Permanents and the power of entropy

Chapter 37

In Chapter 24 we discussed Van der Waerden's conjecture, which established a *lower* bound for the permanent of a doubly stochastic matrix. There is also a wonderful theorem that gives an *upper* bound for integral matrices with prescribed row sums.

Consider an $n \times n$ matrix $M = (m_{ij})$ with $0/1$-entries. To M we associate a simple bipartite graph $G_M = (U \cup V, E)$, whose vertices are given by $U = \{u_1, \dots, u_n\}$ and $V = \{v_1, \dots, v_n\}$, and where

$$u_i v_j \in E \quad :\Longleftrightarrow \quad m_{ij} = 1.$$

Conversely, every bipartite graph G on $n + n$ nodes gives rise to a square $0/1$-matrix M of size $n \times n$ with $G = G_M$. Now look at the permanent

$$\operatorname{per} M \;:=\; \sum_{\sigma} m_{1\sigma(1)} \cdots m_{n\sigma(n)}.$$

Every term $m_{1\sigma(1)} m_{2\sigma(2)} \cdots m_{n\sigma(n)}$ equals 0 or 1, and it is equal to 1 if and only if the set of edges $\{u_1 v_{\sigma(1)}, \dots, u_n v_{\sigma(n)}\}$ is a *perfect matching* of G_M, that is, a set of edges that covers each vertex exactly once. Hence the number $m(G_M)$ of perfect matchings in G_M is just the permanent, that is, $\operatorname{per} M = m(G_M)$.

The all 1's matrix J_n corresponds to the complete bipartite graph $K_{n,n}$, with $\operatorname{per}(J_n) = m(K_{n,n}) = n!$.

The correspondence $G \longleftrightarrow M_G$ stimulated a lot of the early research on permanents. One of the first difficult problems was a conjecture posed by Henryk Minc in 1967: Suppose the $0/1$-matrix M has row sums $d_1, \dots, d_n$ (or equivalently the vertices $u_1, \dots, u_n$ have degrees $d_1, \dots, d_n$), then

$$\operatorname{per} M \;\leq\; \prod_{i=1}^{n} \left(d_i!\right)^{1/d_i}.$$

Observe that we can have equality, as seen from the example in the margin.

If k divides n, the block diagonal matrix

$$M = \begin{pmatrix} J_k & & \\ & \ddots & \\ & & J_k \end{pmatrix}$$

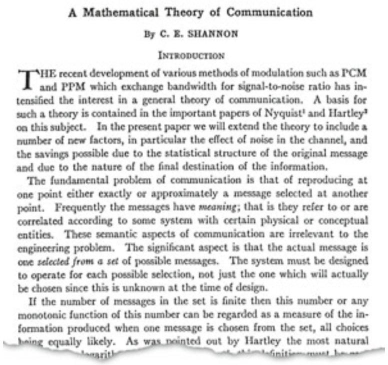
with $\frac{n}{k}$ blocks has $d_1 = \cdots = d_n = k$ and $\operatorname{per} M = (k!)^{n/k}$.

Minc's conjecture was proved by Lev M. Brégman in 1973. A few years later Alexander Schrijver gave a short and sweet proof, with a randomized version appearing in the book of Alon and Spencer. But in our view the proof straight from the BOOK is due to Jaikumar Radhakrishnan. It is not much different, but it uses just the right tool — *entropy* from information theory. Before we come to this, let us state Brégman's theorem again.

M. Aigner, G. M. Ziegler, *Proofs from THE BOOK*, https://doi.org/10.1007/978-3-662-57265-8_37

Theorem 1. *Let $M = (m_{ij})$ be an $n \times n$ matrix with entries in $\{0,1\}$, and let $d_1, \dots, d_n$ be the row sums of M, that is, $d_i = \sum_{j=1}^{n} m_{ij}$. Then*

$$\operatorname{per} M \;\leq\; \prod_{i=1}^{n} (d_i!)^{1/d_i}.$$



A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

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

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural

It does not happen often that a single paper gives birth to a whole field. Claude Shannon's *A Mathematical Theory of Communication* from 1948 was such a singular achievement: It laid the foundations of information theory and coding, and thereby initiated one of the great mathematical success stories of the twentieth century.

Suppose X is a random variable taking values in $\{a_1, \dots, a_n\}$ with probabilities $\text{Prob}(X = a_i) = p_i$. It helps to think of X as an experiment with possible outcomes a_i, like throwing a die with outcomes $1, 2, \dots, 6$. How much information do we receive (on the average) from performing the experiment? Shannon's ingenious idea was the "equation"

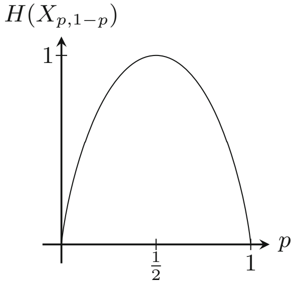
$$\text{information after} \;=\; \text{uncertainty before}.$$

For example, when a coin is rigged and heads comes up most of the time, then there is little information to be gained from throwing it, certainly less than when the coin is fair, in which case the uncertainty (and information) is largest.

It is said that Shannon, following the advice of John von Neumann, used the name "entropy" because nobody knew exactly what this meant anyway ...

By postulating certain natural conditions that an uncertainty measure for X should satisfy, Shannon arrived at his famous definition of *entropy*, which he denoted by $H(X)$:

$$H(X) \;=\; H(X_{p_1,\dots,p_n}) \;:=\; -\sum_{i=1}^{n} p_i \log_2 p_i.$$



For example, if X is a throw of a biased coin with $\text{Prob}(X = \text{heads}) = p$, then the Shannon formula yields the function $H(X_{p,1-p}) = -p \log_2 p$ $-(1-p)\log_2(1-p)$ graphed in the margin.

In the following we always use the binary logarithm $\log_2 p$ with the convention $p \log_2 p = 0$ for $p = 0$. The *support* of the random variable X is $\operatorname{supp} X := \{a : \text{Prob}(X = a) > 0\}$.

Later in his paper Shannon gave an alternative interpretation of $H(X)$ as the expected length of an optimal question strategy for the outcome of X. The appendix to this chapter contains a sketch of this approach.

Suppose X and Y are two random variables with value ranges $\{a_1, \dots, a_m\}$ and $\{b_1, \dots, b_n\}$. A key ingredient for Radhakrishnan's proof is the concept of *conditional entropy of Y under knowledge of X*. To shorten the writing, let us set $p(a_i) := \text{Prob}(X = a_i)$, $p(b_j) := \text{Prob}(Y = b_j)$, and similarly $p(a_i, b_j) := \text{Prob}(X = a_i \wedge Y = b_j)$ for the joint distribution of the pair (X, Y), which may be viewed as a single random variable, and $p(b_j \,|\, a_i) := \text{Prob}(Y = b_j \,|\, X = a_i)$ for the conditional probabilities. Let

$$H(Y \,|\, a_i) \;:=\; -\sum_{j=1}^{n} p(b_j \,|\, a_i) \log_2 p(b_j \,|\, a_i)$$

be the entropy (uncertainty) of Y if we know that the outcome of X is a_i. Now we take the expected value of this quantity over all possible outcomes of X and thus arrive at

$$H(Y \,|\, X) \;:=\; \sum_{i=1}^{m} p(a_i) H(Y \,|\, a_i)$$

as the conditional entropy of Y under knowledge of X.

In particular, $H(Y \,|\, X) = 0$ if and only if the outcome of Y is determined once the result of X is known.

All we need for the proof of Brégman's theorem are three facts about entropy, whose (easy) proofs are given in the appendix; the rest is clever and beautiful probabilistic reasoning. Here are the facts:

(A) $H(X) \leq \log_2(|\text{supp}\, X|)$, *with equality if and only if* X *is uniformly distributed on the support of* X*, that is,* $\text{Prob}(X = a) = \frac{1}{n}$ *for* $a \in \text{supp}\, X$*, where* $n = |\text{supp}\, X|$.

(B) $H(X, Y) = H(X) + H(Y \,|\, X)$*, and more generally* $H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \,|\, X_1) + \cdots + H(X_n \,|\, X_1, \ldots, X_{n-1})$.

(C) *If* $\text{supp}\, X$ *is partitioned into the* d *sets* $E_1, \ldots, E_d$*, where* $E_j := \{a \in \text{supp}\, X : |\text{supp}\,(Y \,|\, a)| = j\}$*, then*

$$H(Y \,|\, X) \;\leq\; \sum_{j=1}^{d} \text{Prob}(X \in E_j) \log_2 j.$$

■ **Proof of Theorem 1.** Let $G = (U \cup V, E)$ be the bipartite graph associated with M, where d_i is the degree of the vertex u_i, and denote by $\mathfrak{S}$ the set of perfect matchings of G. As per $M = m(G) = |\mathfrak{S}|$, we will prove the upper bound of the theorem for the number of perfect matchings of G. We may assume $\mathfrak{S} \neq \varnothing$ because otherwise there is nothing to show. We view each $\sigma \in \mathfrak{S}$ as the corresponding permutation $\sigma(1)\sigma(2)\ldots\sigma(n)$ of the indices. Hence the vertex $u_i \in U$ is matched to $v_{\sigma(i)} \in V$ under σ.

The first idea is to pick $\sigma \in \mathfrak{S}$ uniformly at random and to consider the vector of random variables $X = (X_1, \ldots, X_n) = (\sigma(1), \ldots, \sigma(n))$.
By **(A)**,

$$H(\sigma(1), \ldots, \sigma(n)) = \log_2(|\mathfrak{S}|);$$

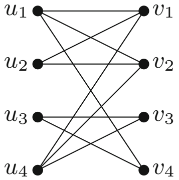
hence it suffices to show that

$$H(\sigma(1), \ldots, \sigma(n)) \;\leq\; \log_2\Big(\prod_{i=1}^{n} (d_i!)^{1/d_i}\Big) \;=\; \sum_{i=1}^{n} \frac{1}{d_i} \log_2(d_i!). \qquad (1)$$

Next we use **(B)** to get

$$H(\sigma(1),\dots,\sigma(n)) \;=\; \sum_{i=1}^{n} H(\sigma(i)\,|\,\sigma(1),\dots,\sigma(i-1)). \qquad (2)$$

Let's find out what the conditional entropy $H(\sigma(i)\,|\,\sigma(1),\dots,\sigma(i-1))$ means. It measures the uncertainty about the matching mate of u_i *after* the mates of $u_1,\dots,u_{i-1}$ have been revealed. In particular, the support of the random variable $\sigma(i)$ under knowledge of $(\sigma(1),\dots,\sigma(i-1))$ is contained in the set of indices of the neighbors of u_i that have *not* already been matched to one of $u_1,\dots,u_{i-1}$.



$\mathfrak{S} = \{1243, 2143, 4132, 4231\}$

For example, let us check the formula in **(B)** for the graph in the margin, which has $|\mathfrak{S}| = 4$. Since all permutations in $\mathfrak{S}$ are equally likely, we have $H(\sigma(1),\dots,\sigma(4)) = \log_2 4 = 2$. Now, $H(\sigma(1)) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}$ $-\frac{1}{2}\log_2\frac{1}{2} = \frac{3}{2}$. Let us compute the conditional entropy $H(\sigma(2)\,|\,\sigma(1))$: For $\sigma(1) = 1$ we get $H(\sigma(2)\,|\,1) = 0$ since $\sigma(2) = 2$ is then determined; similarly $H(\sigma(2)\,|\,2) = 0$, but for $\sigma(1) = 4$ we have $H(\sigma(2)\,|\,4) = 1$, since there are two equally likely outcomes $\sigma(2) = 1, \sigma(2) = 2$. For the expected value we thus compute $H(\sigma(2)\,|\,\sigma(1)) = \frac{1}{2}\cdot 1 = \frac{1}{2}$. The next conditional entropies $H(\sigma(3)\,|\,\sigma(1),\sigma(2))$ and $H(\sigma(4)\,|\,\sigma(1),\sigma(2),\sigma(3))$ are both 0, since the values are determined. So summing up we again get $H(\sigma(1)) + H(\sigma(2)\,|\,\sigma(1)) + H(\sigma(3)\,|\,\sigma(1),\sigma(2)) + H(\sigma(4)\,|\,\sigma(1),\sigma(2),\sigma(3)) = \frac{3}{2} + \frac{1}{2} + 0 + 0 = 2$, in accordance with **(B)**.

Radhakrishnan's wonderful idea was to examine the vertices $u_1,\dots,u_n$ in a *random order* τ, where all τ are equally likely with probability $\frac{1}{n!}$, and then to take the average over the entropies. In other words, we reveal the matching mates in the order $\sigma(\tau(1)),\sigma(\tau(2)),\dots,\sigma(\tau(n))$. Let us look at a fixed τ. If $k_i = \tau^{-1}(i)$, that is, if in the ordering τ the vertex u_i appears in k_ith place, then equation (2) becomes

$$H(\sigma(1),\dots,\sigma(n)) = \sum_{i=1}^{n} H\big(\sigma(i)\,\big|\,\sigma(\tau(1)),\dots,\sigma(\tau(k_i-1))\big).$$

As this holds for all τ, taking the average we get

$$H(\sigma(1),\dots,\sigma(n)) = \frac{1}{n!}\sum_{\tau}\Big(\sum_{i=1}^{n} H\big(\sigma(i)\,\big|\,\sigma(\tau(1)),\dots,\sigma(\tau(k_i-1))\big)\Big).$$

Let us fix τ and look at a summand

$$H\big(\sigma(i)\,\big|\,\sigma(\tau(1)),\dots,\sigma(\tau(k_i-1))\big). \qquad (3)$$

To upper bound (3) we use fact **(C)** from above, applied to the random variables $X = \big(\sigma(\tau(1)),\dots,\sigma(\tau(k_i-1))\big)$ and $Y = \sigma(i)$. For each σ let $N_i(\sigma,\tau)$ be the set of indices of the neighbors of u_i that are *not* among $\{\sigma(\tau(1)),\dots,\sigma(\tau(k_i-1))\}$. Since u_i has d_i neighbors and σ is a perfect matching we have $1 \le |N_i(\sigma,\tau)| \le d_i$ for all σ. Now partition $\operatorname{supp} X$ into the sets $E_{i,j}^{(\tau)}$, where $\big(\sigma(\tau(1)),\dots,\sigma(\tau(k_i-1))\big)$ lies in $E_{i,j}^{(\tau)}$ if and only if $|N_i(\sigma,\tau)| = j$, for $1 \le j \le d_i$. Considering $|N_i(\sigma,\tau)|$ as a random variable on $\mathfrak{S}$, we thus have

$$\text{Prob}\big(X \in E_{i,j}^{(\tau)}\big) \;=\; \text{Prob}\big(|N_i(\sigma,\tau)| = j\big),$$

and fact **(C)** tells us that for fixed τ

$$H\big(\sigma(i)\,\big|\,\sigma(\tau(1)),\ldots,\sigma(\tau(k_i-1))\big) \;\le\; \sum_{j=1}^{d_i} \text{Prob}\big(|N_i(\sigma,\tau)| = j\big)\log_2 j.$$

Hence we get altogether

$$H\big(\sigma(1),\ldots,\sigma(n)\big) \;\le\; \frac{1}{n!}\sum_{i=1}^{n}\sum_{j=1}^{d_i}\log_2 j \sum_{\tau}\text{Prob}\big(|N_i(\sigma,\tau)| = j\big). \quad (4)$$

This seems to get more complicated as we go along — but wait! Looking at (1) it suffices to show that the innermost sum in (4) equals $n!\frac{1}{d_i}$ for all j, because then the right-hand side simplifies to $\sum_{i=1}^{n}\frac{1}{d_i}\log_2(d_i!)$.

And this assertion about the inner sum is easy! Fix σ, and let $\ell_1,\ldots,\ell_{d_i}$ be the indices of the neighbors of u_i, $D_\sigma = \{\sigma^{-1}(\ell_1),\ldots,\sigma^{-1}(\ell_{d_i})\}$ is the set of indices of the U-vertices that are matched onto the neighbors of u_i, including of course i itself, and they appear according to the ordering of D_σ under τ. If i comes first in D_σ, then no neighbors had been taken so far, whence $|N_i(\sigma,\tau)| = d_i$. If i is second, then one neighbor is gone, thus $|N_i(\sigma,\tau)| = d_i - 1$, and so on.

Now the power of averaging comes into play. With τ running through all $n!$ permutations, all possible orderings of the list D_σ occur with equal frequency, which means that i appears in all d_i places of D_σ with the same frequency $\frac{n!}{d_i}$. But this, in turn, implies that $|N_i(\sigma,\tau)| = j$ occurs with frequency $\frac{n!}{d_i}$ for all j, and this holds for all σ, whence

$$\sum_{\tau}\text{Prob}\big(|N_i(\sigma,\tau)| = j\big) \;=\; \frac{n!}{d_i},$$

for all j, and we are done. □

We cannot end this chapter without deriving a stunning asymptotic formula for the number $L(n)$ of Latin squares of order n. (See Chapter 36 for the definition of Latin squares.) The small examples

$$L(1) = 1,\ L(2) = 2,\ L(3) = 12,\ L(4) = 576,\ L(5) = 161280$$

suggest that $L(n)$ grows exceedingly fast. So, all we can hope for are good bounds — and these are miraculously supplied by Brégman's Theorem and by the permanent theorem discussed in Chapter 24.

$n = 1$:

1

$n = 2$:

1	2
2	1

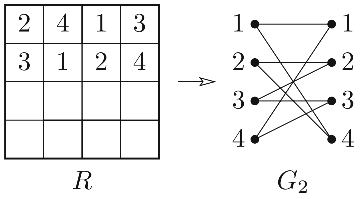
2	1
1	2

$n = 3$:

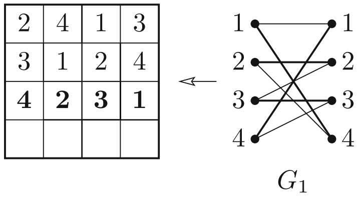
1	2	3
2	3	1
3	1	2

There are $3!2! = 12$ fillings of the first row and the first column; the rest is then determined.

Take an empty $n \times n$ square and fill it row by row with the numbers $1,\ldots,n$, so that the resulting configuration is a Latin square. There are $n!$ ways to fill the first row, since we may take every permutation. Suppose the first



2	4	1	3
3	1	2	4



2	4	1	3
3	1	2	4
4	**2**	**3**	**1**

$$M_2 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix},$$

$$\text{per}\, M_2 = 2.$$

$n-k$ rows are properly filled to give an $(n-k)\times n$ Latin rectangle R. In how many ways can we fill the next row? Consider the following bipartite graph $G_k = (U \cup V, E)$, where U is the set of elements and V the set of column positions, with

$$ij \in E \quad :\Longleftrightarrow \quad i \text{ does } \textit{not} \text{ appear in the } j\text{th column of } R.$$

So, exactly the numbers that are joined to j can be used in column j of the $(n-k+1)$st row. In other words, a proper filling of the next row corresponds to a perfect matching of G_k. Now, every element $i \in U$ appears $n-k$ times in R, hence it is available in k columns for the next row. Thus i has degree k in G_k and similarly $d(j) = k$ for $j \in V$. (We used this argument already in the proof of Lemma 1 in Chapter 36.)

Let M_k be the $0/1$-matrix corresponding to G_k, thus

$$\text{per}\, M_k = \text{the number of proper fillings of row } n-k+1.$$

Every row and column in M_k sums to k; let us denote the set of $0/1$-matrices with this property by $\mathcal{M}(n,k)$. The permanent per M_k depends, of course, on the setup of R, but if we have general lower and upper bounds for matrices in $\mathcal{M}(n,k)$, then by taking the product over all k, we obtain lower/upper bounds for $L(n)$.

By Brégman's Theorem with $d_1 = d_2 = \cdots = d_n = k$ we get right away

$$\text{per}\, M \;\le\; k!^{\frac{n}{k}} \qquad \text{for all } M \in \mathcal{M}(n,k).$$

Now to the lower bound: If M is in $\mathcal{M}(n,k)$, then $\frac{1}{k}M$ is doubly stochastic, which implies by the permanent theorem in Chapter 24 that

Note that per $\lambda M = \lambda^n$per M for an $n \times n$ matrix M.

$$\text{per}\, M = k^n \text{per}\left(\frac{1}{k}M\right) \ge k^n \frac{n!}{n^n}.$$

In summary, we have proved the following remarkable bounds.

Theorem 2. *The number $L(n)$ of Latin squares of order n is bounded by*

$$\frac{n!^{2n}}{n^{n^2}} \;\le\; L(n) \;\le\; \prod_{k=1}^{n} k!^{n/k}.$$

The case $k = n$ corresponds to the $n!$ fillings of the first row.

Using the approximations for $n!$ from page 13

$$\left(\frac{n}{e}\right)^n \;<\; n! \;<\; en\left(\frac{n}{e}\right)^n, \tag{5}$$

we can easily derive from this the following astonishingly simple asymptotic formula.

Corollary. *In the limit, the number $L(n)$ of Latin squares of order n satisfies*

$$\lim_{n\to\infty} \frac{L(n)^{1/n^2}}{n} \;=\; \frac{1}{e^2}.$$

■ **Proof.** For the lower bound we get

$$L(n) \geq \frac{n!^{2n}}{n^{n^2}} > \frac{(\frac{n}{e})^{2n^2}}{n^{n^2}} = \Big(\frac{n}{e^2}\Big)^{n^2},$$

so

$$\frac{L(n)^{1/n^2}}{n} > \frac{1}{e^2} \quad \text{and thus} \quad \lim_{n\to\infty} \frac{L(n)^{1/n^2}}{n} \geq \frac{1}{e^2}.$$

The upper bound needs a little more work. We will show that for any $\varepsilon > 0$

$$\frac{L(n)^{1/n^2}}{n} < \frac{1}{e^2}(1+\varepsilon)$$

holds when n is large enough. For convenience we set $\mathcal{L}(n) = L(n)^{1/n^2}$. Using (5) for k in place of n, we have

$$\begin{aligned} \log \mathcal{L}(n) &\leq \frac{1}{n}\log\prod_{k=1}^{n}(k!)^{\frac{1}{k}} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{k}\log k! \\ &< \frac{1}{n}\sum_{k=1}^{n}\frac{1}{k}\log\big(ek\Big(\frac{k}{e}\Big)^k\big) \\ &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{k}\big(1+\log k + k\log k - k\big) \\ &= \frac{1}{n}\Big[\sum_{k=1}^{n}\frac{1}{k} + \sum_{k=1}^{n}\frac{\log k}{k} + \sum_{k=1}^{n}\log k - n\Big]. \qquad (6) \end{aligned}$$

Now, look at page 13. The first sum is the harmonic number H_n, where $H_n < \log n + 1$. The third sum was also treated there, with

$$\sum_{k=1}^{n}\log k \leq (n+1)\log(n+1) - n \leq (n+2)\log n - n,$$

where the second inequality is derived in the margin, for $n \geq 6$. For the second sum in (6) the same integration method that we had used for the third one yields, using that $\frac{\log x}{x}$ is positive for $x > 1$ and monotonically decreasing for $x > e$, that

$$\sum_{k=4}^{n}\frac{\log k}{k} < \int_1^n \frac{\log x}{x}dx = \Big[\tfrac{1}{2}(\log x)^2\Big]_1^n = \tfrac{1}{2}(\log n)^2.$$

The inequality $(n+1)^{n+1} \leq n^{n+2}$ holds for $n \geq 6$: It may be rewritten as

$$\Big(1+\frac{1}{n}\Big)^n\Big(1+\frac{1}{n}\Big) \leq n,$$

where $(1+\frac{1}{n})^n < e$ and $1+\frac{1}{n} \leq 2$; thus the left side is less than $2e < 6 \leq n$.

Thus the second sum in (6) is smaller than $2 + \frac{1}{2}(\log n)^2$.

Putting everything together, we get

$$\log \mathcal{L}(n) < \frac{3\log n}{n} + \frac{3}{n} + \frac{(\log n)^2}{2n} + \log n - 2.$$

The first three terms go to 0 as n gets large, and we conclude that for every $\delta > 0$

$$\log \mathcal{L}(n) \leq \delta + \log n - 2$$

will hold if n is large enough. Thus we get $L(n)^{1/n^2} \leq \frac{n}{e^2}e^{\delta}$ for all large enough n, and this is what we wanted to prove. □

Appendix: More about entropy

What was Shannon's alternative approach to entropy?

As before, let X be a random variable with value set $\{a_1, \dots, a_n\}$ and $p_i = \text{Prob}(X = a_i)$. We employ a certain strategy $\mathcal{S}$ of yes/no questions until we know the value of X for sure. If our strategy leads us to ask ℓ_i questions in the case of the outcome $X = a_i$, then $\overline{L}(\mathcal{S}) := \sum_{i=1}^n p_i \ell_i$ is the expected number of questions. Of course, a good strategy will want to ask few questions for very likely outcomes a_i (when p_i is large), so as to minimize the average number.

As an example, suppose that the probabilities for throwing a loaded die are $p_1 = \frac{1}{3}$, $p_2 = p_3 = \frac{1}{8}$, $p_4 = \frac{1}{6}$, and $p_5 = p_6 = \frac{1}{8}$. A strategy might be the following. First question: "Is the outcome ≤ 3?" If yes, which happens with probability $\frac{7}{12}$, ask the second question: "Is it 1?" If yes again, we are done, otherwise we need one more question to decide whether the throw shows 2 or 3. Proceeding in analogous fashion if the first answer was no, we get $\ell_1 = 2$, $\ell_2 = \ell_3 = 3$, $\ell_4 = 2$, $\ell_5 = \ell_6 = 3$, thus

$$\overline{L}(\mathcal{S}) \;=\; 2(\tfrac{1}{3} + \tfrac{1}{6}) + 3(\tfrac{1}{8} + \tfrac{1}{8} + \tfrac{1}{8} + \tfrac{1}{8}) = \tfrac{5}{2}.$$

Shannon now proved that the entropy $H(X) = -\sum_{i=1}^n p_i \log_2 p_i$ is a lower bound for the expected number of questions $\overline{L}(\mathcal{S}) = \sum_{i=1}^n p_i \ell_i$ for every *conceivable* strategy $\mathcal{S}$. Let us check this! First we have that $\sum_{i=1}^n 2^{-\ell_i} = 1$ (why?), and the inequality $\log_2 x \leq x - 1$ for $x > 0$ together with $\sum_{i=1}^n p_i = 1$ yields

$$\sum_{i=1}^n p_i \log_2 \frac{2^{-\ell_i}}{p_i} \;\leq\; \sum_{i=1}^n p_i \Big(\frac{2^{-\ell_i}}{p_i} - 1\Big) \;=\; \sum_{i=1}^n 2^{-\ell_i} - \sum_{i=1}^n p_i \;=\; 0.$$

But this means that $-\sum_{i=1}^n p_i \ell_i \leq \sum_{i=1}^n p_i \log_2 p_i$, or $\overline{L}(\mathcal{S}) \geq H(X)$.

Conversely, it is easy to find a strategy $\mathcal{S}_0$ with $\overline{L}(\mathcal{S}_0) < H(X) + 1$, hence

$$H(X) \;\leq\; \overline{L}(X) \;=\; \min_{\mathcal{S}} \overline{L}(\mathcal{S}) \;<\; H(X) + 1.$$

The actual minimum $\overline{L}(X)$ can e. g. be computed by Huffman's algorithm, a classic in computer science.

Looking at n-fold repetitions X^n of the experiment X, Shannon went on to show that the expected number of questions per experiment $\frac{1}{n}\overline{L}(X^n)$ used by optimal strategies for X^n converges to $H(X)$ for $n \to \infty$. (Shannon called this the "Fundamental theorem for a noiseless channel.")

Now to the three facts that we used in the proof of Theorem 1.

(A) $H(X) \leq \log_2(|\text{supp}\, X|)$.

■ **Proof.** Assume without loss of generality that $p_i > 0$ for all i. Consider the general form of the AM-GM inequality $a_1^{p_1} \cdots a_n^{p_n} \leq p_1 a_1 + \cdots + p_n a_n$ on page 144. Set $a_i = \frac{1}{p_i}$ and take the logarithm to obtain

Remember $0 \cdot \log_2 0 = 0$.

$$\sum_{i=1}^n p_i \log_2 \frac{1}{p_i} \;\leq\; \log_2 \Big(\sum_{i=1}^n p_i \frac{1}{p_i}\Big) \;=\; \log_2 n.$$

Equality holds if and only if $p_1 = \cdots = p_n = \frac{1}{n}$, that is, if we have uniform distribution. □

(B) $H(X,Y) = H(X) + H(Y\,|\,X)$.

■ **Proof.** We use the same notation as before and compute

$$\begin{aligned} H(X,Y) &= -\sum_{i,j} p(a_i,b_j)\log_2 p(a_i,b_j) \\ &= -\sum_{i,j} p(a_i,b_j)\log_2 \big(p(a_i)p(b_j\,|\,a_i)\big) \\ &= -\sum_{i,j} p(a_i,b_j)\log_2 p(a_i) - \sum_{i,j} p(a_i)p(b_j\,|\,a_i)\log_2 p(b_j\,|\,a_i) \\ &= -\sum_{i=1}^{m} p(a_i)\log_2 p(a_i) + H(Y\,|\,X) = H(X) + H(Y\,|\,X). \end{aligned}$$

The general formula follows by induction. □

(C) $H(Y\,|\,X) \leq \sum\limits_{j=1}^{d} \text{Prob}(X \in E_j)\log_2 j$.

■ **Proof.** We have $H(Y\,|\,X) = \sum_{i=1}^{m} p(a_i)H(Y\,|\,a_i)$. Partitioning the set $\{a_1,\ldots,a_m\}$ into the subsets E_j given by the assumption and using **(A)** we get

$$\begin{aligned} H(Y\,|\,X) &= \sum_{j=1}^{d}\sum_{a\in E_j} p(a)H(Y\,|\,a) \\ &\leq \sum_{j=1}^{d}\sum_{a\in E_j} p(a)\log_2 j \;=\; \sum_{j=1}^{d}\text{Prob}(X \in E_j)\log_2 j. \end{aligned}$$ □

References

[1] N. ALON & J. SPENCER: *The Probabilistic Method,* Third edition, Wiley-Interscience 2008.

[2] L. BRÉGMAN: *Some properties of nonnegative matrices and their permanents,* Soviet Math. Doklady **14** (1973), 945-949.

[3] A. KHINCHIN: *Mathematical Foundations of Information Theory,* Dover Publications 1957.

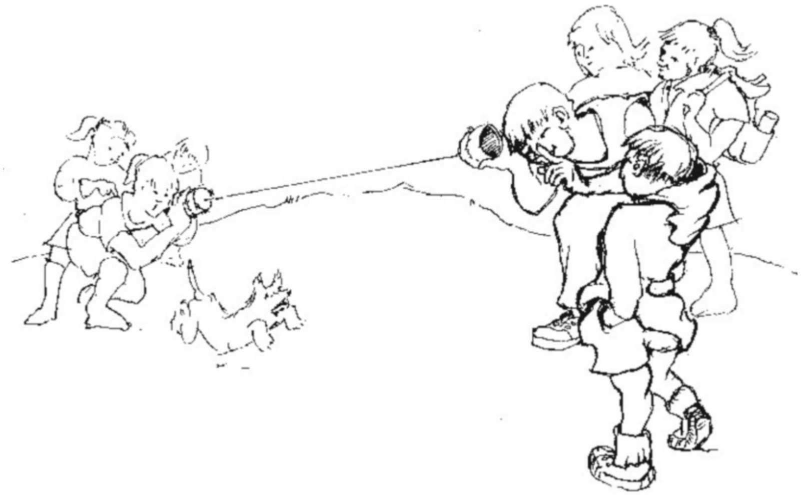
[4] B. D. MCKAY & I. M. WANLESS: *On the number of Latin squares,* Annals of Combinatorics **9** (2005), 335-344.

[5] J. RADHAKRISHNAN: *An entropy proof of Bregman's theorem,* J. Combinatorial Theory, Ser. A **77** (1997), 161-164.

[6] A. SCHRIJVER: *A short proof of Minc's conjecture,* J. Combinatorial Theory, Ser. A **25** (1978), 80-83.

[7] H. MINC: *Permanents,* Encyclopedia of Mathematics and its Applications, Vol. 6, Addison-Wesley, Reading MA 1978; reissued by Cambridge University Press 1984.

[8] C. SHANNON: *A Mathematical Theory of Communication,* Bell System Technical Journal **27** (1948), 379-423, 623-656.



"Do you get any news?"

"Sure! $-\sum_i p_i \log_2 p_i$ *of them!"*