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

1.10.2009.

A precious bit of information

Consider the following puzzle[1]. 100 prisoners wear hats: red or blue. The process of getting these hats was completely random. No one knows his own hat, although everybody sees the hats of others. In order to save their life, *all* prisoners should correctly guess the color of their hats. No communication is possible, and all answers should be given *at once*.

Sure winning is clearly impossible but, perhaps surprisingly, the prisoners have $\frac{1}{2}$ chance to survive. Note that one bit of information is enough to save their life: for example, the *parity* of the number of blue hats. By assumption, this can be 0 or 1 with the same probability. If the prisoners adopt a common strategy *let's assume it is even* (say) then they have $\frac{1}{2}$ chance to win.

Exercise Show that this is also a *lower bound*, i.e., no strategy can give a guarantee better than $\frac{1}{2}$.

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

Fact 1. *If $|S| = m > 0$ and $|\Sigma| = r \geq 2$ then, for some $s \in S$,*

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

[1] It is inspired by *Mathematical Puzzles* by Peter Winkler [4], although the author of these notes has not checked if this problem is in the book.

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. □

Corollary 1. *If $\alpha : \mathbb{N} \to \Sigma^*$ is notation for natural numbers then, for infinitely many n's, $\alpha(n) \geq \lfloor \log_r n \rfloor$.*

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 $|\alpha(i_1)| \geq \lfloor \log_r m' \rfloor \geq \lfloor \log_r i_1 \rfloor$, and, by assumption, $i_1 > i_0$. And so on. □

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

Proposition 1 (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 $|\alpha(n)| \geq \log_3 n$, for infinitely many n's. □

Codes

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

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

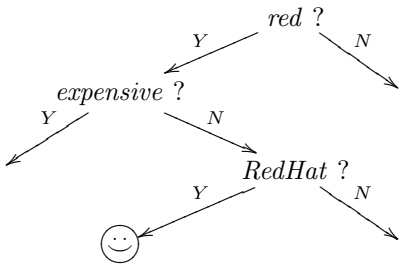
Definition 2. *A notation $\varphi : S \to \Sigma^*$ for a finite non-empty set S is a* code *if $\hat{\varphi}$ is 1:1. A code is* instantaneous *(prefix–free) if moreover $\neg\hat{\varphi}(s) \leq \hat{\varphi}(s')$, for $s \neq s'$.*

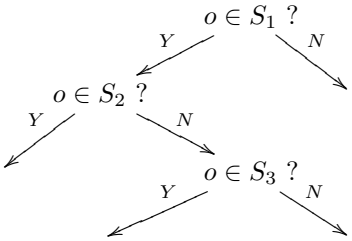
Note that the property of being an (instantaneous) code depends only on the set $\hat{\varphi}(S)$. Notice that $\varepsilon \notin \hat{\varphi}(S)$ (why ?). Any prefix-free set is a code, the set $\{aa, baa, ba\}$ is example of a non-instantaneous code, while $\{a, ab, ba\}$ is not a code at all.

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 $S' \subseteq S$.

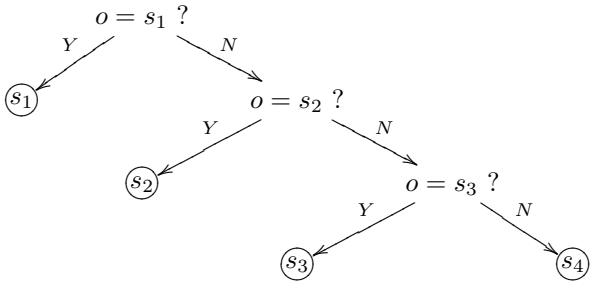




Clearly, $\lceil \log_2 |S| \rceil$ questions suffice to identify any object in S. Can we do better?

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 1 (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 $i < m$. □

8.10.2009.

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

Theorem 2 (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 $K \leq 1$. □

Properties of convex functions

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

Definition 3. *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 1. *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

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

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

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 $EX = p_1 x_1 + \ldots + p_m x_m$. 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 3 (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

$$p_1 f(x_1) + p_2 f(x_2) \quad \geq (>) \quad f(p_1 x_1 + p_2 x_2)$$

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. □

Convention We let

$$0 \log_r 0 = 0 \log_r \frac{1}{0} = 0. \tag{4}$$

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,

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

Lemma 2 (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

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

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 6), 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 $y_{q+1} \neq x_{q+1}$. □

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 $\log_2 \frac{1}{p(s_i)}$. It is possibly since probabilities in that game are powers of $\frac{1}{2}$.

So, the expected number of questions is $\sum_{i=1}^{m} p(s_i) \cdot \log_2 \frac{1}{p(s_i)}$. 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

$$\sum_{i=1}^{m} p(s_i) \cdot \ell(s_i) \quad \geq \quad \sum_{i=1}^{m} p(s_i) \cdot \log_2 \frac{1}{p(s_i)}. \tag{5}$$

Clearly, a similar inequality holds whenever the probabilities are powers of $\frac{1}{2}$. Note that in this case we can precisely "translate" probabilities on the number of questions needed to guess an object. Namely, if $p(s) > p(s')$ then in order to guess s' we need $\log_2 \frac{p(s)}{p(s')}$ questions more than to guess s.

The right-hand side of the inequality (5) makes sense also if the probabilities are not powers of $\frac{1}{2}$. We thus arrive to the central concept of Information Theory.

Definition 4 (Shannon entropy)**.** *The* entropy *of a (finite) probabilistic space S (with parameter $r > 1$) is*

$$H_r(S) \quad = \quad \sum_{s \in S} p(s) \cdot \log_r \frac{1}{p(s)} \tag{6}$$

$$\quad = \quad -\sum_{s \in S} p(s) \cdot \log_r p(s). \tag{7}$$

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 The use of the function log in the definition of entropy can be seen in a more general context of 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. If we view probability as the measure of frequency, and hence its inverse $\frac{1}{p(s)}$ as the

measure of seldomness – or maybe *strangeness* – then the function $\log \frac{1}{p(s)}$ occurring in the equation (6) can be read as our "perception of strangeness".

What values entropy can take, depending on the function 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 2.

$$H_r(S) \quad \leq \quad \log_r |S| \tag{8}$$

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. □

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 5 (Minimal code length)**.** *For a code* φ*, let*

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

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 4. *For any finite probabilistic space* S

$$H_r(S) \quad \leq \quad L_r(S) \tag{9}$$

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

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

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. □

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 $L_2(S) = 1$. However, $H_2(S) < 1$, since the probabilities are not the powers of $\frac{1}{2}$.

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:

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

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 encoding. 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 $n \to \infty$.

15. 10. 2009

Shannon's coding theorem

We first compute 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 $X : S \to \mathbb{R}$ (over a finite probabilistic space S) can be presented in two ways, readily equivalent to each other:

$$EX \quad = \quad \sum_{s \in S} p(s) \cdot X(s) \tag{10}$$

$$\quad = \quad \sum_{t \in \mathbb{R}} t \cdot p(X = t). \tag{11}$$

In the last equation we assume that the sum of arbitrarily many 0's is 0, and

$$p(X = t) \quad = \quad \sum_{s : X(s) = t} p(s). \tag{12}$$

The last notation 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 $\psi(X(s))$ holds.

We recall a basic fact from Probability Theory, which follows immediately from the first presentation of the expected value (10).

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

$$E(\alpha X + \beta Y) \quad = \quad \alpha EX + \beta EY. \tag{13}$$

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 Definition 4, this is precisely the random variable whose expected value amounts to the entropy of the space $S \times Q$, i.e.,

$$H_r(S \times Q) = E(\hat{X} + \hat{Y}).$$

Hence, the equation above gives us

$$H_r(S \times Q) \quad = \quad H_r S + H_r Q. \tag{14}$$

Consequently,

$$H_r S^n \quad = \quad n \cdot H_r S. \tag{15}$$

In order to estimate $\frac{L_r(S^n)}{n} - H_r(S)$, we first complete the inequality of Theorem 4 by the upper bound.

Theorem 5 (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 $H_r(S) = 0$).

[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 $\psi(a_1,\ldots,a_k)$.

Proof. For $|S| = 1$, we have trivially $H_r(S) = 0$ and $L_r(S) = 1$. Assume $|S| \geq 2$. We only construct an appropriate length function ℓ; the existence of a desired code will follow from Kraft's inequality (Theorem 1). 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 ℓ is defined on the whole S, and the above coincides with the Kraft inequality. 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

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

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 $0 < p(s') < 1$. □

We are ready to state Shannon's coding theorem, sometimes also called Shannon's First Theorem.

Theorem 6 (Shannon's coding 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. □

Conditional entropy and mutual information

Entropy of random variable We often consider a random variable (over a finite domain) that takes values in some abstract set $\mathcal{X}$, e.g., a set of words, rather than in real numbers.

We define the *entropy of a random variable* $X : S \to \mathcal{X}$, by

$$H_r(X) \quad = \quad \sum_{t\in\mathcal{X}} p(X = t) \cdot \log_r \frac{1}{p(X = t)} \tag{16}$$

Note that $H_r(X)$ amounts to the expected value

$$H_r(X) \quad = \quad E\left(\log_r \frac{1}{p(X)}\right), \tag{17}$$

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

$$\begin{aligned}\sum_{t\in\mathcal{X}} p(X = t) \cdot \log_r \frac{1}{p(X = t)} &= \sum_{t\in\mathcal{X}} \sum_{X(s)=t} p(s) \cdot \log_r \frac{1}{p(X = t)} \\ &= \sum_{s\in S} p(s) \cdot \log_r \frac{1}{p(X = X(s))},\end{aligned}$$

which yields (17) by the equation (10).

22. 10. 2009

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 $b \in \mathcal{B}$ with $p(b) > 0$, let

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

If $p(b) = 0$, we let, by convention, $H_r(A|b) = 0$. 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 $p(a|b) = p(a)$, and hence $H_r(A|B) = H_r(A)$. On the other hand, $H_r(A|A) = 0$; more generally, if $\varphi : \mathcal{A} \to \mathcal{B}$ is any function then

$$H_r(\varphi(A)|A) \quad = \quad 0. \tag{18}$$

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 (17) $H_r(X) = E\left(\log_r \frac{1}{p(X)}\right)$, we have, by linearity of expectation (13),

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

In general case we have the following.

Theorem 7.

$$H_r(A, B) \quad \leq \quad H_r(A) + H_r(B). \tag{19}$$

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 Golden Lemma. We use the obvious equalities $p(a) = \sum_{b \in \mathcal{B}} p(a \wedge b)$, and $p(b) = \sum_{a \in \mathcal{A}} p(a \wedge b)$.

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

Note that the last 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

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

Clearly, equation (20) will not change if we restrict the sum to $(\mathcal{A} \times \mathcal{B})^+$, i.e.,

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

Then, applying the Golden Lemma (Lemma 2) 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. □

Definition 6 (information)**.** *The value*

$$I_r(A; B) \quad = \quad H_r(A) + H_r(B) - H_r(A, B). \tag{20}$$

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.

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

$$I(A; B) \quad = \quad \sum_{a\in\mathcal{A},b\in\mathcal{B}} p(a \wedge b) \left(\log \frac{1}{p(a)p(b)} - \log \frac{1}{p(a \wedge b)} \right). \tag{21}$$

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 non-negative, 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 arbitrary (possibly dependent) variables.

Fact 3 (Chain rule)**.**

$$H(A,B) \quad = \quad H(A|B) + H(B). \tag{22}$$

Proof. Let $\mathcal{B}^+ = \{b : p(b) > 0\}$. We calculate:

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

□

From equation (22) and Theorem 7, we immediately get the following.

Corollary 2. *For random variables A and B,*

$$H(A|B) \quad \leq \quad H(A). \tag{23}$$

Moreover, the equality holds iff A and B are independent.

The above can be interpreted that the entropy of A may only decrease if we additionally know B. Note however, that this inequality holds *on average*, and may not be true for a particular value of B.

Exercise Show an example where $H(A|b) > H(A)$.

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

$$\begin{aligned}
I(A;B) &= H(A) - H(A|B) &(24)\\
&= H(B) - H(B|A). &(25)
\end{aligned}$$

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 $A_1, A_2, \ldots, A_n$.

$$\begin{aligned}
H(A_1,\ldots,A_n) &= H(A_1|A_2,\ldots,A_n) + H(A_2,\ldots,A_n) \\
&= H(A_1|A_2,\ldots,A_n) + H(A_2|A_3,\ldots,A_n) + H(A_3,\ldots,A_n) \\
&= \sum_{i=1}^{n} H(A_i|A_{i+1},\ldots,A_n) &(26)
\end{aligned}$$

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

Fact 4 (Conditional chain rule)**.**

$$H(A,B|C) \quad = \quad H(A|B,C) + H(B|C). \tag{27}$$

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

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

Then we have

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

To make sure that the respective conditional probabilities are defined, we may assume that b ranges over those values for which $p(b \wedge c) > 0$. (The equations still hold, since the other summands disappear.)

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. □

We leave to the reader to show that

$$H(A, B|C) \quad \leq \quad H(A|C) + H(B|C) \tag{28}$$

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

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

The proof can go along the same lines as on the page 13.

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

$$\begin{aligned} I(A; B|C) &= H(A|C) + H(B|C) - \underbrace{H(A, B|C)}_{H(A|B,C)+H(B|C)} \qquad (29) \\ &= H(A|C) - H(A|B, C). \qquad (30) \end{aligned}$$

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

$$R(A; B; C) \quad = \quad I(A; B) - I(A; B|C). \tag{31}$$

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*.

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 easily make sure that $H(A|C) > 0$.

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.)

29. 10. 2009

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

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

(unique decodability).

Note that we do not require that C be a function of M and K, since the encoding need not, in general, be functional. It can, for example, use random bits (like in the Elgamal cryptosystem, see, e.g., [3]).

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

$$C \quad = \quad M \oplus K$$

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 7 and Definition 6), and to this end, it is enough to show

$$p(C = w|M = u) \quad = \quad p(C = w). \tag{32}$$

We have

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

On the other hand, we have

$$\begin{aligned} p(C = w|M = u) &= \frac{p(C = w \wedge M = u)}{p(M = u)} && (33)\\ &= \frac{p(K = u \oplus w \wedge M = u)}{p(M = u)} && (34)\\ &= \frac{p(K = u \oplus w) \cdot p(M = u)}{p(M = u)} && (35)\\ &= \frac{1}{2^n}. && (36) \end{aligned}$$

To infer (34) from (33), we used the fact that, in One time pad, the values of M and C determine the value of K; hence we have the equivalences

$$C = w \wedge M = u \iff K = u \oplus w \wedge C = w \wedge M = u \iff K = u \oplus w \wedge M = u.$$

This proves (32).

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

Theorem 8 (Shannon's Pessimistic Theorem)**.** *Any perfectly secret cryptosystem satisfies*

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

Consequently (c.f. Theorem 5)

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

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. □

As another application of the quantitative concept of information, we observe a property 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 3. *Suppose A and C are conditionally independent, given B (see page 16). 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. □

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

Corollary 3. *If f is a function then*

$$I(A;f(B)) \quad \leq \quad I(A;B). \tag{37}$$

Proof. Follows from the Lemma, since

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

□

Channels

Definition 7 (channel)**.** *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

$$p(B = b) \quad = \quad \sum_{a \in \mathcal{A}} P(a \to b) \cdot p(A = a). \tag{38}$$

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

$$C_\Gamma \quad = \quad \max_A I(A; B), \tag{39}$$

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

5.11.2009

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

$$(p(a_1), \ldots, p(a_m)) \cdot \begin{pmatrix} P_{11} & \ldots & P_{1n} \\ \ldots & \ldots & \ldots \\ P_{m1} & \ldots & P_{mn}, \end{pmatrix} \quad = \quad (p(b_1), \ldots, p(b_n)) \,. \tag{40}$$

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

$$0 \longrightarrow 0$$

$$1 \longrightarrow 1$$

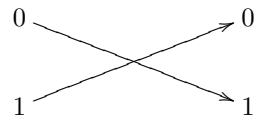
The matrix representation of this channel is

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

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

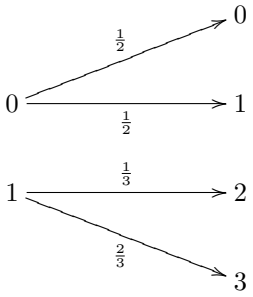
$$C_\Gamma = \max_A I(A; B) = \max_A H(A) = \log_2 |\mathcal{A}| = 1.$$

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 Here[3] 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 $next(a) = b$, $next(b) = c$, $\ldots$, $next(y) = z$, $next(z) = a$.
We leave to the reader to draw graphical and matrix representation.
To compute the capacity, first observe that, for any α,

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

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., $C_\Gamma = 0$?

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:

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

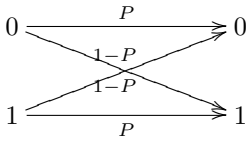
$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{6} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{6} & \frac{1}{3} \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

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

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

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 (4) that $0 \log_r 0 = 0 \log_r \frac{1}{0} = 0$, and let h denote the mapping

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

defined for $0 \leq x \leq 1$. 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 Lemma 1 (page 5), the function $h(x)$ is strictly convex on $[0, 1]$, and it readily implies that so is the function

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

Taking in the definition of convexity (2) $x_1 = q$, $x_2 = \bar{q}$, and $\lambda = P$ (hence $\lambda x_1 + (1-\lambda)x_2 = r$), and noting that $h(q) = h(\bar{q})$, we obtain that

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

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

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

$$H(s) = -s\log_2 s - (1-s)\log_2(1-s) \tag{41}$$

(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

$$C_\Gamma = \max_A H(B) - H(B|A) = 1 - H(P). \tag{42}$$

Decision rules

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

In some cases the answer is simple. For example, in the inverse faithful channel (page 21), we should just interchange 0 and 1. However, for the noisy typewriter (page 21), no "sure" decoding exists. For instance, an output word *afu*, can result from input *zet*, but also from *aft*, and many others[4] (but not, e.g., from input *abc*).

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

The quality of the rule is measured by

$$Pr_C(\Delta, A) =_{def} p(\Delta \circ B = A), \tag{43}$$

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

where (A, B) forms an input–output pair[5]. We have from definition

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

The last term can be decomposed in two ways using conditional probabilities.

$$p(B = b \wedge A = \Delta(b)) = p(A = \Delta(b)) \cdot \underbrace{p(B = b | A = \Delta(b))}_{P(\Delta(b) \to b)} = p(B = b) \cdot p(A = \Delta(b) | B = b).$$

This gives us two formulas to compute $Pr_C(\Delta, A)$

$$\begin{aligned} Pr_C(\Delta, A) &= \sum_{b \in \mathcal{B}} p(A = \Delta(b)) \cdot P(\Delta(b) \to b) && (44) \\ &= \sum_{b \in \mathcal{B}} p(B = b) \cdot p(A = \Delta(b) | B = b), && (45) \end{aligned}$$

both useful.

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, e.g., by

$$Pr_E(\Delta, A) = \sum_{a \in \mathcal{A}} p(A = a) \cdot p(\Delta \circ B \neq a | A = a) \quad (46)$$

We are generally 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 the following.

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) by

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

(Formally, this definition requires that $p(B = b) > 0$; but if $p(B = b) = 0$, we can define $\Delta(b)$ arbitrarily, and it will not affect (43).) Clearly, this choice maximizes the right-hand side of (45). Hence, we have

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

for any rule Δ. Note however, that $Pr_C(\Delta_o, A)$ can be smaller than 1 (hence $Pr_E(\Delta_o, A) > 0$) if, for some b, the maximal value of $p(a|b)$ is achieved with two different a's.

Exercise Calculate $Pr_C(\Delta_o, A)$ for the "bad" channels on page 21.

A disadvantage of the ideal observer rule is that it requires some *a priori* knowledge about the message to be sent. If the distribution of A is unknown, a reasonable choice is the following.

[5] In this case, the distribution of B is determined by the distribution of A by the equation (38), hence the definition is correct.

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.[6],

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

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

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 $\mathcal{P}$; hence $\mathbf{p}(a) = p(A = a)$. Now, using (44), 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 intuitively clear that the value of $\int_{\mathbf{p} \in \mathcal{P}} \mathbf{p}(a)\, d\mathbf{p}$ does not depend on a particular choice of $a \in \mathcal{A}$. (A formal argument refers to the concept of Lebesgue integral. Note however 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 achieved with the maximal likelihood rule.

Next lecture:

- Multiple use of channel.
- Improving reliability.

References

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

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

[3] John Talbot and Dominic Welsh. *Complexity and Cryptography*, Cambridge University Press, 2006.

[4] Peter Winkler. *Mathematical Puzzles: A Connoisseur's Collection*. A K Peters, 2004.

Damian Niwiński, Warsaw University. Last modified: 11.11.2009.

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