Capacity of a Burst-Noise Channel

By E. N. GILBERT

(Manuscript received March 15, 1960)

A model of a burst-noise binary channel uses a Markov chain with two states G and B. In state G, transmission is error-free. In state B, the channel has only probability h of transmitting a digit correctly. For suitably small values of the probabilities, p, P of the B → G and G → B transitions, the model simulates burst-noise channels. Probability formulas relate the parameters p, P, h to easily measured statistics and provide run distributions for comparison with experimental measurements. The capacity C of the model channel exceeds the capacity C(sym. bin.) of a memoryless symmetric binary channel with the same error probability. However, the difference is slight for some values of h,p,P; then, time-division encoding schemes may be fairly efficient.

I. INTRODUCTION

In information theory the symmetric binary channel is the classical model of a noisy binary channel. This channel generates a sequence of binary noise digits z_n, which it adds (modulo 2) to input digits x_n to produce output digits $y_n = x_n + z_n$. The symmetric binary channel is memoryless; a sequence of independent trials produces the noise digits z_n. Each trial has the same probability $P(1)$ of producing an error and probability $1 - P(1) = P(0)$ of no error. The capacity C(sym. bin.) of this channel is well known (see Shannon[1]):

$$C(\text{sym. bin.}) = 1 + P(0) \log_2 P(0) + P(1) \log_2 P(1).$$

Channels with memory occur in practice. If radio static or switching transients produce the noise, the errors group into isolated bursts (several errors close together). Independent trials fail to simulate such a burst-noise. Section II of this paper presents a model of a burst-noise channel that is simple enough to permit calculation of the channel capacity C (see Sections III and VI). Sections IV and V give run distributions, the covariance function and other probability formulas as aids to testing the model's applicability and to picking model parameters which match measured statistical data.

Of all binary channels with a given error probability $P(1)$, the symmetric binary channel has least capacity. Indeed, if an encoding for signaling over the symmetric binary channel at a rate R is known, then N sources can use this encoding in time-division multiplex at rates R/N, each over a burst-noise channel. Here, N must be large enough so that noise digits N apart are nearly independent. Time division protects against other noise patterns besides bursts; still less redundant schemes are possible. The possible increase in signaling rate $C - C(\text{sym. bin.})$ will be seen to be often surprisingly small (see Fig. 4).

II. THE MODEL

A Markov chain with two states can be used to generate bursts. The two states will be called G (for good) and B (for bad or for burst). In state G the noise digit is always $z_n = 0$. In state B a coin is tossed to decide whether z_n will be 0 or 1.

The coin-tossing feature is included because actual bursts contain good digits interspersed with the errors. In the formulas that follow a biased coin is allowed (probability h of making no error in state B). All computations given here take $h = 0.50$, which seems a reasonable value.

After producing the noise digit z_n, the Markov chain makes a transition to prepare for z_{n+1}. To simulate burst noise, the states B and G must tend to persist; i.e., the transition probabilities $P = \text{Prob}(G \rightarrow B)$ and $p = \text{Prob}(B \rightarrow G)$ will be small and the probabilities $Q = 1 - P$, $q = 1 - p$ of remaining in G and B will be large. Fig. 1 is a transition diagram for the Markov chain.

Runs of G will alternate with runs of B. The run lengths have geometric distributions with mean $1/P$ for the G-runs and mean $1/p$ for

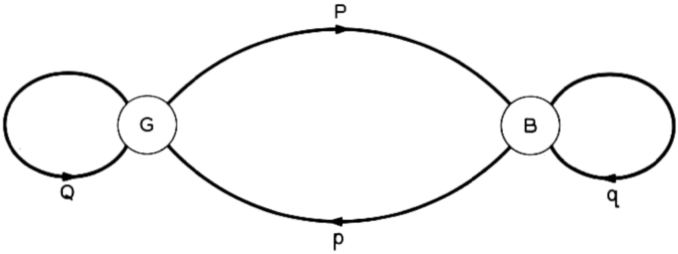

Fig. 1 — Transition diagram for the Markov chain.

the B-runs. The geometric distribution of G-runs seems reasonable. If the various clicks, pops and crashes, which might cause errors on a real channel, are not related to one another, then the times between such events will have the geometric distribution (see Feller,[2] Section XIII.9). Only mathematical simplicity justifies the geometric distribution of B-runs; one might construct more accurate models. Section III mentions one way of elaborating this one; however, complicated models may be useless without adequate statistical data to determine all the model parameters. Section V will illustrate some of the difficulties in determining just the three parameters P, p and h.

The following 500 digits form a typical sample of burst-noise with parameters $P = 0.03$, $p = 0.25$, $h = 0.5$, produced by using random numbers:

$$0^{62}110^{17}10^{46}110101110^{11}110^{15}10^{42}10^{28}110^{90}10^{37}$$

$$110^{5}10010^{35}1011010^{23}110^{4}10^{18}10^{15}11011101101110^{5}.$$

The exponents are run lengths; i.e., 0^{62} denotes a run of 62 consecutive zeros. As expected, long runs of good digits separate the bursts.

The 500-digit sample illustrates the impossibility of reconstructing the sequence of states from the sequence of digits. In portions of some of the long runs of zeros, the Markov chain was in state B; this went unnoticed because the coin tosses produced only zeros. The sample also contains one burst $110^{4}1$ in which a short sojourn into state G produced three of the four zeros.

The fraction of time spent in state B is $P(\mathrm{B}) = P/(p + P)$. Since errors occur only in state B, and then just with probability $1 - h$, the error probability is

$$P(1) = (1 - h)P\ (\mathrm{B}) = (1 - h)\,\frac{P}{p + P}. \qquad (1)$$

III. THE CAPACITY

Let H denote the entropy of the sequence of noise digits $\cdots, z_1, z_2, \cdots$. For all inputs x to the burst-noise channel, the conditional entropy, $H_x(y)$, of the output y knowing the input x is the same:

$$H_x(y) = H.$$

A simple argument then shows that the capacity C of the burst-noise channel is $C = 1 - H$ (a monogram source with probabilities 0.5 for 0 and 0.5 for 1 attains the rate C).

Shannon[1] (Section 7) gives a simple way of computing an entropy H from state probabilities [$P(\mathrm{G})$, $P(\mathrm{B})$ here] and transition probabilities. McMillan[3] (Section 2.0) notes that this result tacitly assumes that the state sequence is reconstructible from the digit sequence. Since a reconstruction is impossible here, H has a more complicated formula.

A definition of H is

$$H = \lim_{N\to\infty} \sum_{z_i=0,1} P(z_1, \cdots, z_N) h(z_1, \cdots, z_N), \tag{2}$$

with

$$h(z_1, \cdots, z_N) = -\sum_{z_{N+1}=0}^{1} P(z_{N+1} \mid z_1, \cdots, z_N) \log_2 P(z_{N+1} \mid z_1, \cdots, z_N). \tag{3}$$

If $z_i = 1$, the corresponding state is certainly B and

$$P(z_{i+1}, \cdots, z_{i+j} \mid z_1, \cdots, z_{i-1}, 1) = P(z_{i+1}, \cdots, z_{i+j} \mid 1) \tag{4}$$

follows for all $j \geqq 1$. Then,

$$P(z_{N+1} \mid z_1, \cdots, z_{i-1}, 1, z_{i+1}, \cdots, z_N) = P(z_{N+1} \mid 1, z_{i+1}, \cdots, z_N)$$

follows and also

$$h(z_1, \cdots, z_{i-1}, 1, z_{i+1}, \cdots, z_N) = h(1, z_{i+1}, \cdots, z_N).$$

Thus, just the number of consecutive zeros at the end of the block $(z_1, \cdots, z_N)$ determine $h(z_1, \cdots, z_N)$ completely. Each of the $2^N h$'s in the sum (2) is one of the $N + 1$ numbers

$$h(1), h(10), \cdots, h(10^k), \cdots, h(10^{N-1}), h(0^N)$$

(again exponents denote run lengths). After using this simplification in (2), summing and letting $N \to \infty$, the result is

$$H = \sum_{K=0}^{\infty} P(10^K) h(10^K). \tag{5}$$

The terms of (5) involve probabilities of runs of zeros. Section IV will give a formula for the conditional probability, $u(K)$, of a run of K or more zeros following a one, that is, $u(K) = P(0^K \mid 1)$. The convention $u(0) = 1$ will be adopted. Then, in (5),

$$P(10^K) = P(1) u(K)$$

[(1) gives $P(1)$]. Also, (3), together with $P(0 \mid 10^K) = u(K + 1)/u(K)$, provides an expression for $h(10^K)$:

$$h(10^K) = -\frac{u(K+1)}{u(K)}\log_2\frac{u(K+1)}{u(K)} - \left[1-\frac{u(K+1)}{u(K)}\right]\log_2\left[1-\frac{u(K+1)}{u(K)}\right]. \tag{6}$$

Using (6), the terms of (5) rearrange into

$$C = 1 + P(1)\sum_{K=0}^{\infty} v(K)\log_2 v(K), \tag{7}$$

where $v(K) = u(K) - u(K+1)$. Section IV contains formulas for $v(K)$. Although (7) seems simpler than (5) and (6), it converges slowly. In Section V the computation method uses a modification of (5) and (6).

Note that $v(K) = P(0^K 1 \mid 1)$. Another derivation of (7) proceeds by showing that the noise sequence consists of successive blocks of digits of the form $1, 01, 001, \cdots, 0^K 1, \cdots$, chosen independently, and with probability $v(K)$ for the block $0^K 1$. Then $-\sum v(K)\log_2 v(K)$ is the information per block and $P(1)$ is the average number of blocks per digit.

Equations (5), (6) and (7) apply to certain other channels. These formulas followed just from (4), which holds whenever the lengths of successive runs of zero are independent. Whenever such independence can be assumed, a more elaborate model might use $v(0), v(1), v(2), \cdots$, directly as parameters. Then $P(1)$ in (7) is

$$P(1) = \left[\sum_{K=0}^{\infty}(K+1)v(K)\right]^{-1}.$$

As a check, the symmetric binary channel has $v(K) = P(1)[P(0)]^K$ and (7) sums to C(sym. bin.).

IV. PROBABILITIES

Recurrent events theory (Feller,[2] Section XIII) provides some probabilities needed in Sections V and VI.

4.1 *Recurrence Times for State B*

Let f_K denote the conditional probability, in state B, that the first return to B will happen at step K:

$$f_K = P(\mathrm{G}^{K-1}\mathrm{B} \mid \mathrm{B}).$$

Then $f_1 = q$, $f_2 = pP$ and $f_K = pQ^{K-2}P$ for $K \geqq 2$. It is convenient to make these probabilities the coefficients of a generating function $F(t)$ of recurrence time probabilities:

$$F(t) = \sum f_K t^K = qt + \frac{pPt^2}{1 - Qt}. \tag{8}$$

For example, the probability $f_K^{(m)}$ that the mth return to B happens at step K has the generating function

$$\sum_{K=1}^{\infty} f_K^{(m)} t^K = [F(t)]^m. \tag{9}$$

The probability of no return to B in k steps is pQ^{k-1}. Then the probability $s(K,m)$ of exactly m returns to B in K steps (but not necessarily a return on step K) is

$$s(K,m) = f_K^{(m)} + \sum_{K=1}^{K-m} f_{K-k}^{(m)} pQ^{k-1}.$$

The corresponding generating function is

$$\sum_{K=1}^{\infty} s(K,m)t^K = \left(1 + \frac{pt}{1 - Qt}\right)[F(t)]^m. \tag{10}$$

4.2. *Recurrence Times for Ones*

Starting from a one (and hence from B), the next one must occur at a return to B, but not necessarily the first return. The probability that the next one occurs at the mth return to B and at step K is

$$h^{m-1}(1 - h)f_K^{(m)}.$$

Then, recurrence time probabilities for ones are

$$v(K - 1) = P(0^{K-1}1 \mid 1) = \sum_{m=1}^{\infty} h^{m-1}(1 - h)f_K^{(m)}.$$

Equation (9) now provides the generating function $V(t) = \sum v(K)t^K$:

$$tV(t) = \frac{(1 - h)F(t)}{1 - hF(t)}. \tag{11}$$

Likewise, the probability $u(K)$ that no one appears in the next K steps is

$$u(K) = \sum_m s(K,m)h^m,$$

which has generating function

$$U(t) = \frac{1 + (p - Q)t}{(1 - Qt)[1 - hF(t)]}. \tag{12}$$

By (8),

$$U(t) = \frac{1 + (p - Q)t}{D(t)}, \tag{13}$$

where $D(t) = 1 - (Q + hq)t - h(p - Q)t^2$.

Factor the quadratic $D(t)$:

$$D(t) = (1 - Jt)(1 - Lt),$$

where $2J = Q + hq + \sqrt{(Q + hq)^2 + 4h(p - Q)}$ and L is the same expression with negative square root. Now, (13) becomes

$$U(t) = \frac{1 + (p - Q)t}{J - L}\left(\frac{J}{1 - Jt} - \frac{L}{1 - Lt}\right).$$

The coefficient of t^K in the power series for $U(t)$ is

$$u(K) = \frac{(J + p - Q)J^K - (L + p - Q)L^K}{J - L}. \tag{14}$$

To find a recurrence formula for $u(K)$, write (13) as $D(t)U(t) = 1 + (p - Q)t$ and equate coefficients of t^K:

$$u(K) = (Q + hq)u(K - 1) + h(p - Q)u(K - 2) \tag{15}$$

for $K = 2,3,\cdots$. Initial values are

$$u(0) = 1, \qquad u(1) = p + hq.$$

For calculating, (15) is more convenient than (14).

Similar steps lead from (11) to

$$v(K) = \frac{1 - h}{J - L}[(qJ + p - Q)J^K - (qL + p - Q)L^K]. \tag{16}$$

For $K = 2,3,\cdots$, $v(K)$ also satisfies (15), but with initial values

$$v(0) = (1 - h)q, \qquad v(1) = (1 - \mathrm{h})(pP + hq^2).$$

4.3. *Covariance*

The covariance function of this binary noise is just a joint probability $r(K) = \mathrm{Prob}(z_0 = 1, z_K = 1)$. A formula for the generating function

$R(t) = \sum r(K)t^K$ is

$$\begin{aligned} R(t) &= P(1)\,\{1 + tV(t) + [tV(t)]^2 + \cdots\} \\ &= \frac{P(1)}{1 - tV(t)} \\ &= \frac{P(1)D(t)}{(1 - t)[1 + (p - Q)t]}. \end{aligned}$$

The term $P(1)[tV(t)]^m$ in the sum generates the probabilities of finding $z_0 = z_K = 1$, with exactly $m - 1$ of the digits $z_1, \cdots, z_{K-1}$ equal to 1.

An explicit formula for $r(K)$ follows by expanding $R(t)$ in a power series:

$$\begin{aligned} r(0) &= P(1), \\ r(K) &= P(1)^2\left[1 + \frac{p(q - P)^K}{P}\right], \qquad K = 1,2,\cdots. \end{aligned} \tag{17}$$

V. PARAMETER MATCHING

The three parameters p, P, h are not directly observable, so methods of deducing them from statistical measurements must now be considered. We will express p, P, h as functions of three other easily estimated noise parameters. One suitable set of three parameters (involving only trigram statistics) is

$$a = P(1), \qquad b = P(1 \mid 1), \qquad c = \frac{P(111)}{P(101) + P(111)}.$$

Here, c is the conditional probability of finding the place between two ones filled by a one, and it has the expression

$$\frac{(1 - h)q^2}{q^2 + pP}.$$

Solving for p, P, h in terms of a, b, c,

$$\begin{aligned} 1 - p = q &= \frac{ac - b^2}{2ac - b(a + c)}, \\ h &= 1 - \frac{b}{q}, \\ P &= \frac{ap}{1 - h - a}. \end{aligned} \tag{18}$$

If $h = 0.5$ is assumed, then $q = 2b$ and no c measurement is needed.

For illustration, the 500-digit sample in Section II contains thirty-eight 1's, fifteen 11's, seven 101's, and three 111's. Estimates of a, b, c are $a = 38/500$, $b = 15/38$, $c = 3/10$. With these estimates, (18) gives ridiculous parameters (p is negative). The trouble is that 500 digits provide too small a sample. In particular, the estimate $c = 3/10$, based on only 10 observations, is far from the correct value $c = 0.49$. If $h = 0.50$ is assumed, the estimates become $p = 0.21$, $P = 0.036$ (compare with true values $p = 0.25$, $P = 0.03$).

After finding p, P, and h, the results of Section IV suggest comparisons between run measurements and the probabilities $u(K)$ or $r(K)$. Fig. 2 shows curves of some run probabilities $P(10^K) = P(1)u(K)$ (on a log scale) versus K. As shown by (14), these curves straighten out for large K with slopes determined by J.

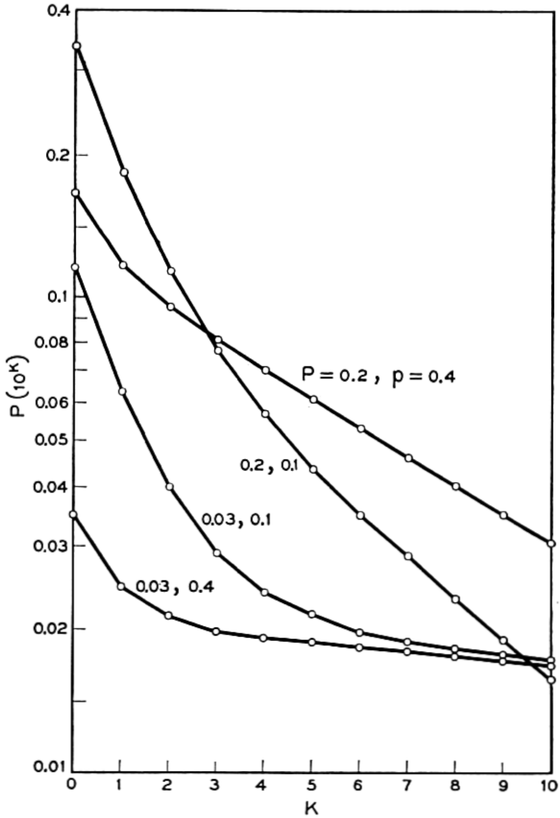

Fig. 2 — Typical run distributions, with $h = \frac{1}{2}$.

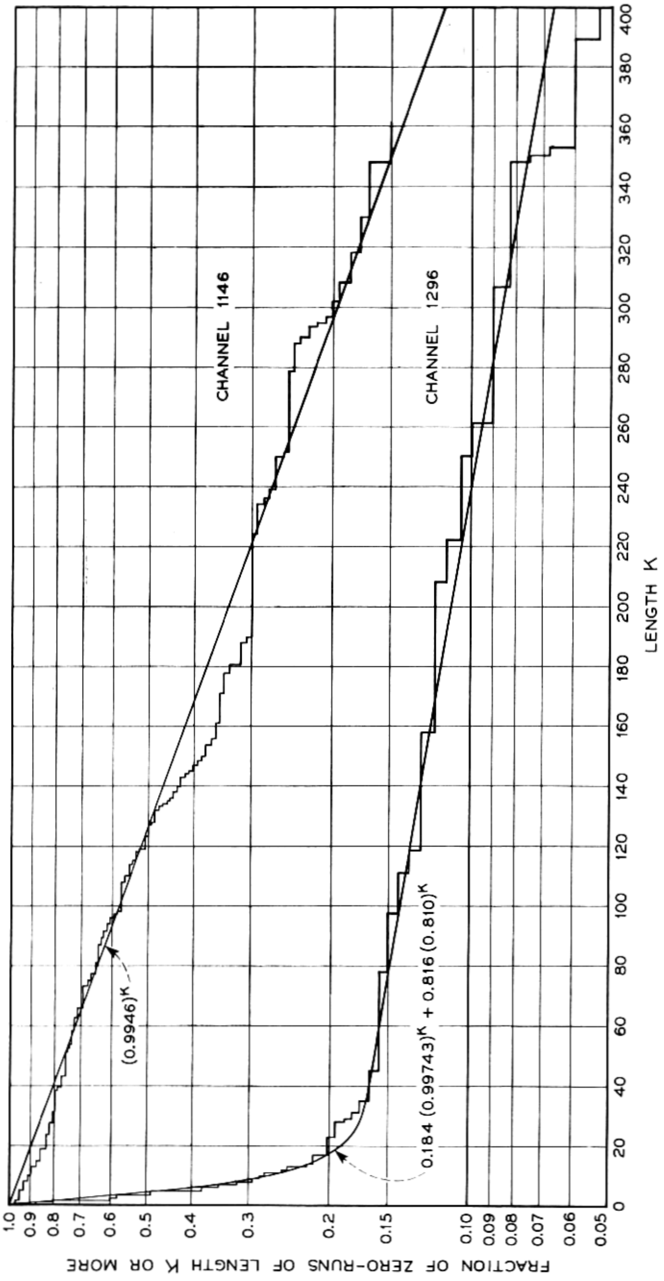

Fig. 3 — Match to experimental run data on telephone channels.

Data on runs of zero can provide another estimate of p, P, h. The fraction of runs of length K or more is an estimate of $u(K)$. By (14), one expects to find constants J, L, A such that

$$u(K) = AJ^K + (1 - A)L^K. \tag{19}$$

These constants are easily found by fitting a curve of the form (19) to the measured run distribution. First, A and J are chosen to give the correct behavior AJ^K for large K. Afterward, L is chosen to improve the fit for small K. Expressions for p, P, h in terms of A, J, L are

$$h = \frac{LJ}{J - A(J - L)},$$

$$P = \frac{(1 - L)(1 - J)}{1 - h},$$

$$p = A(J - L) + (1 - J)\left(\frac{L - h}{1 - h}\right).$$

Fig. 3 shows run distributions for two different telephone circuits transmitting binary data. These were two of the thousands of circuits in a recent large-scale program of telephone circuit measurements (see Alexander, Gryb and Nast.[4] * Channel 1146 carried an exchange call; it used loaded cable and only local exchange switching facilities. Channel 1296 was a toll channel longer than 500 miles; it used K-carrier, a radio path, and loaded cables at the ends. These channels were chosen as examples because they were two of the noisiest cases measured, and thus provided plenty of data. The step functions in Fig. 3 show the fractions of zero runs of lengths K or more from a sample of about 130 consecutive zero runs for each channel. The smooth curves show the curves (19) that fit these distributions. In the case of channel 1146, $u(K) = 0.9946^K$ provided a good fit; then channel 1146 was well approximated by a symmetric binary channel with $p = 0.9946$. The results for channel 1296 look more like Fig. 2. The straight line asymptote is the function AJ^K with parameters $A = 0.184$ and $J = 0.99743$ chosen to approximate the data for large K. The parameter value $L = 0.81$ makes the curve (19) fit the data for small K. These values of A, J, L provide the estimates

$$h = 0.84, \qquad P = 0.003, \qquad p = 0.034.$$

* The curves appearing in Ref. 4 show only combined data from hundreds of channels. Since these channels differ greatly among themselves, the curves in Ref. 4 do not have the form (19).

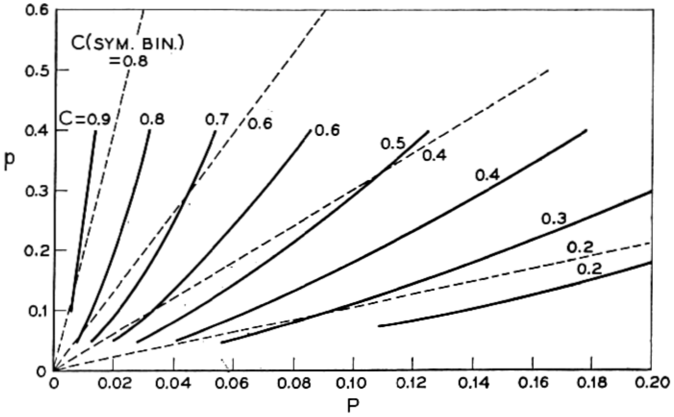

Fig. 4 — Capacities C and C(sym. bin.) as functions of p,P, with $h = \frac{1}{2}$.

The 500-digit sample of Section II provides a run distribution with more statistical fluctuations than in Fig. 3 because of the smaller sample size. The curve fitting yields $A = 0.385$, $J = 0.961$, $L = 0.32$ and $h = 0.432$, $P = 0.047$, $p = 0.232$.

VI. CAPACITY COMPUTATIONS

By (14) and (16), $u(K)$ and $v(K)$ behave like multiples of J^K for large K. In the most interesting cases P is small and J is nearly 1.0 ($J \geqq Q$ always); then (7) converges slowly. However,

$$\frac{u(K+1)}{u(K)} \to J$$

for large K and, by (6),

$$h(10^K) \to -J \log_2 J - (1 - J) \log_2 (1 - J) = h_0 .$$

Here, $h(10^K)$ approaches its limiting value h_0 rapidly; indeed, $L = Q + hq - J \leqq hq$. When $h = 0.5$, typical values of L are about 0.5 or less, and the L^K term in (14) becomes negligible when K reaches 10 or 15. Thus, the approximation $h(10^K) = h_0$ is good for all $K \geqq K_0$ where K_0 is only moderately large. The corresponding terms of the infinite series (5) sum to

$$\begin{aligned} \sum_{K=K_0}^{\infty} P(10^K)h_0 &= h_0 P(1) \sum_{K=K_0}^{\infty} u(K) \\ &= h_0 \left[1 - P(1) \sum_{K=0}^{K_0-1} u(K) \right]. \end{aligned}$$

The last step used the identity

$$P(1)[u(0) + u(1) + u(2) + \cdots] = 1,$$

which follows from (13) with $t = 1$. Then, the first $K_0 - 1$ terms of (5), together with the correction just derived, suffice to compute C accurately.

Fig. 4 shows contours of constant C and C(sym. bin.) versus p,P for $h = 0.5$. [C(sym. bin.) was computed with $P(1)$ given by (1)]. If the average burst length is not large (p not too small), the difference between the two capacities is slight.

The author is indebted to Miss M. A. Lounsberry for the computations shown in Figs. 2 and 4.

REFERENCES

1. Shannon, C. E., A Mathematical Theory of Communications, B.S.T.J., **27,** 1948; pp. 379; 623.
2. Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd Ed., John Wiley & Sons, New York, 1957.
3. McMillan, B., The Basic Theorems of Information Theory, Ann. Math. Stat., **24,** 1953, p. 196.
4. Alexander, A. A., Gryb, R. M. and Nast, D. W., Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **39,** 1960, p. 431.