COLLOQUIUM MATHEMATICUM

VOL. LX/LXI DÉDIÉ À M. ANTONI ZYGMUND 1990

THE SHANNON INFORMATION ON A MARKOV CHAIN APPROXIMATELY NORMALLY DISTRIBUTED

BY

RICHARD O'NEIL (ALBANY, NEW YORK)

TO MY ESTEEMED TEACHER, ANTONI ZYGMUND

I. Stationary languages. An *alphabet* G is a finite set of symbols. A string $C = i_1 \dots i_k$ of length k is called a *k-gram*. Let G^k denote the set of k-grams. Clearly if G contains n symbols, $|G| = n$, then $|G^k| = |G|^k = n^k$.

If to each k-gram we assign a probability:

$$\mathrm{pr}_k(C) \geq 0, \quad \text{for each } C \in G^k, \qquad \sum_{C \in G^k} \mathrm{pr}_k(C) = 1,$$

and if the pr_k are consistent on the right and on the left:

$$\begin{aligned} &\sum_{i_{k+1}} \mathrm{pr}_{k+1}(i_1 \dots i_k i_{k+1}) = \mathrm{pr}_k(i_1 \dots i_k), \\ &\sum_{i_1} \mathrm{pr}_{k+1}(i_1 i_2 \dots i_{k+1}) = \mathrm{pr}_k(i_2 \dots i_{k+1}), \end{aligned} \tag{1}$$

then we say that the pr_k for $k = 1$ to ∞ define a *stationary language* over G.

In what follows we shall usually drop the subscript and write pr instead of pr_k.

Given a k-gram C the *Shannon information* of the k-gram is

$$I(C) = I_k(C) = \log 1/\mathrm{pr}(C),$$

provided $\mathrm{pr}(C) > 0$. The mean of I_k,

$$H_k = \sum_{C \in G^k} \mathrm{pr}(C) I(C),$$

is called the *Shannon entropy*.

It can be shown that the one step entropy defined as

$$H = \lim_{k \to \infty} H_k/k$$

exists for all stationary languages.

Shannon's theory of information is concerned with those stationary languages for which the random variable I_k/k congregates about its approximate mean, H, as k increases. More precisely,

(2) given $\varepsilon > 0$ there is a positive integer K such that for all $k \geq K$, $\sum \mathrm{pr}_k(C) < \varepsilon$, where the sum is taken over all k-grams, C, which do not satisfy $|I_k(C)/k - H| < \varepsilon$.

An $n \times n$ matrix $P = (p_{ij})$ such that $p_{ij} \geq 0$, for all i and j, and $\sum_j p_{ij} = 1$, for each i, is called a *stochastic matrix.*

A probability vector, $Q = (q_1, \ldots, q_n)$, is *stable* under P if $QP = Q$.

Given such a P and Q a stationary language is called a *Markov chain* if

$$\mathrm{pr}_k(i_1 \ldots i_k) = q_{i_1} p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{k-1} i_k}.$$

In this paper we shall be concerned with the information, I_k, on a Markov chain formed by an *aperiodic* stochastic matrix, that is, a stochastic matrix, P, some power of which has all its coefficients strictly positive.

In his famous paper of 1948 [4], Shannon proved (2) for such a language using the law of large numbers. We intend to sharpen his result first by computing the mean and variance of I_k and then by showing that, for large k, I_k is approximately normally distributed.

II. Matrices with positive coefficients. The Perron–Frobenius theorem for a square matrix, A, with strictly positive coefficients states that there exists a positive eigenvalue, λ, called the *principal eigenvalue*, of algebraic multiplicity one, and left and right eigenvectors L and R with strictly positive coefficients:

$$LA = \lambda L, \qquad AR = \lambda R.$$

(For a proof see [2], p. 285.)

We introduce the *oscillation* of A:

$$\omega = a^{**}/a^{*},$$

where a^* and a^{**} denote the minimum and maximum elements of A, $1 \leq \omega < \infty$. If A is $n \times n$ and ℓ^{**} the maximum element of the row vector $L = (\ell_1, \ldots, \ell_n)$ and ℓ^* the minimum, from $\lambda L = LA$ we obtain

$$\lambda \ell^{**} \leq a^{**} \sum \ell_i, \qquad \lambda \ell^{*} \geq a^{*} \sum \ell_i, \tag{3}$$

so that the oscillation of L,

$$\ell^{**}/\ell^{*} \leq a^{**}/a^{*} = \omega.$$

In what follows, to fix ideas we let $\lambda = 1$. Define the *L-norm* on column vectors by

$$\|X\|_L = \sum \ell_i |x_i|.$$

If $Y = AX$, $y_i = \sum_j a_{ij}x_j$, then

$$\begin{aligned}\|Y\|_L &= \sum_i \ell_i|y_i| \le \sum_i \ell_i \sum_j a_{ij}|x_j| = \sum_j |x_j| \sum_i \ell_i a_{ij} \\ &= \sum_j |x_j|\ell_j = \|X\|_L,\end{aligned}$$

so that $\|AX\|_L \le \|X\|_L$; moreover, in the special case that the coefficients of X are nonnegative, we have equality at each step above so that $\|AX\|_L = \|X\|_L$.

Let W_L be the subspace of all those column vectors such that $LX = 0$.

(4) THEOREM. *A is strictly contracting over W_L in the L-norm with contracting factor $\rho \le 1 - 1/\omega$.*

Proof. If $LX = 0$ then $L(AX) = LX = 0$ so that A maps W_L into W_L. Let $s = \min a_{ij}/\ell_j$. Then for $Y = AX$,

$$\begin{aligned}Y_i &= \sum_j a_{ij}x_j = \sum_j \left(\frac{a_{ij}}{\ell_j} - s\right)\ell_j x_j + \sum_j s\ell_j x_j \\ &= \sum_j \left(\frac{a_{ij}}{\ell_j} - s\right)\ell_j x_j, \\ \|Y\|_L &= \sum_i \ell_i|y_i| \le \sum_i \ell_i \sum_j \left(\frac{a_{ij}}{\ell_j} - s\right)\ell_j|x_j| \\ &= \sum_j \ell_j|x_j| \sum_i \ell_i \left(\frac{a_{ij}}{\ell_j} - s\right).\end{aligned}$$

But $\sum_i \ell_i(a_{ij}/\ell_j - s) = 1 - s\sum \ell_i = \rho$, so that $\|Y\|_L \le \rho\|X\|_L$. Clearly $\rho < 1$. By (3) and the definition of s,

$$s\sum \ell_i \ge \frac{a^*}{\ell^{**}}\sum \ell_i \ge \frac{a^*\sum \ell_i}{a^{**}\sum \ell_i} = \frac{1}{\omega},$$

i.e. $\rho \le 1 - 1/\omega$. ∎

Let X be any column vector with non-negative coefficients such that $\|X\|_L = \|R\|_L$. We wish to show that the sequence A^kX approaches R geometrically. We have

$$A^kX - R = A^k(X - R),$$

but $L(X - R) = LX - LR = \|X\|_L - \|R\|_L = 0$ so that $X - R \in W_L$. Therefore

$$\|A^kX - R\|_L \le \rho^k\|X - R\|_L \le \rho^k(\|X\|_L + \|R\|_L) = 2\rho^k\|R\|_L.$$

Now choose R and L so that

$$\|R\|_L = LR = 1.$$

(5) THEOREM. *The powers of A tend to the rank one matrix, RL, which is the product of the column vector R with the row vector L. Moreover, if $a_{ij}^{(k)}$ is the (i,j)th element of A^k then*

$$|a_{ij}^{(k)} - r_i\ell_j| \le 2\rho^{k-1}.$$

Proof. Let A_j denote the jth column of A. Then $\|A_j\|_L = LA_j = \ell_j$, the jth component of L. Thus for the jth column of A^k which is $A^{k-1}A_j$, we obtain

$$\|A^{k-1}A_j - \ell_j R\|_L \le 2\rho^{k-1}\|\ell_i R\|_L = 2\rho^{k-1}\ell_j,$$
$$|a_{ij}^{(k)} - \ell_j r_i|\ell_j \le \|A^{k-1}A_j - \ell_j R\|_L \le 2\rho^{k-1}\ell_j. \blacksquare$$

Apply this to square stochastic matrices with positive coefficients. In this case there is a unique row probability vector Q which is stable. We choose the right eigenvector to be the column vector $\mathbf{1}$ all of whose coefficients are 1. Then $QP = Q$, $P\mathbf{1} = \mathbf{1}$, and $Q\mathbf{1} = 1$. We find that

(6) P^k tends to the matrix $\mathbf{1}Q$ and the (i,j)th entry of P^k is within $2\rho^{k-1}$ of q_j.

Finally, we consider matrices $A(t)$ whose coefficients are positive analytic functions on the real line. The principal eigenvalue $\lambda(t)$ is a function. The eigenpolynomial,

$$\phi(t,x) = \det(xI - A(t)) = x^n + a(t)x^{n-1} + \dots,$$

is a polynomial with analytic coefficients and with $(\partial/\partial x)\phi(t,x)$ evaluated at $x = \lambda(t)$ different from zero (since $\lambda(t)$ is an eigenvalue of algebraic multiplicity one). By the implicit function theorem, $\lambda(t)$ is analytic with

$$\lambda'(t) = -\frac{\frac{\partial}{\partial t}\phi(t,x)}{\frac{\partial}{\partial x}\phi(t,x)} \quad \text{at } x = \lambda(t).$$

A principal right eigenvector, $R(t)$, is a non-trivial solution of the matrix equation

$$(A(t) - \lambda(t)I)X = 0$$

which is a system of n equations in n unknowns with analytic coefficients. But analytic functions belong to the field of meromorphic functions. Solving by Gauss–Jordan elimination yields a non-trivial solution $X = R(t)$ whose components are meromorphic functions which for each value of t are positive (and finite) and so must be analytic.

III. The mean and variance of information on a Markov chain. Consider a Markov chain whose stochastic matrix, $P = (p_{ij})$, has strictly positive elements. We first calculate and write in matrix form the mean of the random variable I_k on the k-grams, G^k:

$$\begin{aligned} H_k &= \sum_{C \in G^k} \mathrm{pr}(C) I_k(C) = \sum \mathrm{pr}(C) \log 1/P(C) \\ &= \sum_{i_1} \cdots \sum_{i_k} \mathrm{pr}(i_1 \ldots i_k)(\log 1/q_{i_1} + \log 1/p_{i_1 i_2} + \ldots + \log 1/p_{i_{k-1} i_k}) \end{aligned}$$

(we break the sum into k parts and on each part use the consistency relations (1) and then write the terms as matrix products)

$$\begin{aligned} &= \sum_i q_i \log 1/q_i + (k-1) \sum_i \sum_j q_i (p_{ij} \log 1/p_{ij}) \\ &= (Q \log 1/Q)\mathbf{1} + (k-1) Q (P \log 1/P)\mathbf{1} \end{aligned}$$

where $(Q \log 1/Q)$ is the row vector whose ith component is $q_i \log 1/q_i$, $(P \log 1/P)$ is the matrix whose (i,j)th component is $p_{ij} \log 1/p_{ij}$ and $\mathbf{1}$ is the column vector all of whose components are 1. We obtain the well known theorem of Shannon:

(7) If $H = Q(P \log 1/P)\mathbf{1}$ then $\lim_{k \to \infty} H_k/k = H$; moreover, there exists a constant B such that $|H_k/k - H| < B/k$.

In a similar way we compute the second moment V_k:

$$\begin{aligned} V_k &= \sum_{C \in G^k} \mathrm{pr}(C) I_k(C)^2 \\ &= \sum_{i_1} \cdots \sum_{i_k} \mathrm{pr}(i_1 \ldots i_k)(\log 1/q_{i_1} + \log 1/p_{i_1 i_2} + \ldots + \log 1/p_{i_{k-1} i_k})^2 \end{aligned}$$

(expand the square term into k^2 terms on which we use (1) whenever possible and then interpret the sums as matrix multiplications)

$$\begin{aligned} =& (Q \log^2 1/Q)\mathbf{1} + 2 \sum_{\ell=0}^{k-2} (Q \log 1/Q) P^\ell (P \log 1/P)\mathbf{1} \\ &+ 2 \sum_{\ell=0}^{k-3} (k - \ell - 2) Q (P \log 1/P) P^\ell (P \log 1/P)\mathbf{1} + (k-1) Q (P \log^2 1/P)\mathbf{1}. \end{aligned}$$

If we denote by S_k^2 the variance of the random variable I_k,

$$S_k^2 = \sum \mathrm{pr}(C)(I_k(C) - H_k)^2 = V_k - H_k^2$$

$$
\begin{aligned}
&= (Q\log^2 1/Q)1 - [(Q\log 1/Q)1]^2 \\
&\quad + 2\sum_{\ell=0}^{k-2}(Q\log 1/Q)[P^\ell - 1Q](P\log 1/P)1 \\
&\quad + 2\sum_{\ell=0}^{k-3}(k-\ell-2)Q(P\log 1/P)[P^\ell - 1Q](P\log 1/P)1 \\
&\quad + (k-1)\{Q(P\log^2 1/P)1 - [Q(P\log 1/P)1]^2\}.
\end{aligned}
$$

By (6) the difference of the (i,j)th component of P^ℓ and $1Q$ is less than $2\rho^{\ell-1}$, $\rho < 1$, so that the series

$$
E = \sum_{\ell=0}^{\infty}(Q\log 1/Q)[P^\ell - 1Q](P\log 1/P)1 \tag{8}
$$

is dominated by a geometric series and so converges absolutely. (For $\ell \geq 1$, $P^\ell - 1Q = (P - 1Q)^\ell$ so that using $I + \sum_{\ell=1}^{\infty}(P^\ell - 1Q) = (I - P + 1Q)^{-1}$ one could write E in a form not using infinite series.) The expression

$$
E_k = \frac{1}{k}\sum_{\ell=0}^{k-3}(k-\ell-2)(Q\log 1/Q)[P^\ell - 1Q](P\log 1/P)1
$$

is a modified $(C,1)$ sum of the series for E and so tends to E as k tends to ∞. Moreover, the terms of the series for E are dominated by a convergent geometric series so that it is easily shown that there is a constant, B, such that $|E_k - E| \leq B/k$. Thus we have proved:

(9) THEOREM. *Let*

$$
\begin{aligned}
S^2 = 2&\sum_{\ell=0}^{\infty}(Q\log 1/Q)[P^\ell - 1Q](P\log 1/P)1 \\
&+ Q(P\log^2 1/P)1 - [Q(P\log 1/P)1]^2.
\end{aligned}
$$

Then

$$
\lim_{k\to\infty} S_k^2/k = S^2;
$$

moreover, there is a constant, B, *such that*

$$
|S_k^2/k - S^2| < B/k.
$$

Chebyshev's inequality may be used to give a proof of Shannon's theorem (2).

The fact that H_k and S_k^2 are approximately equal to kH and kS^2 suggests an underlying central limit theorem which we now proceed to prove. A central limit theorem in a more general setting which holds for a certain class of strictly stationary strongly mixing sequences has been proved by

Ibragimov [2]. The interest of our result lies in the elementary natur; of the proof as well as in precise formulas such as those of Theorem (9).

IV. Information is an approximately normal random variable. We form the moment generating function (the Laplace transform) of the random variable I_k:

$$\begin{aligned}\phi_k(t) &= \sum_{C\in G^k} \mathrm{pr}(C)e^{tI_k(C)} \\ &= \sum \mathrm{pr}(C)\exp(t\log 1/\mathrm{pr}(C)) = \sum \mathrm{pr}(C)^{1-t} \\ &= \sum_{i_1}\cdots\sum_{i_k} q_{i_1}^{1-t}p_{i_1i_2}^{1-t}\cdots p_{i_{k-1}i_k}^{1-t}.\end{aligned}$$

Writing the multiple sum above in the form of matrix multiplication we obtain

$$\phi_k(t) = Q^{1-t}(P^{1-t})^{k-1}\mathbf{1}$$

where Q^{1-t} is the row vector whose ith component is q_i^{1-t}, P^{1-t} the matrix whose (i,j)th component is P_{ij}^{1-t} and $\mathbf{1}$ the column vector all of whose components are 1.

P^{1-t} is a matrix with positive analytic coefficients; let $\lambda(t)$ be its principal eigenvalue. $\lambda(t)$ is an analytic function for each t, in particular it is analytic at $t = 0$, therefore

$$\lambda(t) = 1 + at + \frac{1}{2}bt^2 + O(t^3) \quad \text{as } t \to 0.$$

Let $L(t)$ be a left row eigenvector and $R(t)$ a right column eigenvector for the matrix P^{1-t} with eigenvalue $\lambda(t)$. In particular, we can choose $L(t)$ and $R(t)$ with positive analytic coefficients and such that $L(t)R(t) = 1$. Observe $L(0) = Q$ and $R(0) = \mathbf{1}$. We have

$$\frac{\phi_k(t)}{\lambda(t)^{k-1}} = Q^{1-t}\left[\left(\frac{P^{1-t}}{\lambda(t)}\right)^{k-1} - R(t)L(t)\right]\mathbf{1} + Q^{1-t}R(t)L(t)\mathbf{1}.$$

Let $h(t) = Q^{1-t}R(t)L(t)\mathbf{1}$. Then $h(t)$ is an analytic function so that in a neighborhood of $t = 0$

$$h(t) = 1 + O(t) \quad \text{as } t \to \infty.$$

If ω, the ratio of the largest element of $P = (p_{ij})$ to the smallest, is the oscillation of P, then the oscillation of P^{1-t} is ω^{1-t} so that in a neighborhood of $t = 0$, say $|t| \leq 1$, the oscillation of P^{1-t} is uniformly bounded by ω^2. Thus for $|t| \leq 1$ the matrix $P(t)/\lambda(t)$ tends uniformly to the matrix $R(t)L(t)$ and indeed the individual terms of the matrix difference $(P^{1-t}/\lambda(t))^{k-1} -$

$R(t)L(t)$ are dominated by $2\rho^{k-2}$ where $\rho \leq 1 - 1/\omega^2$. Thus for $|t| \leq 1$,

$$\frac{\phi_k(t)}{\lambda(t)^{k-1}} - h(t) \to 0 \quad \text{uniformly as } k \to \infty.$$

Multiply the numerator and denominator of the above fraction by e^{-kat} and replace t by $t/\sqrt{k}$. Since given any large number δ eventually $k > \delta^2$ we find that given $\delta > 0$, for $|t| \leq \delta$

$$\frac{e^{-kat/\sqrt{k}}\phi_k(t/\sqrt{k})}{[e^{-at/\sqrt{k}}\lambda(t/\sqrt{k})]^{k-1}e^{-at/\sqrt{k}}} - h(t/\sqrt{k}) \to 0 \tag{10}$$

uniformly as k tends to ∞. But the numerator of the above fraction,

$$\sigma_k(t) = e^{-\sqrt{k}at}\phi_k(t/\sqrt{k}),$$

is easily recognized to be the moment generating function of the random variable which to a k-gram C assigns the value $(I_k(C) - ka)/\sqrt{k}$. But $h(t/\sqrt{k})$ and $e^{-at/\sqrt{k}}$ tend uniformly to 1 for $|t| \leq \delta$ and

$$e^{-at}\lambda(t) = 1 + \frac{1}{2}(b - a^2)t^2 + O(t^3)$$

so that from (10) it follows that

$$\begin{aligned}\lim_{k\to\infty}(e^{-at/\sqrt{k}}\lambda(t/\sqrt{k}))^{k-1} &= \lim\left(1 + \frac{1}{2}\frac{b^2 - a^2}{k}t^2 + O(t^3/k^{3/2})\right)^{k-1} \\ &= e^{(b-a^2)t^2/2}.\end{aligned}$$

Thus $\sigma_k(t)$ tends uniformly for $|t| \leq \delta$ to $e^{(b-a^2)t^2/2}$, which we recognize as the moment generating function of a normal distribution of mean 0 and variance $b - a^2$.

Combining the results of this section with those of Section III we see that

$$\phi_k(t) = 1 + H_k t + \frac{1}{2}V_k t^2 + \text{ higher order terms}$$

so that equating coefficients in the limit we know precisely the first terms of $\lambda(t)$,

$$\lambda(t) = 1 + Ht + \frac{1}{2}(S^2 + H^2) + \text{ higher order terms.}$$

Thus using a well known theorem in probability theory (see [1], Theorem 6.2.24, p. 309) we have proved:

(11) THEOREM. *The distribution function of the random variable*

$$(I_k - kH)/\sqrt{k}S$$

over a Markov chain whose transition matrix has positive entries tends to the distribution function of a normal distribution of mean 0 *and variance* 1.

Actually a bit more is true: If A is a matrix with non-negative coefficients such that some power A^m has positive coefficients and if ρ is the contracting factor of A^m acting on W_L, then letting $\sigma = \rho^{1/m}$ and $B = 2/\rho$ the reader will easily verify that the conclusion of Theorem (5) goes through except that the final formula is replaced by

$$|a_{ij}^{(k)} - r_i\ell_i| \leq B\sigma^k.$$

The above theory now goes through for all aperiodic Markov chains *mutatis mutandis*.

We close with a simple example of stationary language for which the theory fails. Let $G = \{1, 2\}$. If the k-gram C does not consist entirely of a string of 1's we let $\mathrm{pr}_k(C) = 1/2^{k+1}$, and $\mathrm{pr}_k(11\ldots1) = 1/2 + 1/2^{k+1}$. Then $I_k(11\ldots1) \approx \log 2$, while for the other k-grams, C, $I_k(C) = (k+1)\log 2$. Thus

$$\frac{I_k(11\ldots1)}{k} \approx 0, \qquad \mathrm{pr}_k(11\ldots1) \approx \frac{1}{2}$$

while for the other k-grams

$$\frac{I_k(C)}{k} \approx \log 2, \qquad \sum \mathrm{pr}(C) \approx \frac{1}{2}.$$

Property (2) fails and indeed in this case we easily verify that S_k^2 is $O(k^2)$ rather than $O(k)$.

Acknowledgments: To Dr. Louis Brickman for helpful suggestions and to Dr. Alexander Volberg for informing me about reference [2].

REFERENCES

[1] E. J. Dudewicz and S. N. Mishra, *Modern Mathematical Statistics*, Wiley, 1988.

[2] I. A. Ibragimov, *Some limit theorems for stationary processes*, Theor. Probab. Appl. 7 (1962), 349–382.

[3] P. Lancaster, *Theory of Matrices*, Academic Press, 1969.

[4] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J., July 1948, p. 379 ff.

DEPARTMENT OF MATHEMATICS AND STATISTICS
THE UNIVERSITY AT ALBANY
ALBANY, NEW YORK 12222, U.S.A.

Reçu par la Rédaction le 17.4.1990