and B form an *input-output* pair then

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

Hence, the distribution of (A, B) forming an input-output pair is uniquely determined by A (for fixed Γ). In particular, a suitable B exists and its distribution is determined by

$$p(B = b) \quad = \quad \sum_{a \in \mathcal{A}} P(a \to b) \cdot p(A = a). \tag{25}$$

Knowing this, the reader may easily calculate $H(A, B)$, $H(B|A)$, $I(A; B)$, etc. (depending on Γ and A).

We define the *capacity* of the channel Γ by

$$C_\Gamma \quad = \quad \max_A I(A; B), \tag{26}$$

where, for concreteness, $I = I_2$. Here A ranges over all random variables with values in $\mathcal{A}$, and (A, B) forms an input-output pair for Γ. The maximum exists because $I(A; B)$ is a continuous mapping from the compact set $\{p \in [0, 1]^{\mathcal{A}} : \sum_{a \in \mathcal{A}} p(a) = 1\}$ to $\mathbb{R}$, which moreover is bounded since $I(A; B) \leq H(A) \leq \log |\mathcal{A}|$.

If $\mathcal{A} = \{a_1, \ldots, a_m\}$, $\mathcal{B} = \{b_1, \ldots, b_n\}$, then the channel can be represented by a matrix

$$\begin{pmatrix} P_{11} & \ldots & P_{1n} \\ \ldots & \ldots & \ldots \\ P_{m1} & \ldots & P_{mn}, \end{pmatrix}$$

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

The formula for distribution of B in matrix notation is

$$(p(a_1), \ldots, p(a_m)) \cdot \begin{pmatrix} P_{11} & \ldots & P_{1n} \\ \ldots & \ldots & \ldots \\ P_{m1} & \ldots & P_{mn}, \end{pmatrix} \quad = \quad (p(b_1), \ldots, p(b_n)) \,. \tag{27}$$

Examples

We can present a channel as a bipartite graph from $\mathcal{A}$ to $\mathcal{B}$, with an arrow $a \to b$ labeled by $P(a \to b)$ (if $P(a \to b) = 0$, the arrow is not represented).

Faithful (noiseless) channel Let $\mathcal{A} = \mathcal{B} = \{0, 1\}$.

$$0 \longrightarrow 0$$

$$1 \longrightarrow 1$$

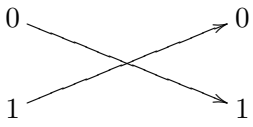
The matrix representation of this channel is

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

Since A is always a function of B, we have $I(A; B) = H(A)$, and hence the capacity is

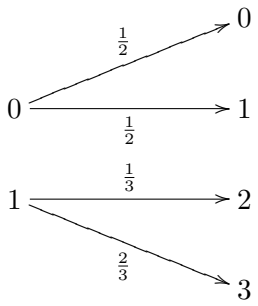
$$C_\Gamma = \max_A I(A; B) = \max_A H(A) = \log_2 |\mathcal{A}| = 1.$$

Inverse faithful channel



matrix representation: $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, capacity: $C_\Gamma = 1$.

Noisy channel without overlap Here $\mathcal{A} = \{0, 1\}$, $\mathcal{B} = \{0, 1, 2, 3\}$.



The matrix representation is

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

Here again A is a function of B, hence $I(A; B) = H(A) - H(A|B) = H(A)$, and therefore $C_\Gamma = 1$.

Noisy typewriter [4] Here we assume $\mathcal{A} = \mathcal{B} = \{a, b, \ldots, z\}$ (26 letters, say), and

$$p(\alpha \to \alpha) = p(\alpha \to next(\alpha)) = \frac{1}{2}$$

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

We leave to the reader to draw graphical and matrix representation.

To compute the capacity, first observe that, for any α,

$$H(B|\alpha) = p(\alpha|\alpha) \cdot \log \frac{1}{p(\alpha|\alpha)} + p(next(\alpha)|\alpha) \cdot \log \frac{1}{p(next(\alpha)|\alpha)} = (\frac{1}{2} + \frac{1}{2}) \cdot \log_2 = 1.$$

Hence

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

(the maximum is achieved for A with uniform distribution).

The reader may have already grasped that capacity is a desired value, like information, and unlike entropy. What are the channels with the minimal possible capacity, i.e., $C_\Gamma = 0$?

[4] *Typewriter* had been a manual device for typing, before a computer-served printers were invented (see old movies).

Bad channels Clearly $C_\Gamma = 0$ whenever $I(A;B) = 0$ for all input-output pairs, i.e., all such pairs are independent. This requires that $p(B = b|A = a) = p(B = b)$, for all $a \in \mathcal{A}$, $b \in \mathcal{B}$ (unless $p(A = a) = 0$), hence for a fixed b, all values $p(B = b|A = a)$ (i.e., all values in a column in the matrix representation) must be equal.

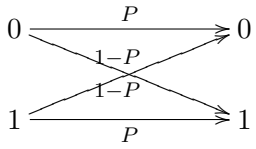
For example, the following channels have this property:

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

The last example is a particularly dull channel, which always outputs the same value. Note that in this case $H(B)$ is always 0, which means that the entropy may sometimes decrease while sending a message through a channel. However, in most interesting cases it actually increases. The following example is most important in our further studies.

Binary symmetric channel (BSC)

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



Letting $\bar{P} = 1 - P$, the matrix representation is

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

Prior to calculating C_Γ, we note the important property.

Fact If (A, B) forms an input-output pair for a BSC then

$$H(B) \;\geq\; H(A).$$

Moreover, the equality holds only if $P \in \{0, 1\}$ (i.e., the channel is faithful or inverse-faithful), or if $H(A) = 1$ (i.e., the entropy of A achieves the maximal value).

Proof Let $q = p(A = 0)$. Then $p(A = 1) = \bar{q}$, and we calculate the distribution of B by the formula

$$(q, \bar{q}) \cdot \begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix} = (\underbrace{qP + \bar{q}\bar{P}}_{p(B=0)}, \underbrace{q\bar{P} + \bar{q}P}_{p(B=1)})$$

Let $r = p(B = 0)$. Then

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

Recall our convention (page 13) that $0 \log_r 0 = 0 \log_r \frac{1}{0} = 0$, and let h denote the mapping

$$h(x) \;=\; x \ln x + (1 - x) \ln(1 - x),$$

defined for $0 \le x \le 1$. We easily calculate (for $0 < x < 1$)

$$\begin{aligned} h'(x) &= 1 + \ln x - 1 - \ln(1-x) \\ h''(x) &= \frac{1}{x} + \frac{1}{1-x} > 0. \end{aligned}$$

Hence by the lemma on convex functions (page 5), the function $h(x)$ is strictly convex on $[0,1]$, and it readily implies that so is the function

$$\log_2 e \cdot h(x) = x \log_2 x + (1-x)\log_2(1-x).$$

Taking in the definition of convexity $x_1 = q$, $x_2 = \bar{q}$, and $\lambda = P$ (hence $\lambda x_1 + (1-\lambda)x_2 = r$), and noting that $h(q) = h(\bar{q})$, we obtain that

$$\begin{aligned} q\log q + \bar{q}\log\bar{q} &\ge r\log r + \bar{r}\log\bar{r} \\ \text{i.e., } H(A) &\le H(B) \end{aligned}$$

and, moreover, the equality holds only if $P \in \{0,1\}$ or if $q = \bar{q}$, which holds iff $H(A) = \log_2|\{0,1\}| = 1$. QED

We are going to calculate C_Γ. It is convenient to use notation

$$H(s) = -s\log_2 s - (1-s)\log_2(1-s) \tag{28}$$

(justified by the fact that $H(s) = H(X)$, whenever $p(X=0) = s$, $p(X=1) = \bar{s}$). Note that $H(0) = H(1) = 0$, and the maximum of H in $[0,1]$ is $H(\frac{1}{2}) = 1$.

By the definition of conditional entropy, we have

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

Hence, $H(B|A)$ does not depend on A.

Now, by the calculation of the distribution of B above, we have

$$H(B) = H(qP + \bar{q}\bar{P})$$

which achieves the maximal value $1 = H(\frac{1}{2})$, for $q = \frac{1}{2}$. Hence

$$C_\Gamma = \max_A H(B) - H(B|A) = 1 - H(P). \tag{29}$$

4.04.2006

Decision rules

Suppose we receive a sequence of letters $b_{i_1}, \ldots, b_{i_k}$, transmitted through a channel Γ. Knowing the mapping $(P(a \to b))$ (for $a \in \mathcal{A}$, $b \in \mathcal{B}$), can we decode the message?

In some cases the answer is simple. For example, in the inverse faithful channel (page 22), we should just interchange 0 and 1. However, for the noisy typewriter (page 22), no unique decoding exists. For instance, an output word *afu*, can result from input *zet*, but also from *aft*, and many others[5] (but not, e.g., from input *abc*).

In general, the objective of the receiver is, given an output letter b, to "decide" what input symbol a has been sent. Clearly the receiver wants to maximize $p(A = a|B = b)$. A *decision rule* is any mapping $\Delta : \mathcal{B} \to \mathcal{A}$.

The quality of the rule is measured by

$$Pr_C(\Delta, A) \quad =_{def} \quad p(\Delta \circ B = A), \tag{30}$$

where (A, B) forms the input–output pair[6]. Using conditional probabilities we can compute it in several ways, for example by

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

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

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

We can compute it by

$$Pr_E(\Delta, A) \quad = \quad \sum_{a \in \mathcal{A}} p(A = a) \cdot p(\Delta \circ B \neq a | A = a) \tag{31}$$

We are interested in rules maximizing $Pr_C(\Delta, A)$, and thus minimizing $Pr_E(\Delta, A)$.

If the distribution of A is known, the above objective is realized by

Ideal observer rule This rule sends $b \in \mathcal{B}$ to $\Delta_o(b) = a$, such that $p(a|b)$ is maximal, where $p(a|b)$ can be calculated (knowing A)

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

[5] The reader is encouraged to find some "meaningful" examples.

[6] We have noted that in this case the distribution of B is determined (eq. (25)), hence the definition is correct.

It follows easily from definition that

$$Pr_C(\Delta_o, A) \quad \geq \quad Pr_C(\Delta, A),$$

for any rule Δ.
If the distribution of A is unknown, a reasonable choice is

Maximal likelihood rule This rule sends $b \in \mathcal{B}$ to $\Delta_{\max}(b) = a$, such that $p(a \to b) = p(b|a)$ is maximal. If A has uniform distribution (i.e., $p(a) = \frac{1}{|\mathcal{A}|}$) then this rule acts as Δ_o, i.e.[7],

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

Indeed, maximizing $p(a|b)$ given b amounts to maximizing $p(a \wedge b) = p(a|b) \cdot p(b)$, which in the uniform case is $p(a \wedge b) = p(a \to b) \cdot \frac{1}{|\mathcal{A}|}$ (i.e., depends on $p(a \to b)$, but not on $p(a)$).

If A is not uniform, the maximal likelihood rule need not be optimal (the reader may easily find an example). However, it is in some sense *globally optimal*. We only sketch the argument informally.

Let $\mathcal{A} = \{a_1, \ldots, a_m\}$, and let $\mathcal{P}$ be the set of all possible probability distributions over $\mathcal{A}$,

$$\mathcal{P} \quad = \quad \{\mathbf{p} : \sum_{a \in \mathcal{A}} \mathbf{p}(a) = 1\}.$$

We identify a random variable A taking values in $\mathcal{A}$ with its probability distribution $\mathbf{p}$ in $\mathcal{P}$. Now, the global value of a rule Δ can be calculated by

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

But it should be clear (at least intuitively, as the formal argument should refer to the concept of Lebesgue integral) that $\int_{\mathbf{p} \in \mathcal{P}} \mathbf{p}(a)\, d\mathbf{p}$ may not depend on a. (Note that $\mathbf{p}(a)$ is just a projection of $\mathbf{p}$ on one of its components, and no component is *a priori* privileged.) Thus $\int_{\mathbf{p} \in \mathcal{P}} \mathbf{p}(\Delta(b))\, d\mathbf{p}$ is always the same. Hence, maximization of $\int_{\mathbf{p} \in \mathcal{P}} Pr_C(\Delta, \mathbf{p})\, d\mathbf{p}$ amounts to maximization of $\sum_{b \in \mathcal{B}} p(\Delta(b) \to b)$, and this is realized by the maximal likelihood rule.

11.04.2006

Multiple use of channel

Recall (page 20) that if A and B form an input-output pair for a channel Γ then $p(b|a) = P(a \to b)$. Now suppose that we subsequently send symbols $a_1, a_2, \ldots, a_k$; what is the probability that the output is $b_1, b_2, \ldots, b_k$? One may expect that this is just the product of the $P(a \to b)$'s, we shall see that this is indeed the case if the transmissions are independent.
Recall that random variables $X_1, \ldots, X_k$ are independent[8] if

$$p(X_1 = x_1 \wedge \ldots \wedge X_k = x_k) \quad = \quad p(X_1 = x_1) \cdot \ldots \cdot p(X_k = x_k)$$

Extending our notational convention (see page 13), we often abbreviate $p(X_1 = x_1 \wedge \ldots \wedge X_k = x_k)$ by $p(x_1 \ldots x_k)$, etc.

[7] We have $\Delta_{\max} = \Delta_o$, assuming that both rules make the same choice if there are more a's with the same maximal $p(a \to b)$.

[8] The reader should note that this assumption is stronger than pairwise independence; an easy example consists of $X_1, \ldots, X_k$ $(k > 2)$, with values in $\{0, 1\}$, where $X_1, \ldots, X_{k-1}$ are independent and $X_k = \bigoplus_{i=1}^{k-1} X_i$.

Lemma If the random variables (X, Y) and (X', Y') are independent then

$$p(Y = y \wedge Y' = y' | X = x \wedge X' = x') \quad = \quad p(Y = y | X = x) \cdot p(Y' = y' | X' = x'),$$

whenever $p(X = x \wedge X' = x') > 0$.

Proof Observe first that independence of (X, Y) and (X', Y') implies independence of X and X'; indeed we have

$$p(x \wedge x') = p\left((x \wedge \bigvee \mathcal{Y}) \wedge (x' \wedge \bigvee \mathcal{Y}')\right) = \sum_{y,y'} p(x \wedge y) \cdot p(x' \wedge y') = p(x) \cdot p(x').$$

Hence

$$p(y \wedge y' | x \wedge x') = \frac{p(y \wedge y' \wedge x \wedge x')}{p(x \wedge x')} = \frac{p(y \wedge x) \cdot p(y' \wedge x')}{p(x) \cdot p(x')} = p(y|x) \cdot p(y'|x').$$

Corollary Suppose that $(A_1, B_1), \ldots, (A_k, B_k)$, are independent random variables with the same distribution, such that each (A_i, B_i) forms an input-output pair for a channel Γ. Then

$$p(b_1 \ldots b_k | a_1 \ldots a_k) \quad = \quad p(b_1|a_1) \cdot \ldots \cdot p(b_k|a_k). \tag{32}$$

Proof Clearly the independence of $(A_1, B_1), \ldots, (A_k, B_k)$ implies that (A_1, B_1) is independent from the random variable $(A_2, \ldots, A_k, B_2, \ldots, B_k)$. Hence, we prove the desired equality by repeated application of the Lemma. QED

The independence assumption in the above corollary may appear unrealistic in some applications. Indeed, it can be replaced by somewhat weaker hypotheses. We discuss it briefly.
Suppose that $(A_1, B_1), \ldots, (A_k, B_k)$ satisfy the following two conditions.

Memorylessness:

$$p(b_k | a_1 \ldots a_k, b_1 \ldots b_{k-1}) \quad = \quad p(b_k | a_k) \tag{33}$$

Absence of feedback:

$$p(a_k | a_1 \ldots a_{k-1}, b_1 \ldots b_{k-1}) \quad = \quad p(a_k | a_1 \ldots a_{k-1}) \tag{34}$$

Then the equation (32) is also satisfied.
Indeed, we can use induction on k to show

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

whenever the last probability is > 0. The case of $k = 1$ is trivial. To show the induction step, we have, from (33),

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

and from (34)

$$p(a_1 \ldots a_k, b_1 \ldots b_{k-1}) \quad = \quad p(a_1 \wedge \ldots \wedge a_{k-1} \wedge b_1 \wedge \ldots \wedge b_{k-1}) \cdot \frac{p(a_1 \wedge \ldots \wedge a_k)}{p(a_1 \wedge \ldots \wedge a_{k-1})}$$

But, by the induction hypothesis,

$$\frac{p(a_1 \wedge \ldots \wedge a_{k-1} \wedge b_1 \wedge \ldots \wedge b_{k-1})}{p(a_1 \wedge \ldots \wedge a_{k-1})} \quad = \quad p(b_1|a_1) \cdot \ldots \cdot p(b_{k-1}|a_{k-1}),$$

which gives the claim.

Proviso In what follows, unless stated otherwise, we always assume that the equation (32) holds, whenever a BSC is used several times.

Improving reliability

Suppose we use a binary symmetric channel (see page 23) Γ given by the matrix $\begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, where $P > Q$. In this case $\Delta_{\max}(i) = i$, 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 $Pr_E(\Delta_{\max}, A) = Q$. As it does not depend on A, we simply write $Pr_E(\Delta_{\max}) = Q$.
Can we achieve a better result, using the same channel in a more clever way? A natural solution is to send each bit of the message more than once, say 3 times. As the correct transmission is more likely than the error (since $P > Q$), the receiver should decode the message looking at the majority:

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

Then the whole procedure behaves as a (new) channel

$$\begin{array}{ccccc} 0 & \to & \boxed{\Gamma'} & \to & 0 \\ 1 & \to & & \to & 1 \end{array}$$

What is the matrix of this channel?
Using the Corollary above, we can see that, e.g., the probability $p(0|0)$ that the output is 0 if the input has been 0, amounts to

$$p(000|000) + p(001|000) + p(010|000) + p(100|000) = P^3 + 3P^2Q.$$

Similar calculations made for the remaining $p(i|j)$, easily show that Γ' is again a binary symmetric channel, with the matrix

$$\begin{pmatrix} P^3 + 3P^2Q & Q^3 + 3Q^2P \\ Q^3 + 3Q^2P & P^3 + 3P^2Q \end{pmatrix}$$

Clearly $Q^3 + 3Q^2P < P^3 + 3P^2Q$, hence the error probability of Γ' is

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

To see that this is indeed less than Q, it is enough to examine the function $Q^3+3Q^2(1-Q)-Q$, which turns out to be negative for $Q < \frac{1}{2} + \frac{1}{\sqrt{12}}$.

More generally, if the sender sends each bit n times and the receiver decides by majority (for simplicity, suppose that n is odd), we obtain the BSC channel with the matrix

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

Now the probability of error is

$$Pr_E(\Delta_{\max}) = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i} P^i \cdot Q^{n-i} \leq \underbrace{\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i}}_{2^{n-1}} P^{\lfloor \frac{n}{2} \rfloor} \cdot Q^{\lfloor \frac{n}{2} \rfloor}$$

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 $Pr_E(\Delta_{\max}) \to 0$ if $n \to \infty$.

This means that we can make the probability of error arbitrarily small, but it comes at the expense of longer and longer messages. The celebrated Shannon's theorem (which we will learn at the next lecture) shows that, in some sense, this expense is not necessary. To get the intuition that this *may* be possible, observe that our choice of repeating the same symbol has been made for simplicity, but other choices are also possible. For example, while spelling a difficult word (e.g., by phone), one often says the names, e.g., Bravo, Alpha, November, Alpha, Charlie, Hotel (here I have used the *International Radio Operators Alphabet*).

Hamming distance

For a finite set $\mathcal{A}$ and $n \in \mathbb{N}$, the *Hamming distance* between $u, v \in \mathcal{A}^n$ is defined by

$$d(u,v) \quad = \quad |\{i : u_i \neq v_i\}| \tag{35}$$

It is easy to see that the axioms of the metric space are satisfied:

positivity $d(u,v) = 0 \iff u = v$,

symmetry $d(u,v) = d(v,u)$,

triangle inequality $d(u,w) \leq d(u,v) + d(v,w)$

(the last follows from the fact that $\{i : u_i \neq w_i\} \subseteq \{i : u_i \neq v_i\} \cup \{i : v_i \neq w_i\}$).

25.04.2006

Consider a BSC Γ given by a matrix $\begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$ with $P > Q$. The Hamming distance defined above allows for a succinct notation of the conditional probability that the sequence of outputs is $\vec{b} = b_1 \dots b_k$ if the sequence of inputs is $\vec{a} = a_1 \dots a_k$. The equation (32) gives us

$$p(b_1 \dots b_k | a_1 \dots a_k) \quad = \quad Q^{d(\vec{a},\vec{b})} \cdot P^{1-d(\vec{a},\vec{b})}. \tag{36}$$

Channel coding

For an input-output pair (A, B), we consider an auxiliary random variable

$$E \quad = \quad A \oplus B,$$

it can be viewed as the error of the transmission by channel. We have

$$p(b|a) \quad = \quad p(E = a \oplus b) \tag{37}$$

Indeed, by definition of BSC

$$p(b|a) = \begin{cases} P & a = b \quad (a \oplus b = 0) \\ Q & a \neq b \quad (a \oplus b = 1) \end{cases}$$

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

$$p(E = 1) = p(A = 0) \cdot p(0 \to 1) + p(A = 1) \cdot p(1 \to q) = Q,$$

so the both sides of (37) coincide, for all a, b.

Now consider a sequence of input-output pairs $(A_1, B_1), \ldots, (A_k, B_k)$, satisfying the equation (32). This implies that the random variables $E_1, \ldots, E_k$ (where $E_i = A_i \oplus B_i$) are *independent*[9]. Indeed, we have (in what follows, $p(\vec{E} = \vec{e})$, or simply $p(\vec{e})$, abbreviate $p(E_1 = e_1 \wedge \ldots \wedge E_k = e_k)$, etc.)

$$p(e_1 \ldots e_k) = \sum_{\vec{a}} p(\vec{A} = \vec{a} \wedge \vec{B} = \vec{a} \oplus \vec{e}) = \sum_{\vec{a}} p(\vec{A} = \vec{a}) \cdot p(\vec{B} = \vec{a} \oplus \vec{e} | \vec{A} = \vec{a}),$$

where $\vec{a}$ ranges over those vectors for which $p(\vec{a}) > 0$. But, using (32) and (37), we have

$$\begin{aligned} p(\vec{B} = \vec{a} \oplus \vec{e} | \vec{A} = \vec{a}) \quad &= \quad p(B_1 = a_1 \oplus e_1 | A_1 = a_1) \cdot \ldots \cdot p(B_k = a_k \oplus e_k | A_k = a_k) \quad (38) \\ &= \quad p(E_1 = e_1) \cdot \ldots \cdot p(E_k = e_k) \quad (39) \end{aligned}$$

for any $\vec{a}$, hence

$$p(e_1 \ldots e_k) \quad = \quad p(e_1) \cdot \ldots \cdot p(e_k)$$

as desired.

Suppose we dispose of a binary symmetric channel Γ as above ($P > Q$), and wish to send a value of a random variable X with values in $\mathcal{X} = \{x_1, \ldots, x_m\}$. In the early lectures we have studied how to efficiently encode the values of X. If the channel is faithful, all we need is to find an optimal encoding $\varphi : \mathcal{X} \to \{0, 1\}^*$ and then send the message bit by bit. The average length (time) of transmission will be bounded by $H(X) + 1$ (c.f. the Shannon-Fano coding, page 12). On the other hand, we can always encode $\mathcal{X}$ using strings of length $\lceil \log m \rceil$, which gives the bound for the worst-case time of the transmission. (The two bounds may be not achievable by the same encoding.)

However, if the channel is insecure, this method would lead to errors. As the example on the page 28 suggests, we should rather use redundant, and hence non-optimal encoding. In what follows, we will struggle for a method which should re-conciliate two antagonistic objectives:

[9]The converse is not true in general, but the independence of $E_1, \ldots, E_k$, *and* independence of $(E_1, \ldots, E_k)$ from $(A_1, \ldots, A_k)$ implies the equation (32).

- keep redundancy as small as possible,
- keep the error probability as small as possible.

We first describe a general scheme of the method.

Transmission algorithm Suppose we are given a random variable X with values in $\mathcal{X} = \{x_1, \ldots, x_m\}$.

1. Choose $n \in \mathbb{N}$, and $C \subseteq \{0,1\}^n$ with $|C| = m$.
2. Choose $\varphi : \mathcal{X} \stackrel{1:1}{\to} C$. Clearly φ is an instantaneous code.

   We can identify $\mathcal{X}$ and C (via φ). That is, since now on, we assume that X is a random variable with values in C.
3. Send the string $X = a_1 \ldots a_n$ by the channel Γ, bit by bit. Let the output received from the channel be $Y = b_1 \ldots b_n$. Assuming that the use of the channel is memoryless and feedback–free, we have (eq. (36))

$$p(b_1 \ldots b_k | a_1 \ldots a_k) \quad = \quad Q^{d(\vec{a},\vec{b})} \cdot P^{1-d(\vec{a},\vec{b})}.$$

4. To decode, given $Y = b_1 \ldots b_n$, choose $a_1 \ldots a_n \in C$ which maximizes $p(b_1 \ldots b_n | a_1 \ldots a_n)$ (like in the maximal likelihood rule).

   In other words, let $\Delta(b_1 \ldots b_n)$ be a code-word in C nearest to $b_1 \ldots b_n$. (We fix some policy of choice if there is more than one word with this property.)

   This Δ is called the *nearest neighbour rule*.

The method described above can be viewed as a new channel (from C to C)

$$C \ni a_1 \ldots a_n \to \boxed{\Gamma} \to b_1 \ldots b_n \to \Delta(b_1 \ldots b_n) \in C$$

with the probability of error

$$Pr_E(\Delta, X) \quad = p(\Delta \circ Y \neq X).$$

Our first observation is that the worst case is if (the distribution of) X is *uniform*, i.e., $p(x) = \frac{1}{m}$, for $x \in C$.

Fact Let X, U, be two random variables with values in $C \subseteq \{0,1\}^n$, where U is uniform and X arbitrary. Then there is a permutation $\varphi : C \stackrel{1:1}{\to} C$ such that

$$Pr_E(\Delta, \varphi \circ X) \quad \leq \quad Pr_E(\Delta, U).$$

Proof See the 2004 note `http://zls.mimuw.edu.pl/~niwinski/Info/` (in Polish).

Hence, in order to estimate the efficiency of our method in terms of $Pr_E(\Delta, X)$, we may assume without loss of generality that X is uniform.

9.05.2006

In view of the fact just proved, in order to estimate the error probability for arbitrary variable, it is enough to consider X with uniform distribution. In this case $Pr_E(\Delta, X)$ depends only on C, hence we will denote it just by $Pr_E(\Delta, C)$.

The redundancy is measured as the ratio between the binary entropy of C, which is $\log_2 |C|$ (c.f. Fact on page 8) and the actual length of the code, i.e., n.

Definition The *transmission rate* of a code $C \subseteq \{0,1\}^n$ is

$$R(C) \quad = \quad \frac{\log_2 |C|}{n}.$$

The intuition behind it is that, in order to transmit $\log_2 |C|$ bits of information, we need in reality to send n bits, hence the rate is $\frac{\log_2 |C|}{n}$ bits per transmission.

Note that the two objectives stated on the page 30 mean that we want both $Pr_E(\Delta, C)$ and $R(C)$ to be as small as possible.

Examples We start with a noisy typewriter described on the page 22. Although it does not exactly fit to the setting above, the basic concepts are well illustrated.

Clearly this channel can produce many errors. However, if we have used only each second letter, say $a, c, e, g, \ldots, u, w, y$, then the received message can be always decoded in the correct way.

Can we use this observation to transmit faithfully arbitrary texts?

A simple idea is to encode the letters by pairs, still using only a half of the alphabet, e.g.

a	aa
b	ac
c	cc
d	ce
...	

For example, `hhqtfeabtvjjceefbb` should be read as `greatidea`.

Here, in order to transmit one letter, we need to send two, hence the transmission rate is $\frac{1}{2}$. Can we do better?

If we have an auxiliary symbol, # say, which can be decoded without error, we can encode (somewhat like in the musical notation)

a	a
b	# a
c	c
d	# c
...	

Here the average length of the encoding is $\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2 = \frac{3}{2}$, so we can estimate the rate by $\frac{2}{3}$. We can apply this idea without extending the alphabet, by choosing one letter, y say, to play the role of #, and additionally encode y by yy (which decreases the rate slightly). A more interesting problem of that kind is considered in Exercise 1 of Series 1 A.

In the second example we assume a BSC Γ, and explore the improvement we have made on the page 28 to send just two bits 0 and 1. Assume $n = k \cdot \ell$, and let $m = 2^k$. Then every string of bits $a_1 \ldots a_k$ can be encoded by $a_1^\ell \ldots a_k^\ell \in \{0,1\}^n$, which defines a code of rate $\frac{1}{\ell}$. Refining the analysis of page 29 we can see that, for arbitrary k, we can make $Pr_E(\Delta, C)$ arbitrary small

if ℓ is sufficiently large. Roughly speaking, as soon as a BSC is not completely chaotic (i.e., $P \neq Q$), we can transmit arbitrary text with arbitrary small error, but the price to pay is the slowing down the transmission rate almost to 0.

The celebrated Shannon Theorem shows that the situation is much better than that: we can achieve the same thing with the rate close to a *positive* constant, namely the channel capacity C_Γ.

Before stating and proving the theorem, we show a kind of converse (lower bound) result, stating that if no error is permitted, the transmission rate cannot exceed C_Γ. In this fact, Γ can be an arbitrary binary channel (not necessarily[10] BSC), but as usual, we assume equation (32) (c.f. Proviso on page 28).

Fact If $Pr_E(\Delta, C) = 0$ then

$$R(C) \quad \leq \quad C_\Gamma.$$

Proof Let $X = (A_1, \ldots, A_n)$ and $Y = (B_1, \ldots, B_n)$ be as in the transmission algorithm. Using the equation (32) we verify by an easy calculation that

$$H(Y|X) \quad = \quad H(B_1|A_1) + \ldots + H(B_n|A_n). \tag{40}$$

We also have (12)

$$H(Y) \quad \leq \quad H(B_1) + \ldots + H(B_n)$$

Hence

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &\leq \sum_{i=1}^{n} H(B_i) - \sum_{i=1}^{n} H(B_i|A_i) \\ &= \sum_{i=1}^{n} \underbrace{(H(B_i) - H(B_i|A_i))}_{I(A_i,B_i)} \\ &\leq n \cdot C_\Gamma \end{aligned}$$

(by definition of C_Γ).
On the other hand, we have

$$\begin{aligned} I(X,Y) &= H(X) - \underbrace{H(X|Y)}_{0} \\ &= \log_2 m \end{aligned}$$

where $m = |C|$. Here $H(X|Y)$ vanishes since the assumption $Pr_E(\Delta, C) = 0$ implies that X is a function of Y, namely $X = \Delta(Y)$ (c.f. (11)). Next, $H(X) = \log_2 m$, since X is uniform (as assumed in the definition of $Pr_E(\Delta, C) = 0$).

Hence we have

$$R(C) = \frac{\log_2 m}{n} \leq C_\Gamma$$

as required. QED

[10] Actually, the reader may notice that for a BSC, the statement is uninteresting, as the zero–error assumption holds only for a faithful channel. However, the extension of the proof to a sharper result mentioned in *Remark* below makes sense for BSC as well.

Remark Using the continuity argument we could easily show that if we weaken the assumption $Pr_E(\Delta, C) = 0$ to $Pr_E(\Delta, C) \leq \delta$, for some $\delta > 0$, we have in the above proof $H(X|Y) \leq \vartheta(\delta)$ (for some continuous bounded function ϑ), and hence

$$\log_2 m - \vartheta(\delta) \quad \leq \quad n \cdot C_\Gamma$$

implying

$$R(C) \quad \leq \quad C_\Gamma + \frac{\vartheta(\delta)}{n}.$$

Roughly speaking, if we want to make the error probability small, we need to keep the transmission rate close to C_Γ.

We are ready to state the basic result of information theory, due to Claude Shannon (1948). Intuitively it says that message transmission through a noisy channel with arbitrarily small error probability is possible, with the transmission rate arbitrarily close to the channel capacity, provided that the length of code is sufficiently large. We prove the theorem only for BSC, for the general setting see, e.g., [2] (Theorem 8.7.1). The proof below follows [1].

Channel Coding Theorem (Shannon) Let Γ be a binary symmetric channel (BSC) with a matrix $\begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$, where $P > Q$. Then $\forall \varepsilon, \delta > 0 \ \exists n_0 \quad \forall n \geq n_0 \quad \exists C \subseteq \{0,1\}^n$

$$C_\Gamma - \varepsilon \leq \quad R(C) \quad \leq C_\Gamma \tag{41}$$

$$Pr_E(\Delta, C) \quad \leq \delta \tag{42}$$

Proof We first informally describe the basic idea. Suppose an input $X = a_1 \dots a_n$ is turned into the output $Y = b_1 \dots b_n$. What is the *expected* distance between X and Y? As this distance amounts to the number of transmission errors, and the probability of one error is Q, the Law of Large Numbers tells us that $d(X, Y)$ approaches to $Q \cdot n$ if $n \to \infty$. Now, if the decoding fails, i.e., $\Delta(Y) \neq X$, it is useful to distinguish between two possible "reasons" for that:

- Y is "far" from X,
- Y is not that far, but a confusion arises, because some $X' \neq X$ is at least as good as X,

where "far" means: exceeding the expected value $Q \cdot n$.

The first type of failure is caused by the channel, but it is corrected by Nature: the Law of Large Numbers guarantees that a big distance between X and Y happens rarely if n is large. The second issue is, to some extent, responsibility of the code designer. Indeed, to prevent confusion, the code-words should be "reasonably far" one from another. Taking the expected distance as the measure of "far", this means that the balls of radius $Q \cdot n$ (in the Hamming metrics) centered in any two code-words should be disjoint. So the question is: how many disjoint balls of radius $Q \cdot n$ can one "pack" in $\{0,1\}^n$? The size of one such ball, as we will see later, can be estimated by $\approx 2^{n \cdot H(Q)}$. This suggests the number of balls of

$$m \approx 2^n : 2^{n \cdot H(Q)} = 2^{n(1-H(Q))} = 2^{n \cdot C_\Gamma},$$

and hence the transmission rate $R(C) \approx C_\Gamma$. The amazing discovery of Shannon is that this bound is really achievable. However, the proof is non-constructive, i.e., it only shows the existence of the desired code.

In what follows, we use the lower-case letters $u, v, w, x, y, \ldots$ for the "concrete" vectors in $\{0,1\}^n$, to distinguish them from random variables. As usual, $\oplus$ is understood component-wise. We choose $\eta > 0$, whose dependence on ε and δ will be made precise later (intuitively: very small). Let

$$\rho \quad = \quad n(Q + \eta).$$

Now suppose $C \subseteq \{0,1\}^n$ is a code with $|C| = m$. By definition of Δ, if, for some $u \in C$, $e \in \{0,1\}^n$, $d(u, u \oplus e) \leq \rho$ and $\forall v \in C - \{u\}$, $d(v, u \oplus e) > \rho$, then $\Delta(u \oplus e) = u$. Therefore, if $\Delta(u \oplus e) \neq u$ then either $d(u, u \oplus e) > \rho$, or, for some $v \in C - \{u\}$, $d(v, u \oplus e) \leq \rho$. Now we can view the vector e as the value of a random variable $E = (E_1, \ldots, E_n)$, where $E_i = A_i \oplus B_i$ are as in the page 30. Recall that $E_1, \ldots, E_n$ are independent and have the identical distribution $p(E_i = 0) = P$, $p(E_i = 1) = Q$.

Then the above observation induces the following inequality, for a fixed $u \in C$.

$$p(\Delta(u \oplus E) \neq u) \quad \leq \quad p(d(u, u \oplus E) > \rho) + \sum_{v \in C - \{u\}} p(d(v, u \oplus E) \leq \rho). \tag{43}$$

16-05-2006

To estimate the first summands of (43) we use the following.

Weak Law of Large Numbers Let $X_1, X_2, \ldots$, be a sequence of random variables, such that any $X_1, X_2, \ldots, X_n$ are independent, and each X_i takes a finite number of real values with the same distribution. Let $\mu = E(X_i)$. Then, for any $\alpha > 0$,

$$\lim_{n \to \infty} p(|\frac{1}{n} \sum_{i=1}^{n} X_i - \mu| > \alpha) \quad = \quad 0. \tag{44}$$

We apply it to the sequence $E_1, E_2, \ldots$. Clearly $E(E_i) = 0 \cdot P + 1 \cdot Q = Q$. Hence $p(|\frac{1}{n} \cdot \sum_{i=1}^{n} E_i - Q| > \eta) \to 0$ if $n \to \infty$. Therefore

$$p(d(u, u \oplus E) > \rho) \leq p(\frac{1}{n} \cdot \sum_{i=1}^{n} E_i > Q + \eta) \leq p(|\frac{1}{n} \cdot \sum_{i=1}^{n} E_i - Q| > \eta) \leq \frac{\delta}{2}, \tag{45}$$

for n sufficiently large.

Now recall that we wish to estimate $Pr_E(\Delta, C)$, which amounts to (c.f. (31))

$$Pr_E(\Delta, C) \quad = \quad \sum_{u \in C} p(X = u) \cdot p(\Delta \circ Y \neq u | X = u)$$

We have by definition $Y = X \oplus E$, and from (39),

$$p(Y = w | X = u) \quad = \quad p(E = w \oplus u). \tag{46}$$

Hence

$$\begin{aligned} p(\Delta \circ Y \neq u | X = u) \quad &= \quad \sum_{v : \Delta(v) \neq u} p(Y = v | X = u) \\ &= \quad \sum_{e : \Delta(u \oplus e) \neq u} p(Y = u \oplus e) | X = u) \\ &= \quad \sum_{e : \Delta(u \oplus e) \neq u} p(E = e) \\ &= \quad p(\Delta(u \oplus E) \neq u). \end{aligned}$$

Moreover $p(X = u) = \frac{1}{m}$ since, by assumption, X is uniform.
Together with (45), this gives us

$$\begin{aligned} Pr_E(\Delta, C) &\leq \frac{1}{m} \sum_{u \in C} \left(p(d(u, u \oplus E) > \rho) + \sum_{v \in C - \{u\}} p(d(v, u \oplus E) \leq \rho) \right) \\ &\leq \frac{\delta}{2} + \frac{1}{m} \sum_{u \in C} \sum_{v \in C - \{u\}} p(d(v, u \oplus E) \leq \rho), \end{aligned} \tag{47}$$

if n is sufficiently large.
Before proceeding further, we will estimate the size of a ball of radius $\lambda \cdot n$, where $\lambda \leq \frac{1}{2}$. More specifically, we show that

$$\sum_{i \leq \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \leq 2^{n \cdot H(\lambda)}, \tag{48}$$

where H is the function defined by (28).
Let $\kappa = 1 - \lambda$. Note first that

$$\begin{aligned} \log_2 \lambda^{\lambda n} \cdot \kappa^{\kappa n} &= n \cdot (\lambda \cdot \log_2 \lambda + \kappa \cdot \log_2 \kappa) \\ &= -n \cdot H(\lambda) \end{aligned}$$

Now it is enough to show that, for all $i \leq \lambda n$,

$$\lambda^i \kappa^{n-i} \geq \lambda^{\lambda n} \cdot \kappa^{\kappa n}. \tag{49}$$

Indeed, this will give us

$$1 \geq \sum_{i \leq \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \lambda^i \kappa^{n-i} \geq \sum_{i \leq \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \lambda^{\lambda n} \cdot \kappa^{\kappa n}$$

and consequently

$$\sum_{i \leq \lambda \cdot n} \left(\begin{array}{c} n \\ i \end{array} \right) \leq \frac{1}{\lambda^{\lambda n} \cdot \kappa^{\kappa n}} = 2^{n \cdot H(\lambda)},$$

as required.
If λn is integer, the inequality (49) is obvious (just replace smaller by bigger). Otherwise, we have $\lambda n = \lfloor \lambda n \rfloor + \Delta\lambda$, $\kappa n = \lfloor \kappa n \rfloor + \Delta\kappa$, $\lfloor \lambda n \rfloor + \lfloor \kappa n \rfloor = n - 1$, and $\Delta\lambda + \Delta\kappa = 1$. Since $\kappa \geq \lambda$, we have, for $i \leq \lambda n$,

$$\lambda^i \kappa^{n-i} \geq \lambda^{\lfloor \lambda n \rfloor} \cdot \kappa^{\lfloor \kappa n \rfloor + 1} = \lambda^{\lfloor \lambda n \rfloor} \cdot \kappa^{\lfloor \kappa n \rfloor} \underbrace{\kappa^{\Delta\lambda + \Delta\kappa}}_{\geq \lambda^{\Delta\lambda} \cdot \kappa^{\Delta\kappa}} \geq \lambda^{\lambda n} \cdot \kappa^{\kappa n}.$$

This completes the proof of (48).
We come back to the estimation of $Pr_E(\Delta, C)$. Recall that (47) holds for any code C, if only n is sufficiently large. We will now show that, for sufficiently large n, there *exists* a code C satisfying the conditions (41) and (42) of Shannon's theorem; in particular the second term of (47) should be estimated by $\frac{\delta}{2}$.

To this end, rather than searching for a specific code C with the desired property, we will use the *probabilistic method*.

Fix $m < 2^n$. Let $\mathcal{C}$ be the set of all sequences $c_1, \ldots, c_m \in \{0,1\}^n$, with $c_i \neq c_j$, for $i \neq j$. Let $N = |\mathcal{C}|$. Clearly

$$N \;=\; \left(\begin{array}{c} 2^n \\ m \end{array} \right) \cdot m!$$

In what follows, we use symbol $\bar{C}$ for a sequence in $\mathcal{C}$, and let the notation $Pr_E(\Delta, \bar{C})$ stand for $Pr_E(\Delta, C)$, where C is the set of values of $\bar{C}$.

Now the probabilistic argument, due to Claude Shannon, is based on the following simple observation. If

$$\frac{1}{N} \sum_{\bar{C}} Pr_E(\Delta, \bar{C}) \;\leq\; \delta$$

then there exists a code C, such that $Pr_E(\Delta, C) \leq \delta$.

Note that if $\bar{C}$ is a sequence in $\mathcal{C}$ with the set of values $C = \{c_1, \ldots, c_m\}$ then

$$\sum_{u \in C} \sum_{v \in C - \{u\}} p(d(v, u \oplus E) \leq \rho) \;=\; \sum_{i=1}^{m} \sum_{j \neq i} p(d(c_j, c_i \oplus E) \leq \rho).$$

Hence, (47) gives us

$$\begin{aligned} \frac{1}{N} \sum_{\bar{C}} Pr_E(\Delta, \bar{C}) \;&\leq\; \frac{1}{N} \sum_{\bar{C}} \left(\frac{\delta}{2} + \frac{1}{m} \sum_{i=1}^{m} \sum_{j \neq i} p(d(c_j, c_i \oplus E) \leq \rho) \right) \\ &=\; \frac{\delta}{2} + \frac{1}{m} \sum_{i=1}^{m} \sum_{j \neq i} \underbrace{\frac{1}{N} \sum_{\bar{C}} p(d(c_j, c_i \oplus E) \leq \rho)}_{(*)} \end{aligned} \tag{50}$$

We will now estimate (*), for a *fixed* pair of indices $i \neq j$.

For $e \in \{0,1\}^n$, let $S_\rho(e)$ be the ball in $\{0,1\}^n$ of radius ρ centered in e, i.e.,

$$S_\rho(e) \;=\; \{v \in \{0,1\}^n : d(v, e) \leq \rho\}.$$

It is easy to see that

$$d(v, u \oplus e) \leq \rho \;\iff\; v \oplus u \in S_\rho(e).$$

Hence

$$\begin{aligned} \frac{1}{N} \sum_{\bar{C}} p(d(c_j, c_i \oplus E) \leq \rho) \;&=\; \frac{1}{N} \sum_{\bar{C}} p\left(c_i \oplus c_j \in S_\rho(E)\right) \\ &=\; \sum_{e \in \{0,1\}^n} p(E = e) \cdot \underbrace{\frac{1}{N} \sum_{\bar{C}} \cdot \chi(c_i \oplus c_j \in S_\rho(e))}_{(**)} \end{aligned} \tag{51}$$

where χ is the truth function, i.e.,

$$\chi(\varphi) = \left\{ \begin{array}{ll} 1 & \text{if } \varphi \text{ holds} \\ 0 & \text{otherwise} \end{array} \right.$$

We now estimate the value of (**), for a fixed e. Obviously any vector different from 0^n occurs as a value of $c_i \oplus c_j$, for some sequence $\bar{C} \in \mathcal{C}$, and it is easy to see that each such vector occurs in this role the same number of times, i.e.,

$$|\{\bar{C} : u = c_i \oplus c_j\}| = |\{\bar{C} : v = c_i \oplus c_j\}| = \frac{N}{2^n - 1}$$

for any $u, v \in \{0,1\}^n - \{0^n\}$. Hence each $u \in S_\rho(e) - \{0^n\}$ contributes the value $\frac{N}{2^n-1}$ to the sum $\sum_{\bar{C}} \cdot \chi(c_i \oplus c_j \in S_\rho(e))$, i.e.,

$$\sum_{\bar{C}} \cdot \chi(c_i \oplus c_j \in S_\rho(e)) \;=\; \frac{N}{2^n - 1}|S_\rho(e) - \{0^n\}|$$

Therefore we further have

$$\begin{aligned}\sum_{e\in\{0,1\}^n} p(E=e) \cdot \frac{1}{N} \sum_{\bar{C}} \cdot \chi(c_i \oplus c_j \in S_\rho(e)) &= \sum_{e\in\{0,1\}^n} p(E=e) \cdot \frac{1}{2^n-1}|S_\rho(e) - \{0^n\}| \\ &= \frac{1}{2^n-1}|S_\rho(e) - \{0^n\}|\end{aligned}$$

(since $\sum_e p(E=e) = 1$). Now, from (48), we have

$$|S_\rho(e) - \{0^n\}| \;\leq\; 2^{n\cdot H(Q+\eta)}$$

(recall that $\rho = Q + \eta$). Coming back to (50), we have

$$\begin{aligned}\frac{1}{N}\sum_{\bar{C}} Pr_E(\Delta, \bar{C}) &\leq \frac{\delta}{2} + \frac{1}{m}\sum_{i=1}^{m}\sum_{j\neq i} \frac{1}{2^n-1} \cdot 2^{n\cdot H(Q+\eta)} \\ &= \frac{\delta}{2} + \frac{1}{m} \cdot m \cdot \underbrace{(m-1)\cdot\frac{1}{2^n-1}}_{\leq \frac{m}{2^n}} \cdot 2^{n\cdot H(Q+\eta)} \\ &\leq \frac{\delta}{2} + \frac{m}{2^n} \cdot 2^{n\cdot H(Q+\eta)} \\ &= \frac{\delta}{2} + 2^{n\cdot\left(\frac{\log_2 m}{n} + H(Q+\eta) - 1\right)} \qquad (52)\end{aligned}$$

Intuitively, we are very close to the goal, as the term $\left(\frac{\log_2 m}{n} + H(Q+\eta) - 1\right)$ is "almost" $R(C) - C_\Gamma$, which[11] we want to estimate in (41).

More specifically, from all previous considerations, we know that (52) holds for all sufficiently large n, say $n \geq n_1$, and all $2 \leq m < 2^n$, $0 < \eta < \frac{1}{2} - Q$. We claim, that we can further choose $n_0 \geq n_1$, m, and η, in such a way that, for all $n \geq n_0$, the following holds. (Recall that ε and δ are given in the theorem.)

$$C_\Gamma - \varepsilon \;\leq\; \frac{\log_2 m}{n} \;\leq C_\Gamma \qquad (53)$$

$$\frac{\log_2 m}{n} + H(Q+\eta) - 1 \;\leq\; -\frac{\varepsilon}{3}. \qquad (54)$$

[11] We have calculated in (29), $C_\Gamma = 1 - H(P)$, but $H(P) = H(Q)$, by symmetry of H.

Note first that (54) implies

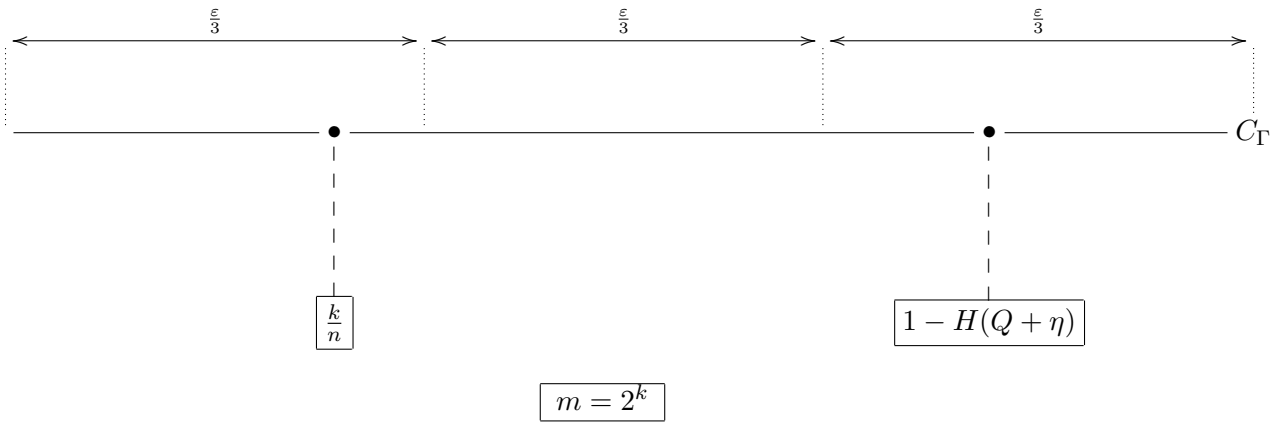
$$2^{n\cdot\left(\frac{\log_2 m}{n}+H(Q+\eta)-1\right)} \;\leq\; \frac{1}{2^{n\cdot\frac{\varepsilon}{3}}},$$

hence, if n is sufficiently large, we finally have

$$\frac{1}{N}\sum_{\bar{C}} Pr_E(\Delta,\bar{C}) \;\leq\; \frac{\delta}{2}+\frac{\delta}{2}=\delta.$$

By the probabilistic argument, there exists a code C of size m, satisfying $Pr_E(\Delta,C)\leq\delta$, as required by (42). Since $R(C)=\frac{\log_2 m}{n}$, (53) gives us the condition (41) as well.

It remains to show that the choices satisfying (53) and (54) indeed can be made. This is best illustrated by the picture:



First, using the continuity of function H, we choose η such that $C_\Gamma-\frac{1}{3}\cdot\varepsilon \leq 1-H(Q+\eta) \leq C_\Gamma$. Next, if n is sufficiently large, we can find k, such that $C_\Gamma-\varepsilon \leq \frac{k}{n} \leq C_\Gamma-\frac{2}{3}\cdot\varepsilon$. Then (53) and (54) are fulfilled by $m=2^k$. QED

Next lecture:

- Error correcting codes.
- Algorithmic (Kolmogorov) complexity.

Bibliography

[1] Gareth A. Jones and J. Mary Jones, *Information and Coding Theory*, Springer 2000.

[2] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

Damian Niwiński, Warsaw University. Last modified: 17.05.2006.