To appear in Proc. ISCTA '99

The Discrete Noiseless Channel Revisited.*

Aamod Khandekar, Robert McEliece, and Eugene Rodemich
California Institute of Technology
Pasadena, California, USA

Abstract: *In this paper, we will discuss two of the lesser-known gems, viz., Theorems 1 and 8, from Shannon's 1948 classic paper, A Mathematical Theory of Communication [9], both dealing with the capacity of what he called the* discrete noiseless channel. *Although our paper is largely tutorial, it contains a few new ideas and results, including: a careful treatment of the case of symbols of different durations, including non-integer durations, the introduction of a new proof technique, which we call the "partition function" technique, and a theorem on the maximum entropy of a Markov chain defined on a digraph, subject to a constraint on the average edge duration.*

1. Introduction.

This paper deals with the *discrete noiseless channel*, which was the first topic treated in Shannon's classic 1948 paper [9, Section 1]. To paraphrase Shannon, a discrete noiseless channel is a channel which allows the noiseless transmission of a sequence of symbols chosen from a finite alphabet A, each symbol having a certain duration in time, possibly different for different symbols. Furthermore, there may be restrictions on the allowed sequences of symbols from A. The question is, what is the capacity of such a channel to transmit information?

To be more precise, we suppose A is a q-letter alphabet, and that associated with each letter $a \in A$ is a positive number $\tau(a)$ called the *duration* of a. Here are two examples: (1) the *standard q-ary alphabet*, with $A = \{0, 1, \ldots, q-1\}$, and $\tau(0) = \tau(1) = \cdots \tau(q-1) = 1$; and (2) *Shannon's telegraphy alphabet*, with $A = \{d, D, s, S\}$ and durations given in the following table: [1]

(1.1)

a	$\tau(a)$
d	2
D	4
s	3
S	6

A *word of length k* over A is a finite string of k letters from A. If $\alpha = a_1 a_2 \cdots a_k$ is such a word, its duration is defined to be $\tau(\alpha) = \tau(a_1) + \cdots + \tau(a_k)$. For example, 010110 is a word of length 6 and duration 6 over the standard binary alphabet, 01221022 is a word of length 8 and duration 8 over the standard 3-ary alphabet, and

$$dddsddddsdDsDdsDdsDDDsDdS$$

is a word of length 25 and duration 74 over Shannon's telegraphy alphabet.

A *language* $\mathcal{L}$ over A is a collection of words over A. The discrete noiseless channel associated with $\mathcal{L}$, the $\mathcal{L}$-channel for short, is the channel which is only allowed to transmit sequences from $\mathcal{L}$, although it transmits them without error. We ask, what is the capacity of the $\mathcal{L}$-channel?

For a general language $\mathcal{L}$, not much can be said, and so we must restrict the class of languages, as did Shannon. In this paper we will consider only *Shannon* languages. Roughly, a Shannon language is defined by a directed graph whose edges are labelled with letters from the alphabet A. The corresponding language $\mathcal{L}$ is then defined to be the set of words that result by reading off the edge labels on paths of the graph. (We will give a more precise definition in Section 4.)

For example, if the graph consists of a single vertex v and q self-loops at v, each labelled with a different element of A, the resulting language consists of all possible sequences over A. We will call this the *free*

* This work was partially supported by NSF grant no. CCR-9804793, and grants from Sony and Qualcomm.

[1] In Shannon's terminology, d is "dot", D is "dash", s is "letter space", and S is "word space".

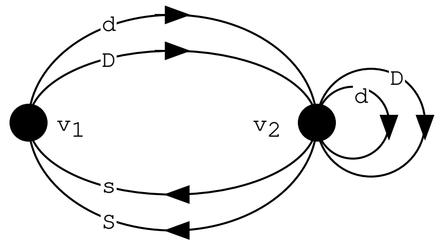

Figure 1. A directed graph with the branches labelled by symbols from Shannon's telegraphy alphabet.

language over A. As an example of a more resrictive type, consider the directed graph of Figure 1, in which each of the edges is labelled with an element from Shannon's telegraphy alphabet. The resulting language consists of all sequences of d's, D's, s's and S's, except that no word containing two consecutive spaces is permitted.

Shannon gave, in effect, two different definitions for the capacity of a noiseless channel corresponding to a Shannon language $\mathcal{L}$, and then showed that the two definitions gave the same value. This common value was thereby established unambiguously as the as the maximum rate, in bits[2] per second, that information can be transmitted over the channel. We shall call these two definitions the *combinatorial* capacity and the *probabilistic* capacity.

If $\mathcal{L}$ is a Shannon language, we denote the total number of words in $\mathcal{L}$ of duration t by $N(t)$. The combinatorial capacity of the $\mathcal{L}$-channel is defined as[3]

$$C_{\text{comb}} \triangleq \limsup_{t\to\infty} \frac{1}{t} \log N(t). \tag{1.2}$$

Similarly, let (X_n) be a stationary discrete Markov chain defined on the labelled graph defining $\mathcal{L}$. If the entropy of (X_n) is H and the average branch duration is T, then the *information rate* of (X_n), in bits per second, is H/T. The probabilistic capacity of the $\mathcal{L}$-language is defined to be the maximum of this rate, over all possible Markov chains:

$$C_{\text{prob}} \triangleq \sup_{(X_n)} \frac{H}{T}. \tag{1.3}$$

As a very simple example, for the free language over the standard q-ary alphabet,

$$N(t) = \begin{cases} q^t & \text{if } t \text{ is an integer} \\ 0 & \text{if } t \text{ is not an integer,} \end{cases}$$

and so by (1.2), $C_{\text{comb}} = \log_2 q$. Similarly, it is easy to see in this case that the maximum-rate Markov chain is a sequence of i.i.d. uniform random variables, for which $H = \log q$ and $T = 1$, so $C_{\text{prob}} = \log q$ as well.

In his original paper, Shannon gave a simple algebraic method for computing C_{comb} (Theorem 1), and a showed that the same value held for C_{prob} (Theorem 8). His proofs, however, were brief and in places quite cryptic. In this paper, we will give complete, and we hope clear, proofs of both of these important results. We will first treat the important special case of the free languages. The proofs in this case are quite simple, but they serve to introduce the more difficult ideas needed in the general case. Here is a guide to our paper:

• Shannon's Theorem 1 (C_{comb}): For free languages, Section 2; for general Shannon languages, Section 4.

[2] Assuming the implied logarithms are base-2. In this paper, however, we will mostly use natural logarithms, and so information will be measured in nats.

[3] Actually, Shannon used "lim" instead of "lim sup," but in order to be fully rigorous, we will need the more precise term.

• Shannon's Theorem 8 (C_{prob}): For free languages, Section 3; for general Shannon languages, Section 5.

In Section 6, we briefly sketch a kind of generalization of Shannon's Theorem 8 which we hope will be of interest, and then in Section 7 we give some concluding remarks.

2. The Combinatorial Capacity of Free Languages.

The *free language* A^* over the alphabet A is simply the set of all possible strings of all lengths over A. To find the combinatorial capacity of the A^*-channel, we first define the *partition function*[4] for A as follows:

$$P(s) = \sum_{a \in A} e^{-s\tau(a)}, \quad \text{for } s \geq 0.$$

For example, for the standard q-ary alphabet, we have $P(s) = e^{-s} + e^{-s} + \cdots + e^{-s} = qe^{-s}$; and for Shannon's telegraphy alphabet (1.1), we have

$$P(s) = e^{-2s} + e^{-3s} + e^{-4s} + e^{-6s}. \tag{2.1}$$

Here is the main result about the combinatorial capacity of free languages.

2.1 Theorem. *The combinatorial capacity of the free language A^* is given by*

$$C_{\mathrm{comb}} = s_0 \text{ nats},$$

where s_0 is the unique solution of the equation $P(s) = 1$.

To prove Theorem 2.1, we study the convergence of the series

$$Q(s) = \sum_{k \geq 0} P(s)^k \tag{2.2}$$

in two ways. On the one hand, it is clear that $Q(s)$ converges if and only if $P(s) < 1$. Thus, since $P(s)$ is strictly decreasing from $P(0) = q \geq 1$ to $P(\infty) = 0$, it follows that $Q(s)$ converges for $s > s_0$ and diverges for $s < s_0$. In short, s_0 is the *abscissa of convergence* for $Q(s)$.

The second approach to studying the convergence of $Q(s)$ involves an alternative expression for $Q(s)$, which appears as a corollary to the following proposition.

2.2 Proposition. *If k is a positive integer, then*

$$P(s)^k = \sum_{t \in T} N_k(t) e^{-st}, \tag{2.3}$$

where $N_k(t)$ denotes the number of words from A^* of length k and duration t, and T is the set of possible durations for words from A^*.

Proof: We have

$$\begin{aligned} P(s)^k &= P(s) \cdot P(s) \cdots P(s) \\ &= \sum_{a_1 \in A} e^{-s\tau(a_1)} \cdot \sum_{a_2 \in A} e^{-s\tau(a_2)} \cdots \sum_{a_k \in A} e^{-s\tau(a_k)} \\ &= \sum_{a_1 a_2 \cdots a_k} e^{-s(\tau(a_1)+\tau(a_2)+\cdots+\tau(a_k))} \\ &= \sum_{a_1 a_2 \cdots a_k} e^{-s\tau(a_1 a_2 \cdots a_k)} \\ &= \sum_{t \in T} N_k(t) e^{-st}, \end{aligned}$$

as asserted. ▪

[4] A term borrowed from statistical mechanics, where a partition function is of the form $Z = \sum_r e^{-\beta E_r}$, with E_r being the energy associated with state r, and β is inversely proportional to the temperature [11].

2.3 Corollary. *If $Q(s)$ is as defined in (2.2), then*

$$Q(s) = \sum_{t \in T} N(t) e^{-st}, \tag{2.4}$$

where $N(t)$ is the number words from A^* of duration t.

Proof: Summing both sides of (2.3) for $k \geq 0$, we obtain

$$\begin{aligned} Q(s) = \sum_{k \geq 0} P(s)^k &= \sum_{k \geq 0} \sum_{t \in T} N_k(t) e^{-st} \\ &= \sum_{t \in T} e^{-st} \sum_{k \geq 0} N_k(t) \\ &= \sum_{t \in T} N(t) e^{-st}. \quad ▪ \end{aligned}$$

Equation (2.4) expresses $Q(s)$ as a discrete Laplace transform, or DLT. It follows from the convergence properties of DLTs (see Appendix A, Theorem A.2) that the abscissa of convergence of $Q(s)$ is

$$s_0 = \limsup_{t \in T} \frac{1}{t} \log N(t),$$

which by (1.2) is the combinatorial capacity of the A^*-channel. This completes the proof of Theorem 2.1. ▪

2.4 Example. Let's use Theorem 2.1 to find the combinatorial capacity of the free language over Shannon's telegraphy alphabet. As noted in (2.1), we have $P(s) = e^{-2s} + e^{-3s} + e^{-4s} + e^{-6s}$, and *Mathematica* tells us that the unique solution of the equation $P(s) = 1$ is $s_0 = 0.414012$, and so by Theorem 2.1, the capacity is 0.414012 nats (= 0.597294 bits) per second. ▪

2.5 Example. ([5], [8]) Suppose $S = \{\sigma_1, \sigma_2, \ldots\}$ is a uniquely decodable code over a q-letter alphabet, and that the length of σ_i is n_i. Think of S as the alphabet for a discrete noiseless channel. By Theorem 2.1, the combinatorial capacity C (in nats per second) of the free language over S satisfies

$$\sum_i e^{-Cn_i} = 1.$$

On the other hand, since the underlying channel uses the standard q-ary alphabet, $C \leq \log q$, so that

$$\sum_i e^{-(\log q)n_i} \leq 1,$$

i.e.,

$$\sum_i q^{-n_i} \leq 1,$$

which is McMillan's inequality (see e.g. [2], sec. 5.5.) ▪

3. The Probabilistic Capacity of Free Languages.

Suppose we attach a probability $p(a)$ to each letter $a \in A$, and then let X be a random variable with $\Pr\{X = a\} = p(a)$. The average amount of information provided by one observation of X is the entropy $H(X) = -\sum_{a \in A} p(a) \log p(a)$, and the average duration of X is $T(X) = \sum_{a \in A} p(a)\tau(a)$, so that the average information per unit time is $H' = H(X)/T(X)$. The probabilistic capacity of the A^*-channel is defined to be the maximum of this quantity over all choices of the probability density $p(a)$:

$$C_{\text{prob}} \triangleq \max_{p(a)} \frac{H(X)}{T(X)}.$$

It is satisfying that the probabilistic capacity is the same as the combinatorial capacity.

3.1 Theorem. *If A^* is the free language over the alphabet A, then*

$$C_{\text{prob}} = C_{\text{comb}},$$

i.e., $C_{\text{prob}} = s_0$, where s_0 is the unique solution to the equation $P(s) = 1$.

Proof: Since $P(s_0) = \sum_{a \in A} e^{-s_0 \tau(a)} = 1$, it is natural to define the probability density

$$p^*(a) \triangleq e^{-s_0 \tau(a)}.$$

Then if $p(a)$ is an arbitrary probability density on A, by Jensen's inequality

$$\sum_a p(a) \log \frac{p^*(a)}{p(a)} \leq \log \sum_a p^*(a) = \log 1 = 0,$$

with equality if and only if $p(a) = p^*(a)$ for all $a \in A$. Thus

$$H(X) = \sum_a p(a) \log \frac{1}{p(a)} \leq \sum_a p(a) \log \frac{1}{p^*(a)} = \sum_a p(a) s_0 \tau(a) = s_0 T(X),$$

with equality if and only if $p(a) = p^*(a)$ for all $a \in A$. This completes the proof of Theorem 3.1, and also identifies $p(a)^*$ as the optimizing probability density on A. ▪

3.2 Example. We saw in Example 2.4 that for the free language over Shannon's telegraph alphabet, $s_0 = 0.414012$. Thus the optimizing probabilities $p^*(a)$ for achieving the probabilistic capacity are

$$\begin{aligned} p^*(d) &= e^{-2s_0} = 0.43691 \\ p^*(D) &= e^{-4s_0} = 0.19089 \\ p^*(s) &= e^{-3s_0} = 0.28880 \\ p^*(S) &= e^{-6s_0} = 0.08340. \end{aligned}$$ ▪

4. The Combinatorial Capacity of Shannon Languages.

In this section, we will define the full class of languages to which Shannon's theorems on discrete noiseless channels apply, and we will give Shannon's characterization of the combinatorial capacity. Shannon languages are defined by labelled, directed graphs.

A *directed graph* (or digraph) G has vertex set $V = \{v_1, \ldots, v_M\}$ and branch set $B = \{b_1, \ldots, b_N\}$. Each branch $b \in B$ has an *initial vertex* $\text{init}(b) \in V$ and a *final vertex* $\text{fin}(b) \in V$. The set of branches with initial vertex v and final vertex w is denoted by $B_{v,w}$:

$$B_{v,w} \triangleq \{b \in B : \text{init}(b) = v, \text{fin}(b) = w.\} \tag{4.1}$$

Similarly, we define

$$B_{v*} \triangleq \{b \in B : \text{init}(b) = v\} \tag{4.2}$$

$$B_{*w} \triangleq \{b \in B : \text{fin}(b) = w\}. \tag{4.3}$$

For example, Figure 2 shows a digraph with $V = \{v_1, v_2\}$, and $B = \{b_1, \ldots, b_6\}$ described in Table 1 (ignore the entries marked $\lambda(b)$ and $\tau(b)$ for now):

e	$\text{init}(b)$	$\text{fin}(b)$	$\lambda(b)$	$\tau(b)$
b_1	v_1	v_2	d	2
b_2	v_1	v_2	D	4
b_3	v_2	v_2	d	2
b_4	v_2	v_2	D	4
b_5	v_2	v_1	s	3
b_6	v_2	v_1	S	6

Table 1. Tabular description of the digraph in Figures 1 and 2.

The sets $B_{v,w}$ are as follows:

$$B_{1,1} = \emptyset$$
$$B_{1,2} = \{b_1, b_2\}$$
$$B_{2,1} = \{b_5, b_6\}$$
$$B_{2,2} = \{b_3, b_4\}.$$

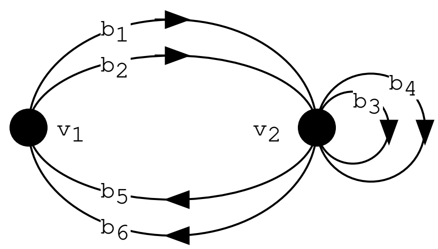

Figure 2. A digraph with two vertices and six branches.

A *path* P of length k from v to w in G is a sequence of k branches $P = b_1 b_2 \cdots b_k$ with $\mathrm{init}(b_1) = v$, $\mathrm{fin}(b_1) = \mathrm{init}(b_2)$, ..., $\mathrm{fin}(b_{k-1}) = \mathrm{init}(b_k)$, $\mathrm{fin}(b_k) = w$. We write $\mathrm{len}(P) = k$, $\mathrm{init}(P) = v$, and $\mathrm{fin}(P) = w$. For future reference, we denote the set of paths of length k from v to w by $B^k_{v,w}$:

$$B^k_{v,w} \triangleq \{P : \mathrm{len}(P) = k, \mathrm{init}(P) = v, \mathrm{fin}(P) = w\}.$$

The set of *all* paths from v to w will be denoted by $B^*_{v,w}$:

$$B^*_{v,w} \triangleq \bigcup_{k \geq 0} B^k_{v,w}. \tag{4.4}$$

Given a digraph G and an alphabet A, we can generate a language by labelling each branch of G with an element of A. Let $\lambda : B \to A$ be such a labelling. We will assume that this labelling is *right-resolving*, i.e., for each vertex v, the labels on all branches with $\mathrm{init}(b) = v$ are distinct [4, Section 3.3]. For example, Figure 1 shows a right-resolving labelling of the digraph from Figure 2, where the labels come from Shannon's telegraphy alphabet—see Table 1.[5]

The label of a path $P = b_1 b_2 \cdots b_k$ is defined to be the concatenation of its branch labels:

$$\lambda(P) = \lambda(b_1)\lambda(b_2)\cdots\lambda(b_k).$$

For each path P, the corresponding label $\lambda(P)$ is therefore a word over A. In what follows, we will use the term *path duration* to mean the duration of the corresponding word, and write $\tau(P)$ instead of $\tau(\lambda(P))$. Note that different paths can have the same label. For example, in Figure 2, the paths $P_1 = b_3 b_3 b_4$ and $P = b_1 b_3 b_4$ both have the label ddD. However, the "right-resolving" assumption guarantees that two paths *with the same initial vertex* and the same label must be identical.

The set of all possible path labels is called *the (Shannon) language generated by* G, and denoted by $\mathcal{L}_{G,\lambda}$. Our goal is to find the combinatorial capacity of $\mathcal{L}_{G,\lambda}$. To do this, we must suitably generalize the notion of partition function introduced in Section 2.

Let s be nonnegative real number, and for each pair of vertices (v, w), define the *branch duration partition function* as follows:

$$P_{v,w}(s) = \sum_{b \in B_{v,w}} e^{-s\tau(b)}. \tag{4.5}$$

[5] Our Figure 1 is essentially the same as Shannon's Figure 2.

The functions $P_{v,w}(s)$ can be thought of as entries in a $M \times M$ matrix $P(s)$. For example, for the labelled graph in Figure 1, the corresponding matrix is

$$P(s) = \begin{pmatrix} 0 & e^{-2s} + e^{-4s} \\ e^{-3s} + e^{-6s} & e^{-2s} + e^{-4s} \end{pmatrix}.$$

We denote the spectral radius (the magnitude of the largest eigenvalue) of the matrix $P(s)$ by $\rho(s)$. We call $\rho(s)$ the *partition function* for the language $\mathcal{L}_{G,\lambda}$. The function $\rho(s)$ has many interesting properties. For example, $\rho(s)$ is strictly decreasing, convex $\cup$, with $\lim_{s\to-\infty} \rho(s) = +\infty$, $\lim_{s\to+\infty} \rho(s) = 0$, with $\rho(0)$ bounded between the maximum and minimum out degree of any vertex of G. Furthermore, $\rho(s)$ is "log-convex," i.e., $\log \rho(s)$ is convex $\cup$. However, the only one of these properties we shall need in our proof is the fact that $\rho(s)$ is strictly decreasing. A proof of this appears in Appendix B, as Theorem B.5.

Here is the complete version of Shannon's Theorem 1 (Cf. Theorem 2.1).

4.1 Theorem (Shannon's Theorem 1). *The combinatorial capacity of the $\mathcal{L}_{G,\lambda}$-language is given by*

$$C_{\text{comb}} = s_0 \text{ nats},$$

where s_0 is the unique solution to the equation $\rho(s) = 1$. Alternatively, C_{comb} is the greatest positive solution of the equation $q(s) = \det(I - P(s)) = 0$.

Proof: Our proof of Theorem 4.1 depends on a calculation of the abscissa of convergence of the matrix series

$$Q(s) = \sum_{k\geq 0} P(s)^k \tag{4.6}$$

in two different ways.[6]

The first method for computing the a.o.c. of $Q(s)$ relies on certain known theorems in matrix theory (see Appendix B). By Theorem B.2, $Q(s)$ converges if and only if $\rho(s) < 1$. Next, by Theorem B.5, $\rho(s)$ is a strictly decreasing function of s. Thus if s_0 is the unique solution to the equation $\rho(s) = 1$, it follows that $\rho(s) < 1$ for all $s > s_0$, and $\rho(s) > 1$ for all $s < s_0$, so that s_0 is the r.o.c. for $Q(s)$. Finally, since by Theorem B.1, $\rho(s)$ is a eigenvalue of $P(s)$, we know that s_0 is the largest value of s for which $P(s)$ has 1 as an eigenvalue, i.e.,

$$s_0 = \sup\{s : \det(I - P(s)) = 0\}.$$

The second method for computing the a.o.c. for $Q(s)$ depends on an alternative expression for $Q(s)$ which appears as a corollary to the following theorem.

4.2 Theorem. *The (v, w)th entry in the matrix $P(s)^k$ is the DLT for durations of paths of length k from v to w, i.e.,*

$$[P(s)^k]_{v,w} = \sum_{t\in T} N^{(k)}_{v,w}(t) e^{-st}, \tag{4.7}$$

where $N^{(k)}_{v,w}(t)$ is the number of paths from v to w of length k and duration t, and T denotes the set of all possible path durations.

Proof: If we define the *weight* of a branch b by $W(b) = e^{-s\tau(b)}$, this follows immediately from from Theorem C.1, Appendix C. ▪

4.3 Corollary. *If $Q(s)$ is defined as in (4.6), then*

$$[Q(s)]_{v,w} = \sum_{t\in T} N_{v,w}(t) e^{-st}, \tag{4.8}$$

[6] The a.o.c. of a matrix $A(s)$ is defined to be the infimum of the values of s for which each $A_{v,w}(s)$ converges.

where $N_{v,w}(t)$ denotes the number of paths from v to w of duration t.

Proof: The proof is virtually identical to that of Corollary 2.3. ▪

Equation (4.8) expresses $[Q(s)]_{v,w}$ as a discrete Laplace transform. It follows from the convergence properties of DLTs (see Appendix A) that the abscissa of convergence of $[Q(s)]_{v,w}$ is

$$C_{v,w} \triangleq \limsup_{t \in T} \frac{1}{t} \log N_{v,w}(t),$$

so that the a.o.c. of $Q(s)$ itself is

$$s_0 = \max_{v,w} C_{v,w}.$$

Our proof of Theorem 4.1 will be complete if we can show that C as defined in eq. (1.2) equals $\max_{v,w} C_{v,w}$, i.e.,

$$\limsup_{t \in T} \frac{1}{t} \log N(t) = \max_{v,w} \left(\limsup_{t \in T} \frac{1}{t} \log N_{v,w}(t) \right),$$

where $N(t)$ denotes the total number of words in $\mathcal{L}_{G,\lambda}$ (= distinct path labels) of duration t. To accomplish this, we note that for all $t \geq 0$, we have

$$\max_{v,w} N_{v,w}(t) \leq N(t) \leq \sum_{v,w} N_{v,w}(t). \tag{4.9}$$

The left-hand inequality in (4.9) follows from the fact that the labelling λ is right-resolving, so that for any pair (v, w) the paths from v to w have distinct labels. The right-hand inequality in (4.9) follows from the fact that every word in $\mathcal{L}_{G,\lambda}$ corresponds to at least one path in G. Since there are M vertices in G, we can upperbound $\sum_{v,w} N_{v,w}(t)$ with $M^2 \cdot \max_{v,w} N_{v,w}(t)$, obtaining

$$\max_{v,w} N_{v,w}(t) \leq N(t) \leq M^2 \cdot \max_{v,w} N_{v,w}(t). \tag{4.10}$$

Taking logarithm of these inequalities, dividing by t, and passing to the "lim sup," we obtain

$$\begin{aligned} \max_{v,w} \left(\limsup_{t \in T} \frac{\log N_{v,w}(t)}{t} \right) \leq \limsup_{t \in T} \frac{\log N(t)}{t} &\leq \max_{v,w} \left(\limsup_{t \in T} \frac{\log M^2 + \log N_{v,w}(t)}{t} \right) \\ &= \max_{v,w} \left(\limsup_{t \in T} \frac{\log N_{v,w}(t)}{t} \right). \end{aligned}$$

Thus $C = \max_{v,w} C_{v,w}$, as required. ▪

4.4 Example. For the telegraphy language defined by Figure 1, the matrix $I_2 - P(s)$ is

$$I_2 - P(s) = \begin{pmatrix} 1 & -e^{-2s} - e^{-4s} \\ -e^{-3s} - e^{-6s} & 1 - e^{-2s} - e^{-4s} \end{pmatrix},$$

whose determinant is $q(s) = 1 - e^{-2s} - e^{-4s} - e^{-5s} - e^{-7s} - e^{-8s} - e^{-10s}$. The greatest positive root of $q(s) = 0$ (in fact, the only root) is $s_0 = 0.373562$, so that the capacity is $C = 0.373562$ nats $= 0.538936$ bits per second. This value should be compared to the capacity of the *free* language over the telegraphy alphabet, which as we saw above was 0.597294 bits per second. The structure imposed by Figure 1 has reduced the capacity by 0.058 bits per second, or about 10%, relative to the free language over the same alphabet. ▪

5. The Probabilistic Capacity of Shannon Languages.

Suppose we have a labelled digraph[7] (G, λ) as in Section 4, and we assign to each branch $b \in B$ a positive probability p_b, such that

$$\sum_{b \in B_{v*}} p_b = 1, \tag{5.1}$$

[7] In this section we will assume that G is strongly connected, i.e., there is at least one directed path between any two vertices. This will ensure that the resulting partition matrices are primitive and the Markov chains are ergodic.

for all vertices $v \in V$. (The set B_{v*} is defined in (4.2).) Then the p_b's describe a finite ergodic Markov chain, so there exists a unique stationary probability density on the vertices, $\{\pi_v : v \in V\}$, which satisfies

$$\sum_{b \in B_{*w}} \pi_{\mathrm{init}(b)} p_b = \pi_w, \tag{5.2}$$

for all $w \in V$. (The set B_{*w} is defined in (4.3).) The average amount of information provided by one observation of this Markov chain is the entropy

$$\begin{aligned} H &= -\sum_{v \in V} \pi_v \sum_{b \in B_{v*}} p_b \log p_b \\ &= -\sum_{b \in B} \pi_{\mathrm{init}(b)} p_b \log p_b, \end{aligned}$$

and the average duration of an observation is

$$\begin{aligned} T &= \sum_{v \in V} \pi_v \sum_{b \in B_{v*}} p_b \tau(b) \\ &= \sum_{b \in B} \pi_{\mathrm{init}(b)} p_b \tau(b). \end{aligned}$$

Thus the average information per unit time is $H' = H/T$. The probabilistic capacity of the language $\mathcal{L}_{G,\lambda}$ is defined to be maximum of this quantity, over all choices of the branch probabilities p_b:

$$C_{\mathrm{prob}} \triangleq \max_{(p_b)} \frac{H}{T}.$$

As was the case for free languages, the probabilistic capacity is the same as the combinatorial capacity.

5.1 Theorem (Shannon's Theorem 8). *The combinatorial and probabilistic capacites of the $\mathcal{L}_{G,\lambda}$-language are the same, i.e.,*

$$C_{\mathrm{prob}} = C_{\mathrm{comb}} = s_0,$$

where s_0 is the unique solution to the equation $\rho(s) = 1$.

Proof: Since $\rho(s_0) = 1$, by the Perron-Frobenius theorem [4, Sec. 4.2] the primitive matrix $P(s_0)$ has positive left and right "Perron" eigenvectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ associated with the eigenvalue 1 such that

$$\boldsymbol{\xi} P(s_0) = \boldsymbol{\xi} \tag{5.3}$$

$$P(s_0)\boldsymbol{\eta} = \boldsymbol{\eta} \tag{5.4}$$

$$\boldsymbol{\xi} \cdot \boldsymbol{\eta} = 1. \tag{5.5}$$

Let us define branch probabilities by the formula

$$p_b^* = \frac{\eta_{\mathrm{fin}(b)}}{\eta_{\mathrm{init}(b)}} e^{-s_0 \tau(b)}. \tag{5.6}$$

The stationary probability distribution for these branch probabilities turns out to be

$$\pi_v^* = \xi_v \eta_v, \ v \in V. \tag{5.7}$$

We pause to prove that indeed the p_b^*'s satisfy (5.1), and the π_v^*'s satisfy (5.2).

5.2 Lemma. $\sum_{b\in B_{v*}} p_b^* = 1$, *for all* $v \in V$.

Proof: We have

$$\begin{aligned}
\sum_{b\in B_{v*}} p_b^* &= \sum_{b\in B_{v*}} \frac{\eta_{\mathrm{fin}(b)}}{\eta_{\mathrm{init}(b)}} e^{-s_0\tau(b)} \qquad \text{(by (5.6))}\\
&= \frac{1}{\eta_v} \sum_{b\in B_{v*}} \eta_{\mathrm{fin}(b)} e^{-s_0\tau(b)}\\
&= \frac{1}{\eta_v} \sum_{w\in V} \eta_w \sum_{b\in B_{v,w}} e^{-s_0\tau(b)}\\
&= \frac{1}{\eta_v} \sum_{w} \eta_w P_{v,w}(s_0) \qquad \text{(by (4.5))}\\
&= \frac{1}{\eta_v}\cdot \eta_v \qquad \text{(by (5.4))}\\
&= 1. \qquad \blacksquare
\end{aligned}$$

5.3 Lemma. $\sum_{b\in B_{*w}} \pi^*_{\mathrm{init}(b)} p_b^* = \pi_w^*$, *for all* $w \in V$.

Proof: We have

$$\begin{aligned}
\sum_{b\in B_{*w}} \pi^*_{\mathrm{init}(b)} p_b^* &= \sum_{b\in B_{*w}} \xi_{\mathrm{init}(b)} \eta_{\mathrm{init}(b)} \frac{\eta_{\mathrm{fin}(b)}}{\eta_{\mathrm{init}(b)}} e^{-s_0\tau(b)} \qquad \text{(by (5.6) and (5.7))}\\
&= \eta_w \sum_{b\in B_{*w}} \xi_{\mathrm{init}(b)} e^{-s_0\tau(b)}\\
&= \eta_w \sum_{v} \xi_v \sum_{b\in B_{v,w}} e^{-s_0\tau(b)}\\
&= \eta_w \sum_{v} \xi_v P_{v,w}(s_0) \qquad \text{(by (4.5))}\\
&= \eta_w \xi_w \qquad \text{(by (5.3)}\\
&= \pi_w^*. \qquad \blacksquare
\end{aligned}$$

Now let (p_b) be an arbitrary set of branch probabilities and let (π_v) be the corresponding stationary vertex probabilities. Then by Jensen's inequality,

$$\sum_{v\in V} \pi_v \sum_{b\in B_{v*}} p_b \log \frac{p_b^*}{p_b} \le 0, \tag{5.8}$$

with equality if and only if $p_b = p_b^*$ for all $b \in B$. If we expand (5.8), we get

$$\begin{aligned}
H = \sum_{v\in V} \pi_v \sum_{b\in B_{v*}} p_b \log \frac{1}{p_b} &\le \sum_{v\in V} \pi_v \sum_{b\in B_{v*}} p_b \log \frac{1}{p_b^*}\\
&= \sum_{v\in V} \pi_v \sum_{b\in B_{v*}} p_b \left\{ s_0\tau(b) + \log \eta_{\mathrm{init}(b)} - \log \eta_{\mathrm{fin}(b)} \right\}\\
&= s_0 T + \sum_{v} \pi_v \sum_{b\in B_{v*}} p_b \log \eta_{\mathrm{init}(b)} - \sum_{v} \pi_v \sum_{b\in B_{v*}} p_b \log \eta_{\mathrm{fin}(b)}\\
&= s_0 T + S_1 - S_2. \qquad (5.9)
\end{aligned}$$

It is easy to see that S_1 and S_2 are equal. For on the one hand

$$\begin{aligned}
S_1 &= \sum_{v} \pi_v \sum_{b\in B_{v*}} p_b \log \eta_{\mathrm{init}(b)}\\
&= \sum_{v} \pi_v \log \eta_v \sum_{b\in B_{v*}} p_b\\
&= \sum_{v} \pi_v \log \eta_v \qquad \text{(by (5.1))},
\end{aligned}$$

while on the other hand

$$\begin{aligned} S_2 &= \sum_v \pi_v \sum_{b \in B_{v*}} p_b \log \eta_{\mathrm{fin}(b)} \\ &= \sum_{b \in B} \pi_{\mathrm{init}(b)} p_b \log \eta_{\mathrm{fin}(b)} \\ &= \sum_w \log \eta_w \sum_{b \in B_{*w}} \pi_{\mathrm{init}(b)} p_b \\ &= \sum_w \log \eta_w \ \pi_w \qquad \text{(by (5.2))}. \end{aligned}$$

Thus (5.9) becomes

$$H \leq s_0 T,$$

with equality if and only if $p_b = p_b^*$, for all $b \in B$. This completes the proof of Theorem 5.1, and furthermore identifies the branch probabilities (p_b^*) defined in (5.6) as the optimal ones. ▪

5.4 Example. We consider the application of Theorem 5.1 to Shannon's telegraphy language. We know from Example 4.4 that the capacity (combinatorial or probabilistic) is $s_0 = 0.373562$. A short calculation then shows that

$$P(s_0) = \begin{pmatrix} 0 & 0.698144 \\ 0.432369 & 0.698144 \end{pmatrix}.$$

The left and right Perron eigenvectors corresponding to the eigenvalue $\rho(s_0) = 1$ are then calculated to be (cf. (5.3)—(5.5))[8]

$$\boldsymbol{\xi} = (0.405048, 0.936811)$$
$$\boldsymbol{\eta} = (0.572441, 0.819946).$$

From these values, we can compute the optimal branch probabilities p_b^* as defined in (5.6), thus "completing" Table 1:

e	init(b)	fin(b)	$\lambda(b)$	$\tau(b)$	p_b^*
b_1	v_1	v_2	d	2	0.67856
b_2	v_1	v_2	D	4	0.32145
b_3	v_2	v_2	d	2	0.47373
b_4	v_2	v_2	D	4	0.22442
b_5	v_2	v_1	s	3	0.22764
b_6	v_2	v_1	S	6	0.07422

Table 2. The optimal branch probabilities for the labelled digraph of Figure 1.

Finally, we note that from (5.7), we have the stationary probabilities

$$\pi_1^* = 0.23187, \qquad \pi_2^* = 0.76813. \qquad ▪$$

6. Maximum Entropy Markov Chains on Digraphs.

In this section we briefly sketch a kind of generalization of Shannon's Theorem 8. (The corresponding result for free languages is stated as Problem 1.8 in [6].)

The setup being the same as in Section 5, we ask the following question: for a given value of $T > 0$, what is the maximum possible entropy $H_{\max}(T)$ among all Markov chains whose average branch duration is T?

[8] Of course $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are only determined up to a multiplicative constant.

The answer to a similar question was given by McEliece and Rodemich [7]. Here we summarize, and slightly extend, their answer.

If we define the branch duration matrix $P(s)$ as in (4.5), with corresponding spectral radius $\rho(s)$, then as the real variable s runs from $-\infty$ to $+\infty$, the $(H_{\max}(T), T)$ pairs are given parametrically by the equations

$$\begin{aligned} H(s) &= \log\rho(s) - s\frac{\rho'(s)}{\rho(s)} \\ T(s) &= -\frac{\rho'(s)}{\rho(s)}. \end{aligned} \tag{6.1}$$

In (6.1), as s runs from $-\infty$ to $+\infty$, T decreases monotonically from its maximum possible value $T_{\max}$ to its minimum possible value $T_{\min}$; $T_{\max}$ and $T_{\min}$ are the maximum and minimum average branch durations, respectively, among the cycles in G.

Furthermore, since $P(s)$ is primitve, $P(s)$ has positive left and right Perron eigenvectors $\boldsymbol{\xi}(s)$ and $\boldsymbol{\eta}(s)$ such that (cf. (5.3)–(5.5))

$$\begin{aligned} \boldsymbol{\xi}(s)P(s) &= \rho(s)\boldsymbol{\xi}(s) \\ P(s)\boldsymbol{\eta}(s) &= \rho(s)\boldsymbol{\eta}(s) \\ \boldsymbol{\xi}(s)\cdot\boldsymbol{\eta}(s) &= 1. \end{aligned}$$

The optimal branch probabilities and the corresponding stationary vertex probabilities are then given by the formulas

$$\begin{aligned} p_b(s)^* &= \frac{1}{\rho(s)}\frac{\eta_{\mathrm{fin}(b)}(s)}{\eta_{\mathrm{init}(b)}(s)}e^{-s\tau(b)} \\ \pi_v^*(s) &= \xi_v(s)\eta_v(s). \end{aligned}$$

As an application of this result, we can give a quick calculation of the probabilistic capacity, which is simply the maximum of the function $h(s) \triangleq H(s)/T(s)$. According to (6.1), we have

$$h(s) = s - \log\rho(s)\frac{\rho(s)}{\rho'(s)}, \tag{6.2}$$

so that

$$h'(s) = \log\rho(s)\left[\frac{\rho(s)\rho''(s) - \rho'(s)^2}{\rho'(s)^2}\right] \tag{6.3}$$

Since $\rho(s)$ is log-convex (proof omitted), the quantity $\rho(s)\rho''(s) - \rho'(s)^2$ is either identically zero (in which case $\tau(b)$ is the same for all branches), or strictly positive for all s. In the former case, $h(s)$ is independent of s, and the maximization problem is trivial. In the latter case, we have

$$\begin{aligned} \rho(s) > 1 &\Longrightarrow h'(s) > 0 \\ \rho(s) = 1 &\Longrightarrow h'(s) = 0 \\ \rho(s) < 1 &\Longrightarrow h'(s) < 0. \end{aligned}$$

Thus if s_0 is defined as the unique solution to the equation $\rho(s) = 1$, then by (6.2), we know that

$$h(s) = \frac{H(s)}{T(s)} \le s_0,$$

with equality if and only if $s = s_0$. This give an alternative proof of Theorem 5.1.

Finally we note that the maximum of $H(s)$ itself occurs at $s = 0$, and is $H_{\max} = \log\rho(0)$. This is because

$$H'(s) = -s\frac{\rho(s)\rho''(s) - \rho'(s)^2}{\rho(s)^2},$$

which is either identically 0, or is zero only at $s = 0$. If G is regular of degree r, $H_{\max} = \log r$. In the general case, $\rho(0)$ lies between the smallest and largest out-degree of any vertex.

6.1 Example. We briefly illustrate these results for the labelled digraph of Figure 1. In Figure 3, we have plotted the function $\rho(s)$, and in Figure 4 we have plotted $H(s)$ and $T(s)$. By eye we can estimate that the solution to $\rho(s) = 1$, i.e. the common value of C_{comb} and C_{prob}, is about 0.4 nats/sec; as we have seen (Example 4.4), the exact value is $s_0 = 0.373562$. In Figure 5, we have plotted the function $H_{\max}(T)$ vs. T using (6.1), and in Figure 6, we have plotted $H(s)/T(s)$. We see that $H_{\max}(T)$ is a convex $\cap$ function of T in the range $T \in [T_{\min}, T_{\max}] = [2.0, 5.0]$. (The self-loop b_2 has average branch duration $T_{\min} = 2$, and the loop (b_2, b_6) has average branch duration $T_{\max} = 5$.) The maximum of H occurs at $s = 0$, $T(0) = 3.4146$, and is $\log \rho(0) = \log(1 + \sqrt{5}) = 1.1744$. The maximum of $H(s)/T(s)$ occurs at $s = s_0$ and its value is also s_0. This example is entirely typical of the general case. ▪

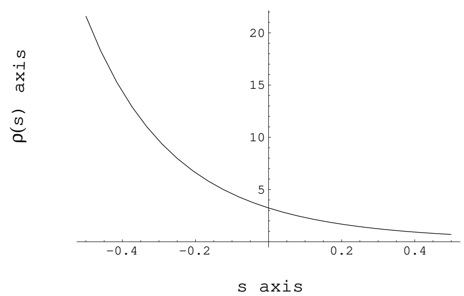

Figure 3. The function $\rho(s)$ for the labelled digraph of Figure 2.

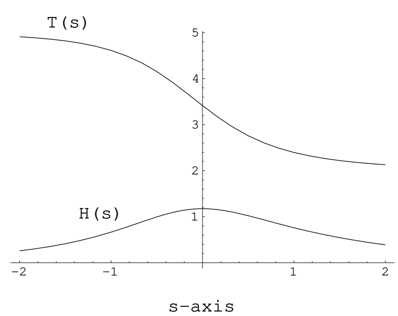

Figure 4. The functions $H(s)$ and $T(s)$.

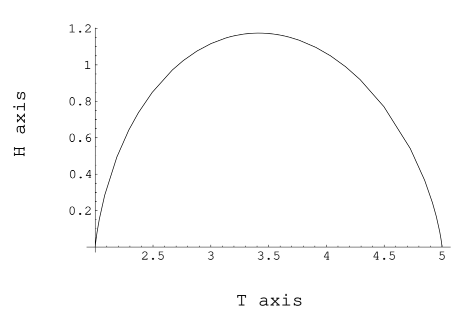

Figure 5. The function $H_{\max}(T)$.

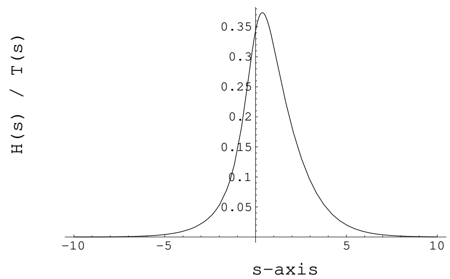

Figure 6. The function $H(s)/T(s)$.

7. Concluding Remarks.

Since (apart from Section 6) this paper contains no new results, the reader may wonder why the authors wrote this paper at all. In this section, we attempt to answer this reasonable question.

In the first place, we wished to reaffirm the continuing importance, after more than 50 years, of a relatively unsung part of Shannon's classic paper. (There is, for example, no mention of the discrete noiseless channel in the current standard information theory textbook [2].)

Second, and more important, we wished to provide clear and complete proofs of Shannon's Theorems 1 and 8. Shannon's own proof of Theorem 1, for example, seems not to apply to the case where the symbol durations are not all integers. His proof of Theorem 8, which appears in Appendix 4 in [9], is very brief and (for us) difficult to follow. The only information theory text that treats the discrete noiseless channel at all is [1, Chapter 2] which gives a complete proof of Theorem 1 in the case of equal symbol durations, and a sketch of a proof in the general case, but does not discuss Theorem 8 at all. The treatment of Theorem 1 in a recent book on symbolic dynamics [4, Theorem 4.3.3] is excellent, but considers only the case of equal symbol durations. Disappointingly, the same book has only a brief treatment of Theorem 8 in its chapter on "Advanced Topics" [4, Section 13.3]. In short, we are not aware of any published fully satisfactory treatment of the discrete noiseless channel, so this paper may be a first. Moreover, while we have followed Shannon's proof of Theorem 8 quite closely, we have found necessary to devise an entirely new method of proof for Theorem 1 (the partition function method), which we believe may be of independent interest.

Appendix A: Discrete Laplace Transforms.

Let $T = \{0 \le t_0 < t_1 < \cdots\}$ be an increasing, countable sequence of positive real numbers which is not too dense, in the sense that for any integer $n \ge 0$,

$$|\{t \in T : t < n\}| = O(n^K) \tag{A.1}$$

for some constant $K \ge 0$. For example, if $\tau_1, \ldots, \tau_q$ is a set of q positive real numbers, then the set of numbers of the form

$$t = n_1\tau_1 + \cdots + n_q\tau_q,$$

where $n_1, \ldots, n_q$ are arbitrary nonnegative integers, satisfies (A.1), with $K = q$.

A.1 Lemma. *If T satisfies (A.1), and if $0 \le \rho < 1$, then the series $\sum_{t \in T} \rho^t$ converges.*

Proof: If $0 \le \rho < 1$, we have

$$\begin{aligned}\sum_{t \in T} \rho^t &\le \sum_{t \in T} \rho^{\lfloor t \rfloor} \\ &= \sum_{n \ge 0} \rho^n \cdot |\{t \in T : n \le t < n+1\}| \\ &\le \sum_{n \ge 0} \rho^n \cdot |\{t \in T : t < n+1\}| \\ &= O\left(\sum_{n \ge 0} \rho^n (n+1)^K\right) \\ &< \infty. \quad \blacksquare\end{aligned}$$

If for each $t \in T$, $N(t)$ is a nonnegative real number, we define the *discrete Laplace transform* of $N(t)$ to be the series

$$f(s) = \sum_{t \in T} N(t) e^{-st}, \tag{A.2}$$

where $s \ge 0$ is a real number. The following theorem identifies the "abscissa of convergence" of $f(s)$.

A.2 Theorem. *Let*

$$s_0 = \limsup_{t \to \infty} \frac{1}{t} \log N(t).$$

Then $f(s)$ diverges for all $s < s_0$, and converges for all $s > s_0$.

Proof: We begin by rewriting the sum defining $f(s)$ as follows.

$$f(s) = \sum_{t \in T} e^{-t(s - \log N(t)/t)}. \tag{A.3}$$

If $s < s_0$, then by the definition of lim sup, $\log N(t)/t > s$ infinitely often. But if $\log N(t)/t > s$, the corresponding term in (A.3) is ≥ 1, so $f(s)$ cannot converge.

On the other hand, if $s > s_0$, there exists $\epsilon > 0$ so that $s - \epsilon > s_0$. Then, again by the definition of lim sup, we know that $\log N(t)/t < s - \epsilon$ for all but a finite number of t's. But if $\log N(t)/t < s - \epsilon$, the corresponding term in (A.3) is $\le e^{-t\epsilon} = \rho^t$, where $\rho = e^{-\epsilon} < 1$. Thus $f(s)$ converges, by Lemma A.1. $\blacksquare$

Appendix B: Theorems on Matrices.

In this appendix we state for reference several theorems from matrix theory that are needed in the body of the paper. All these theorems deal with the *spectral radius*, i.e., the largest eigenvalue of, a matrix. If A is a square matrix, its spectral radius is denoted by $\rho(s)$.

B.1 Theorem. [3, Thm. 5.6.]. *If $A \ge 0$, then $\rho(A)$ is an eigenvalue of A; in short, the largest eigenvalue of a nonegative matrix is real and nonnegative.*

B.2 Theorem. [3, Section 5.6.]. *If A is a square matrix, the geometric series*

$$\sum_{k \ge 0} A^k$$

converges if and only if $\rho(A) < 1$.$\blacksquare$

B.3 Theorem. [3, Corollary 8.1.19] *Let A and B be square matrices. If $0 \le A \le B$, then $\rho(A) \le \rho(B)$.*

B.4 Theorem. [3, Theorem 6.3.12] *If $\rho(s)$ denotes the spectral radius of a primitive matrix function $A(s)$, and if $\boldsymbol{\xi}(s)$ and $\boldsymbol{\eta}(s)$ are the corresponding left and right Perron eigenvectors, then*

$$\rho'(s) = \frac{\boldsymbol{\xi}(s) A'(s) \boldsymbol{\eta}(s)}{\boldsymbol{\xi}(s) \cdot \boldsymbol{\eta}(s)}. \tag{B.1}$$

B.5 Theorem. *The partition function $\rho(s)$ is strictly decreasing.*

Proof: The entries of the partition matrix $P(s)$ are either zero or of the form $\sum_i e^{-st_i}$, so that the derivatives of these entries are all nonpositive. On the other hand, the components of the Perron vectors $\boldsymbol{\xi}(s)$ and $\boldsymbol{\eta}(s)$ are strictly positive, so that by (B.1), $\rho'(s) < 0$. ▪

Appendix C: Path Weights in Digraphs.

In this appendix, we will prove a simple theorem of surprisingly wide applicability. We have quoted [10, Section 4.7], almost verbatim.

Let $G = (V, B)$ be a digraph, and let $W : B \to R$ be a *weight function* defined on the branches of G, taking values in some commutative ring R. (For most purposes, R will be the ring of real numbers or the ring of polynomials over the reals.) If $P = b_1 b_2 \cdots b_n$ is a path of length n, the weight of P is defined to be $W(P) = W(b_1)W(b_2)\cdots W(b_n)$. Now define

$$A_{v,w}(n) \triangleq \sum_{P \in B^n_{v,w}} W(P),$$

i.e., $A_{v,w}(n)$ is the sum of the weights of all paths of length n from v to w. The problem here is to find an efficient way to calculate the numbers $A_{v,w}(n)$. The key to the solution is to define the *branch weight matrix* $A = (A_{v,w})$ by

$$A_{v,w} \triangleq \sum_{b \in B_{v,w}} W(b).$$

C.1 Theorem. *Let n be a positive integer. Then the (v, w)th entry of A^n is $A_{v,w}(n)$.*

Proof: By a straightforward induction argument, the (v, w)th entry of the nth power of A is given by the formula

$$[A^n]_{v,w} = \sum A_{v,v_1} A_{v_1,v_2} \cdots A_{v_{n-1},w},$$

where the sum is over all possible sequences of vertices $(v_1, v_2, \ldots, v_{n-1})$. The summand is zero unless there is a path $b_1 b_2 \cdots b_n$ from v to w, with $\mathrm{init}(b_1) = v, \mathrm{fin}(b_1) = v_1$, $\mathrm{init}(b_2) = v_1, \mathrm{fin}(b_2) = v_2$, $\ldots$, $\mathrm{init}(b_n) = v_{n-1}, \mathrm{fin}(b_n) = w$. If one or more such paths exist, then the summand is the sum of the weights of all such paths. ▪

References.

1. Richard E. Blahut, *Principles and Practice of Information Theory.* Reading, Mass.: Addison-Wesley, 1987.
2. Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory.* New York: John Wiley & Sons, 1991.
3. Roger A. Horn and Charles R. Johnson, *Matrix Analysis.* Cambridge: Cambridge University Press, 1985.
4. Douglas Lind and Brian Marcus, *Symbolic Dynamics and Coding.* Cambridge: Cambridge University Press, 1995.
5. D. J. C. MacKay, personal communication.
6. Robert J. McEliece, *The Theory of Information and Coding.* Reading, Mass.: Addison-Wesley, 1977.
7. Robert J. McEliece and Eugene R. Rodemich, "A maximum entropy Markov chain," *Proc. 17th Conf. Inform. Sciences and Systems* (Johns Hopkins University, March 1983), pp. 245–248.
8. V. C. da Rocha Jr., "Some information-theoretic aspects of uniquely decodable codes," this volume.
9. C. E. Shannon, *The Mathematical Theory of Communication.* Urbana, Illinois: University of Illinois Press, 1963.
10. Richard P. Stanley, *Enumerative Combinatorics, vol. I.* Monterey, Calif.: Wadsworth & Brooks/Cole, 1986.
11. J. M. Yeomans, *Statistical Mechanics of Phase Transitions.* Oxford: Oxford University Press, 1992.