and E_2: there exists an index $r \neq 1$, $r \in R$, such that

$$d(y_2,x(r)) \leq n(\alpha\bar{p} + \bar{\alpha}p) + n\varepsilon.$$

Thus

$$E\{\bar{p}_2(e)\} \leq \Pr\{E_1\} + \Pr\{E_2\}, \tag{63}$$

where here the probability is understood to range over the random choice of code as well as the selection of (r,s). From (61),

$$\Pr\{E_1\} = \Pr\{d(y_2,x(1)) > n(\alpha\bar{p} + \bar{\alpha}p + \varepsilon)\}$$
$$= \Pr\left\{\frac{1}{n}\sum_{i=1}^{n} z(s)_i \oplus e_i > \alpha\bar{p} + \bar{\alpha}p + \varepsilon\right\}. \tag{64}$$

We find the expected value (over e and s)

$$\frac{1}{n}\sum_{i=1}^{n} z(s)_i \oplus e_i = \frac{1}{n}\sum_{i=1}^{n} \Pr\{(z(s)_i,e_i) = (1,0) \text{ or } (0,1)\}$$
$$= \frac{1}{n}\sum_{i=1}^{n} (\alpha\bar{p} + \bar{\alpha}p) = \alpha\bar{p} + \bar{\alpha}p. \tag{65}$$

Also, after some calculation

$$\operatorname{var}\frac{1}{n}\sum_{i=1}^{n} z(s)_i \oplus e_i \leq \frac{p\bar{p}}{n}. \tag{66}$$

It follows that $d(y,x(r)) \to \alpha\bar{p} + \bar{\alpha}p$ in probability and therefore $\Pr\{E_1\} \to 0$ as $n \to \infty$.

We are left with the evaluation of $\Pr\{E_2\}$. We write

$$\Pr\{E_2\} \leq \Pr\{d(x(r),y_2) \leq n(\alpha\bar{p} + \bar{\alpha}p + \varepsilon), \text{ for some } r \neq 1 | x(1) \text{ transmitted}\}$$
$$\leq 2^{nR_{12}} \Pr\{d(x(2),y_2) \leq n(\alpha\bar{p} + \bar{\alpha}p + \varepsilon)\}. \tag{67}$$

But

$$d(x(2),y_2) = wt(x(2) \oplus x(1) \oplus z(s) \oplus e), \tag{68}$$

where wt denotes the number of 1's in the binary n-tuple, and $x(2)$ and $x(1)$ are independent Bernoulli n-sequences with parameter $\frac{1}{2}$. Thus, for any $\varepsilon > 0$,

$$\Pr\{E_2\} \leq 2^{nR_{12}} 2^{n(H(\alpha\bar{p}+\bar{\alpha}p)+O(\ln n/n)+\varepsilon')} 2^{-n}, \tag{69}$$

where $2^{n(H(\alpha\bar{p}+\bar{\alpha}p)+O(\ln n/n)+\varepsilon')}$ denotes the number of points

$$\sum_{i=0}^{n(\alpha\bar{p}+\bar{\alpha}p+\varepsilon)} \binom{n}{i}$$

in the decoding sphere centered at y_2. Consequently, if

$$R_{12} < 1 - H(\alpha\bar{p} + \bar{\alpha}p) - \varepsilon', \tag{70}$$

then $\Pr\{E_2\} \to 0$, as $n \to \infty$. Collecting the constraints of (60) and (70), we see that if

$$R_2 = R_{12} < 1 - H(\alpha\bar{p} + \bar{\alpha}p) \tag{71}$$
$$R_1 < H(\alpha) + R_2 = 1 - H(\alpha\bar{p} + \bar{\alpha}p) + H(\alpha),$$

then

$$E\{\bar{p}_1^{(n)}(e) + \bar{p}_2^{(n)}(e)\} = E\{\bar{p}_1^{(n)}(e)\} + E\{\bar{p}_2^{(n)}(e)\} \to 0. \tag{72}$$

Since the best code behaves better than the average, there must exist a sequence of $[(2^{nR_1},2^{nR_2},2^{nR_{12}}), n]$ codes for $n = 1,2,\ldots,$ with

$$R_1 = C(\alpha\bar{p} + \bar{\alpha}p) + H(\alpha) - \varepsilon$$
$$R_2 = C(\alpha\bar{p} + \bar{\alpha}p) - \varepsilon \tag{73}$$

such that

$$\bar{p}_1^{(n)}(e) + \bar{p}_2^{(n)}(e) \to 0, \tag{74}$$

and thus $\bar{p}_1^{(n)}(e) \to 0$, $\bar{p}_2^{(n)}(e) \to 0$.

Taking the limit of (R_1,R_2) as $\varepsilon \to 0$ proves the theorem.

References

[1] R. B. Ash, *Information Theory*. New York: Interscience, 1965.
[2] J. Wolfowitz, *Coding Theorems of Information Theory*, 2nd ed. Berlin: Springer, and Englewood Cliffs, N.J.: Prentice-Hall, 1964.
[3] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 1948, pp. 379–423 (pt. 1) and pp. 623–656 (pt. 2); also Urbana, Ill.: Univ. Illinois Press, 1949.
[4] C. E. Shannon, "Two-way communication channels," in *Proc. 4th Berkeley Symp. Probability and Statistics*, vol. 1. Berkeley, Calif.: Univ. California Press, 1961, pp. 611–644.
[5] D. Blackwell, L. Breiman, and A. Thomasian, "The capacity of a class of channels," *Ann. Math. Statist.*, vol. 30, 1959, pp. 1229–1241.
[6] C. E. Shannon, "A note on a partial ordering for communication channels," *Inform. Contr.*, vol. 1, 1958, pp. 390–397.

An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels

SUGURU ARIMOTO

Abstract—A systematic and iterative method of computing the capacity of arbitrary discrete memoryless channels is presented. The algorithm is very simple and involves only logarithms and exponentials in addition to elementary arithmetical operations. It has also the property of monotonic convergence to the capacity. In general, the approximation error is at least inversely proportional to the number of iterations; in certain circumstances, it is exponentially decreasing. Finally, a few inequalities that give upper and lower bounds on the capacity are derived.

Manuscript received September 9, 1970.
The author is with the Faculty of Engineering Science, Osaka University, Osaka, Japan.

I. Introduction

It is well known that the capacity of discrete memoryless channels that are symmetric from the input can easily be evaluated. Muroga [1] developed a method for straightforward evaluation of capacity, but unfortunately its usefulness is restricted to the case where 1) the channel matrix is square and nonsingular, and 2) each probability in the resulting input probability distribution is positive. It appears that for nonsquare channel matrices there exist no general algorithms that compute capacity exactly and straightforwardly.

Recently, Meister and Oetti [2] proposed an iterative procedure based upon the method of concave programming and showed that it converges to the capacity. This paper is also concerned with an iterative method for calculation of the capacity. The procedure turns out to be somewhat similar to that given by Meister and Oetti [2], but the idea originates from a concept of generalized equivocation measure due to the author [3]. Therefore, without using the concept of concave or convex functions, it is shown that the algorithm has the property of monotonic convergence to the exact capacity. Finally, along with some results concerning speed of convergence, a few inequalities giving upper and lower bounds on the capacity are presented.

II. Mutual Information and Equivocation

Denote a discrete memoryless channel with n input and m output symbols by the stochastic $m \times n$ matrix P:

$$P = \{p(i/j)\} \qquad i = 1,\cdots,m, j= 1,\cdots,n, \tag{1}$$

where $p(i/j) \geq 0$ and $\sum_{i=1}^{m} p(i/j) = 1$. As a matter of course, we assume that for every i ($i = 1,\cdots,m$) there exists at least one j ($j = 1,\cdots,n$) such that $p(i/j) > 0$. To simplify the notation, two types of sets of probability distribution vectors are introduced as follows:

$$D^k = \left\{ p = (p_1,\cdots,p_k) \,\middle|\, p_i > 0, \sum_{i=1}^{k} p_i = 1 \right\},$$

$$\bar{D}^k = \left\{ p = (p_1,\cdots,p_k) \,\middle|\, p_i \geq 0, \sum_{i=1}^{k} p_i = 1 \right\}. \tag{2}$$

For example, the m-dimensional probability vector

$$p(\cdot/j) = \{(p(1/j),\cdots,p(m/j)\} \tag{3}$$

belongs, in general, to $\bar{D}^k$.

The mutual information concerning the channel P is defined by

$$I(P;p) = H(p) - H(P;p), \tag{4}$$

where

$$H(p) = \sum_{j=1}^{n} -p_j \log p_j, \tag{5}$$

$$H(P;p) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(i/j)p_j \log \frac{p(i/j)p_j}{\sum_{k=1}^{n} p(i/k)p_k}, \tag{6}$$

where $p = (p_1,\cdots,p_n) \in \bar{D}^n$ is a probability vector of input symbols. From Shannon's coding theorem [3] the capacity of the memoryless channel P, which will be denoted by $C(P)$, is given by

$$C(P) = \max_{p \in \bar{D}^n} I(P;p). \tag{7}$$

On the other hand, one can generalize the concept of the equivocation measure $H(P;p)$ defined in (6). In fact, according to [4], a stochastic $n \times m$ matrix ϕ is introduced such that

$$\phi = \{\phi(j/i)\}, \qquad i = 1,\cdots,m, j = 1,\cdots,n, \tag{8}$$

where $\phi(j/i) \geq 0$ and $\sum_{j=1}^{n} \phi(j/i) = 1$ and the generalized equivocation is defined as

$$J(P;p,\phi) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(i/j)p_j \log \phi(j/i). \tag{9}$$

Then, if $\phi(j/i)$ is defined by the Bayes formula

$$\phi(j/i) = \frac{p(i/j)p_j}{\sum_{k=1}^{n} p(i/k)p_k} = p^*(j/i), \tag{10}$$

(9) becomes equal to $H(P;p)$. Furthermore, one can easily prove the inequality

$$J(P;p,\phi) \geq J(P;p,P^*), \tag{11}$$

where P^* is the stochastic matrix whose (j,i)th entry is $p^*(j/i)$ as defined in (10). In view of this fact, one obtains another characterization of the capacity as

$$C(P) = \max_{p \in \bar{D}^n} \max_{\phi \in \Phi} [H(p) - J(P;p,\phi)], \tag{12}$$

where Φ denotes the set of all stochastic matrices satisfying (8).

The following well-known proposition will be utilized subsequently, e.g., see [5, theorem 7.5].

Proposition 1: The probability vector $p^0 = (p_1{}^0,\cdots,p_n{}^0) \in \bar{D}^n$ maximizes $I(P;p)$ if and only if

$$\sum_{i=1}^{m} p(i/j) \cdot \log \frac{p(i/j)}{\sum_{k=1}^{n} p(i/k)p_k{}^0} \begin{cases} = I(P;p^0) = C(P), & p_j{}^0 > 0 \\ \leq I(P;p^0) = C(P), & p_j{}^0 = 0. \end{cases} \tag{13}$$

Moreover, if $p^0,\cdots,p^{k-1} \in \bar{D}^n$ all maximize $I(P;p)$, then any linear combination such that

$$p^k = \alpha_0 p^0 + \cdots + \alpha_{k-1} p^{k-1}, \tag{14}$$

where $\alpha = (\alpha_0,\cdots,\alpha_{k-1}) \in \bar{D}^k$ also maximizes $I(P;p)$ and in addition

$$\sum_{j=1}^{n} p(i/j)p_j{}^0 = \cdots = \sum_{j=1}^{n} p(i/j)p_j{}^k. \tag{15}$$

III. Iterative Procedure and Convergence

Based upon the characterization (12) of capacity, we now propose a procedure for capacity evaluation, which comprises the following steps.

i) Initially, choose an arbitrary probability vector $p^1 \in D^n$ (in practice the uniform probability distribution $p_j{}^1 = 1/n$ for all $j = 1,\cdots,n$ is generally suitable). Then, the following two steps are iterated as $t = 1,2,\cdots$.

ii)$_t$ Maximize $H(p^t) - J(P;p^t,\phi)$ with respect to $\phi \in \Phi$

while fixing p^t. According to (11) the maximizing ϕ is

$$\phi^t(j/i) = \frac{p(i/j)p_j^t}{\sum_{k=1}^n p(i/k)p_k^t}, \tag{16}$$

that is,

$$C(t,t) = \max_{\phi \in \Phi} [H(p^t) - J(P;p^t,\phi)] = H(p^t) - J(P;p^t,\phi^t). \tag{17}$$

iii)$_t$ Maximize $H(p) - J(P;p,\phi^t)$ with respect to $p \in \bar{D}^n$ while fixing ϕ^t. This maximizing probability vector, denoted by p^{t+1}, is given by

$$p_j^{t+1} = \frac{r_j^t}{\sum_{k=1}^n r_k^t}, \quad r_j^t = \exp\left[\sum_{i=1}^m p(i/j)\log\phi^t(j/i)\right]. \tag{18}$$

In fact, the following lemma is quite obvious.

Lemma 1: For any fixed $\phi \in \Phi$,

$$\max_{p \in \bar{D}^n} [H(p) - J(P;p,\phi)] = H(p^*) - J(P;p^*,\phi) = \log\left[\sum_{j=1}^n \exp\left\{\sum_{i=1}^m p(i/j)\log\phi(j/i)\right\}\right] \le C(P), \tag{19}$$

where $p^* \in \bar{D}^n$ is given by

$$p_j^* = \frac{r_j}{\sum_{k=1}^n r_k}, \quad r_j = \exp\left[\sum_{i=1}^m p(i/j)\log\phi(j/i)\right], \quad j = 1,\cdots,n. \tag{20}$$

Proof: To prove this by using a Lagrange multiplier, let

$$F(p) = H(p) - J(P;p,\phi) + \lambda\left(1 - \sum_{j=1}^n p_j\right). \tag{21}$$

Then it follows from a theorem of nonlinear programming that a maximizing vector $p^* \in \bar{D}^n$ satisfies the relation

$$\left.\frac{\partial F}{\partial p_j}\right|_{p_j = p_j^*} = -1 - \log p_j^* + \sum_{i=1}^m p(i/j)\log\phi(j/i) - \lambda = 0, \quad p_j^* > 0. \tag{22}$$

The right-hand equality of (22) reduces to

$$p_j^* = \exp\left[-1 - \lambda + \sum_{i=1}^m p(i/j)\log\phi(j/i)\right], \quad p_j^* > 0. \tag{23}$$

Taking into account the relation $\sum_j p_j^* = 1$ and (23), we have (20) for $p_j^* > 0$. As for the case $p_j^* = 0$, note that from the continuity of $F(p)$ in $p \in \bar{D}^n$ there exists at least one i such that $\phi(j/i) = 0$ and $p(i/j) > 0$. Therefore, to avoid confusion, we assume that $\log 0 = -\infty$ and $0 \cdot \log 0 = 0$. Thus the relation (20) is valid even for $p_j^* = 0$.

Corollary 1:

$$\max_{p \in \bar{D}^n} [H(p) - J(P;p,\phi)] = \log\left(\sum_{j=1}^n r_j\right), \tag{25}$$

where

$$r_j = \exp\left[\sum_{i=1}^m p(i/j)\log\phi(j/i)\right]. \tag{26}$$

Now we prove the convergence of the procedure.

Theorem 1: Let $p^1 \in D^n$. Then the values $C(t,t)$ defined by (17) converge monotonically from below to the capacity $C(P)$ as $t \to \infty$.

Proof: At step iii)$_t$, let

$$C(t+1,t) = \max_{p \in \bar{D}^n} [H(p) - J(P;p,\phi^t)] = H(p^{t+1}) - J(P;p^{t+1},\phi^t). \tag{27}$$

Then, it follows from Corollary 1 that

$$C(t+1,t) = \log\left(\sum_{j=1}^n r_j^t\right) \tag{28}$$

and, furthermore, from the definitions of $C(t,t)$ and $C(t+1,t)$ it follows that

$$C(1,1) \le C(2,1) \le C(2,2) \le \cdots \le C(t,t) \le C(t+1,t) \le \cdots \le C(P). \tag{29}$$

Let p^0 be one of the input probability vectors maximizing the mutual information; i.e., $I(P;p^0) = C(P)$. Then, using the notations

$$q_i^0 = \sum_{j=1}^n p(i/j)p_j^0, \quad q_i^t = \sum_{j=1}^n p(i/j)p_j^t, \tag{30}$$

and taking into account (28) and Proposition 1, we have

$$\begin{aligned}\sum_{k=1}^n p_k^0 \log\frac{p_k^{t+1}}{p_k^t} &= \sum_{k=1}^n p_k^0 \log\left(\frac{1}{p_k^t}\cdot\frac{r_k^t}{\sum_j r_j^t}\right) \\ &= -C(t+1,t) + \sum_{k=1}^n p_k^0\left(\sum_{i=1}^m p(i/k)\log\frac{p(i/k)}{q_i^t}\right) \\ &= -C(t+1,t) + \sum_{k=1}^n\sum_{i=1}^m p(i/k)p_k^0 \cdot\left[\log\frac{p(i/k)}{q_i^0} + \log\frac{q_i^0}{q_i^t}\right] \\ &= C(P) - C(t+1,t) + \sum_{i=1}^m q_i^0 \log\frac{q_i^0}{q_i^t}.\end{aligned} \tag{31}$$

Since the last term of (31) is nonnegative,

$$\sum_{i=1}^m q_i^0 \log(q_i^0/q_i^t) \ge \sum_{i=1}^m q_i^0\left(1 - \frac{q_i^t}{q_i^0}\right) = 0,$$

it is seen that

$$C(P) - C(t+1,t) \le \sum_{k=1}^n p_k^0 \log(p_k^{t+1}/p_k^t). \tag{32}$$

Obviously, this inequality holds for any $t = 1,2,\cdots$. Hence, summing up (32) from $t = 1$ to $t = N$,

$$\sum_{t=1}^{N} [C(P) - C(t+1,t)] \le \sum_{k=1}^{n} p_k^0 \log (p_k^{N+1}/p_k^1)$$
$$\le \sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^1) \quad (33)$$

for any integer $N \ge 1$. Note that the right-hand side of (33) is finite and constant since $p^1 \in D^n$. Thus, since the value $C(P) - C(t+1, t)$ is nonnegative and nonincreasing with increasing t, it has been clearly shown that

$$\lim_{t\to\infty} C(t+1,t) = \lim_{t\to\infty} C(t,t) = C(P). \quad (34)$$

Corollary 1: The approximation error $e(t) = C(P) - C(t+1, t)$ is inversely proportional to the number of iterations. In particular, if p^1 is chosen as the uniform distribution, then

$$C(P) - C(t+1,t) \le [\log n - H(p^0)]/t. \quad (35)$$

Proof: The proof follows straightforwardly from (33).

Corollary 2:

$$\lim_{N\to\infty} q^N = q^0. \quad (36)$$

Proof: First of all, note that from [6, theorem 3.1] the convergence of $\omega(N)$ defined by

$$\omega(N) = \sum_{i=1}^{m} q_i^0 \log (q_i^0/q_i^N) \quad (37)$$

to zero implies (36). Therefore, to show the convergence of $\omega(N)$, we rewrite (31) as follows:

$$\sum_{t=1}^{N} \omega(t) = \sum_{t=1}^{N} \left(\sum_{i=1}^{m} q_i^0 \log (q_i^0/q_i^t)\right)$$
$$\le \sum_{t=1}^{N} \left(\sum_{k=1}^{n} p_k^0 \log (p_k^{t+1}/p_k^t)\right)$$
$$= \sum_{k=1}^{n} p_k^0 \log (p_k^{N+1}/p_k^1)$$
$$\le \sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^1), \quad (38)$$

where the nonnegativeness of $C(P) - C(t+1, t)$ is taken into account. Since $p^1 \in D^n$, the right-hand side of (38) is finite and constant. Thus, (38) implies that $\omega(N) \to 0$ as $N \to \infty$.

IV. Speed of Convergence

In the case where the input distribution p^0 achieving capacity is unique and belongs to D^n, we can deduce more detailed results. In particular, the speed of convergence is considerably improved over that predicted by Corollary 1 of Theorem 1.

Theorem 2: If the input distribution achieving capacity is unique, then the sequence of input distributions p^t converges to p^0 monotonically, in the sense that

$$\sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^t) \ge \sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^{t+1})$$
$$\ge \to 0 \quad \text{as } t \to \infty. \quad (39)$$

Proof: The monotonicity follows directly from (31). To prove convergence, note that the set $\bar{D}^n$ is a closed and bounded subset of n-dimensional Euclidean space. Hence the sequence $\{p^t\}$ has at least a point of accumulation $\bar{p} = (\bar{p}_1,\cdots,\bar{p}_n)$ in $\bar{D}^n$. Theorem 1 implies that $\bar{p}$ also maximizes the mutual information. Thus, by virtue of the uniqueness of p^0, we conclude that $\bar{p} = p^0$.

Theorem 3: If the input distribution p^0 achieving capacity is unique and belongs to D^n, then there exist an integer N and a constant $0 < \sigma \le 1$ such that for all $t \ge N$

$$\sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^t) \le (1-\sigma)^{t-N} \sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^N) \quad (40)$$

and N is independent of σ.

Proof: It follows from Theorem 2 that for an arbitrary number $\delta > 0$ there exists a positive integer $N = N(\delta)$ such that

$$\sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^N) < \delta. \quad (41)$$

Of course, p^N approaches p^0 as $\delta \to 0$. Henceforth, consider the difference vector $d = p^0 - p^N$ and note that

$$\sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^N) = \sum_{k=1}^{n} - p_k^0 \log \left(1 - \frac{d_k}{p_k^0}\right)$$
$$= \tfrac{1}{2}(d, P_0^{-1} d) + \sum_{k=1}^{n} O(d_k^3), \quad (42)$$

$$\sum_{i=1}^{m} q_i^0 \log (q_i^0/q_i^N) = \sum_{i=1}^{m} - q_i^0$$
$$\cdot \log \left[1 + \sum_{k=1}^{n} p(i/k)\, d_k/q_i^0\right]$$
$$= \tfrac{1}{2}(d, P^T Q_0^{-1} P d) + \sum_{k=1}^{n} 0(d_k^3), \quad (43)$$

where p^T denotes the transpose of the channel matrix P, the notation (x,y) for vectors x and y denotes the inner product, and P_0,Q_0 are diagonal matrices such that

$$P_0 = \begin{pmatrix} p_1^0 & & 0 \\ & \ddots & \\ 0 & & p_n^0 \end{pmatrix}, \qquad Q_0 = \begin{pmatrix} q_1^0 & & 0 \\ & \ddots & \\ 0 & & q_m^0 \end{pmatrix}.$$

If $\delta > 0$ is chosen sufficiently small, then from (42) and (43) one can choose an integer N such that the following inequalities are satisfied simultaneously:

$$\sum_{k=1}^{n} p_k^0 \log (p_k^0/p_k^N) \le \tfrac{2}{3}(d, P_0^{-1}d),$$
$$\sum_{i=1}^{m} q_i^0 \log (q_i^0/q_i^N) \ge \tfrac{1}{3}(d, P^T Q_0^{-1} P d). \quad (44)$$

To show the positive definiteness of the matrix $P^T Q_0^{-1} P$, assume that there exists a nonzero vector $\xi \in R^n$ such that $P\xi = 0$. Then the vector ξ clearly has a representation of the form

$$\xi = \alpha e + \beta d^1, \tag{45}$$

where α and β are scalars, e is the n-vector whose components are all equal to unity, and d^1 is an n-vector that is orthogonal to e. Note that the assumption $P\xi = 0$ implies

$$0 = \sum_{i=1}^{m} \sum_{j=1}^{n} p(i/j)\xi_j = \sum_{i=1}^{m} \sum_{j=1}^{n} p(i/j)(\alpha e_j + \beta d_j^1)$$
$$= \sum_{j=1}^{n} \sum_{i=1}^{m} p(i/j)(\alpha + \beta d_j^1) = \sum_{j=1}^{n} [\alpha + (e,d^1)] = n\alpha, \tag{46}$$

which means $\alpha = 0$. Hence, $P\xi = \beta P d^1 = 0$, which implies $\beta = 0$ or $Pd^1 = 0$. If $Pd^1 = 0$, then for a sufficiently small number $\varepsilon > 0$ the vector $p = p^0 + \varepsilon d^1$ becomes also the capacity-achieving distribution in view of Proposition 1, since $Pd^1 = 0$ and $p \in D^n$. Thus, from the uniqueness of the capacity-achieving distribution, it follows that $\beta = 0$ or $d^1 = 0$, but this fact implies $\xi = 0$, which contradicts the assumption. Due to the positive definiteness of $P^T Q_0^{-1} P$, there exists a positive number $\sigma > 0$ such that

$$\tfrac{1}{3}(d, P^T Q_0^{-1} P d) \geq \sigma \tfrac{2}{3} (d, P_0^{-1} d), \qquad \text{for every } d \in R^n. \tag{47}$$

Substituting this inequality into (44), and noting that on account of Theorem 2 the same discussion is valid for all t such that $t \geq N$, we obtain

$$\sum_{i=1}^{m} q_i^0 \log (q_i^0/q_i^t) \geq \sigma \sum_{j=1}^{n} p_j^0 \log (p_j^0/p_j^t), \qquad \text{for all } t \geq N. \tag{48}$$

On the other hand, it follows from (31) that

$$\sum_{j=1}^{n} p_j^0 \log \frac{p_j^0}{p_j^{t+1}} \leq \sum_{j=1}^{n} p_j^0 \log \frac{p_j^0}{p_j^t} - \sum_{i=1}^{m} q_i^0 \log \frac{q_i^0}{q_i^t}. \tag{49}$$

Finally, (40) follows directly from substituting (48) into (49).

V. Upper and Lower Bounds on the Capacity

In this section we will derive some properties of $C(P)$. In particular we give some approximation formulas that give upper and lower bounds on the capacity.

First, let

$$C(P,\phi) = \max_{p \in \bar{D}^n} [H(p) - J(P; p,\phi)]. \tag{50}$$

Note that the exact form of $C(P,\phi)$ is given in (19), and $C(P,\phi) \leq C(P)$. Of course, from Proposition 1

$$\max_{\phi \in \Phi} C(P,\phi) = C(P). \tag{51}$$

Moreover, we can prove the following theorem.

Theorem 4: Let P_1 and P_2 be $m \times n$ channel matrices, respectively, α an arbitrary number such that $0 \leq \alpha \leq 1$, and ϕ an arbitrary $n \times m$ stochastic matrix. Then the following inequality holds:

$$C(\alpha P_1 + (1-\alpha)P_2,\phi) \leq \alpha C(P_1,\phi) + (1-\alpha)C(P_2,\phi). \tag{52}$$

Proof: Let

$$x_j = \exp\left[\sum_{i=1}^{m} p_1(i/j) \log \phi(j/i)\right],$$
$$y_j = \exp\left[\sum_{i=1}^{m} p_2(i/j) \log \phi(j/i)\right]. \tag{53}$$

It then follows that

$$C(\alpha P_1 + (1-\alpha)P_2, \phi) = \log\left[\sum_{j=1}^{n} x_j^{\alpha} y_j^{1-\alpha}\right]. \tag{54}$$

Applying the Hölder inequality to the right-hand side of (54),

$$\log\left[\sum_{j=1}^{n} x_j^{\alpha} y_j^{1-\alpha}\right] \leq \alpha \log\left[\sum_{j=1}^{n} x_j\right] + (1-\alpha)\log\left[\sum_{j=1}^{n} y_j\right]$$
$$= \alpha C(P_1,\phi) + (1-\alpha)C(P_2,\phi). \tag{55}$$

Equations (54) and (55) constitute (52).

Using (51) and (52), we can easily prove the following inequality, which was first given by Shannon [7].

Corollary 1:

$$C(\alpha P_1 + (1-\alpha)P_2) \leq \alpha C(P_1) + (1-\alpha)C(P_2). \tag{56}$$

Proof:

$$C(\alpha P_1 + (1-\alpha)P_2) = \max_{\phi\in\Phi} C(\alpha P_1 + (1-\alpha)P_2,\phi)$$
$$\leq \max_{\phi\in\Phi} [\alpha C(P_1,\phi) + (1-\alpha)C(P_2,\phi)]$$
$$\leq \alpha \max_{\phi\in\Phi} C(P_1,\phi) + (1-\alpha)\max_{\phi\in\Phi} C(P_2,\phi)$$
$$= \alpha C(P_1) + (1-\alpha)C(P_2). \tag{57}$$

Now we will present four types of inequalities that give upper and lower bounds to the capacity.

Theorem 5:

$$\text{i)}\quad C(P) \geq \log n - \left(1 - \frac{\Delta}{n}\right)\log(n-1) - H\left(\frac{\Delta}{n}, 1 - \frac{\Delta}{n}\right), \tag{58}$$

ii) $$C(P) \geq \log \left[\sum_{j=1}^{n} \exp \left\{ \sum_{i=1}^{m} p(i/j) \cdot \log \frac{p(i/j)}{\sum_{k=1}^{n} p(i/k)} \right\} \right], \quad (59)$$

iii) $$C(P) \geq \log n - \frac{1}{n} \sum_{j=1}^{n} H(p(^*/j)) - \frac{1}{n} \sum_{i=1}^{m} \cdot \left(\sum_{j=1}^{n} p(i/j) \log \left(\sum_{j=1}^{n} p(i/j) \right) \right), \quad (60)$$

iv) $$C(P) \leq \log n + \max_{j} \left[\sum_{i=1}^{m} p(i/j) \cdot \log \frac{p(i/j)}{\sum_{k=1}^{n} p(i/k)} \right]. \quad (61)$$

Here Δ is defined as

$$\Delta = \sum_{i=1}^{m} p(i/j_i), \quad (62)$$

where j_i denotes one of integers arbitrarily chosen from 1 to n, correspondingly to each i, and

$$H\left(\frac{\Delta}{n}, 1 - \frac{\Delta}{n}\right) = -\frac{\Delta}{n} \log \frac{\Delta}{n} - \left(1 - \frac{\Delta}{n}\right) \log \left(1 - \frac{\Delta}{n}\right). \quad (63)$$

Proof: To show i), let ε be an arbitrary number such that $0 \leq \varepsilon \leq 1$ and define

$$p_j = 1/n, \qquad j = 1, \cdots, n,$$
$$\phi(j/i) = \begin{cases} 1 - \varepsilon, & j = j_i \\ \varepsilon/(n-1), & j \neq j_i. \end{cases} \quad (64)$$

Then, from (50) and (51) it follows that

$$\begin{aligned} C(P) &\geq H(p) - J(P; p, \phi) \\ &= \log n + \sum_{i=1}^{m} \frac{1}{n} p(i/j_i) \log (1 - \varepsilon) + \sum_{i=1}^{m} \sum_{j \neq j_i} \frac{1}{n} p(i/j) \log \frac{\varepsilon}{n-1} \\ &= \log n - \left(1 - \frac{\Delta}{n}\right) \log (n-1) + \frac{\Delta}{n} \log (1 - \varepsilon) + \left(1 - \frac{\Delta}{n}\right) \log \varepsilon. \end{aligned} \quad (65)$$

Maximizing the right-hand side of (65) with respect to $0 \leq \varepsilon \leq 1$, we obtain i). To derive ii) and iii), let

$$p_j = 1/n, \qquad j = 1, \cdots, n$$
$$\phi(j/i) = p(i/j) \Big/ \sum_{k=1}^{n} p(i/k). \quad (66)$$

Substituting these into (19) yields both ii) and iii). In order to prove iv), let $p \in D^n$ be arbitrary and $p^0 \in \bar{D}^n$ be a distribution achieving capacity. Then it follows from Proposition 1 that

$$\begin{aligned} C(P) &= \sum_{i=1}^{m} \sum_{j=1}^{n} p(i/j) p_j^0 \log p(i/j)/q_i^0 \\ &= -\sum_{j=1}^{n} q_i^0 \log \frac{q_i^0}{q_i} + \sum_{i=1}^{m} \sum_{j=1}^{n} p(i/j) p_i^0 \log \frac{p(i/j)}{q_i} \\ &\leq \max_j \left[\sum_{i=1}^{m} p(i/j) \log \frac{p(i/j)}{q_i} \right]. \end{aligned} \quad (67)$$

The inequality iv) follows immediately from setting $p_j = 1/n$ for $j = 1, \cdots, n$ in (67).

It should be noted that the inequality i) is an extension of Helgert's (see [8, theorem 1]).

Numerical Example: Consider the channel matrix

$$P = \begin{pmatrix} 0.6 & 0.7 & 0.5 \\ 0.3 & 0.1 & 0.05 \\ 0.1 & 0.2 & 0.45 \end{pmatrix}, \quad (68)$$

which is the same as that given by Helgert [9] as a computation example. Table I shows upper and lower bounds on the capacity of the channel given by (68). (Throughout this example, the logarithm to base 2 is employed.) Numerical results corresponding to i)–iv) in Table I are obtained, respectively, by using inequalities i)–iv) in Theorem 5, and those corresponding to Theorems 1–4 are calculated according to Theorems 1–4 presented by Helgert [8]. The tightest bounds are ii) and iv) are

$$0.1257 \leq C(P) \leq 0.1708. \quad (69)$$

In fact, the iterative procedure proposed in Section III yields Table II, where p^0 is the input distribution achieving

TABLE I
UPPER AND LOWER BOUNDS ON THE CAPACITY OF THE CHANNEL GIVEN BY (68)

Lower Bound			
i)	ii)	iii)	Theorem 1
0.0691	0.1257	0.1238	0.0691
Upper Bound			
iv)	Theorem 2	Theorem 3	Theorem 4
0.1708	1.0161	0.5985	0.5547

TABLE II
CAPACITY AND INPUT DISTRIBUTION ACHIEVING THE CAPACITY OF THE CHANNEL GIVEN BY (68)

$C(P)$	0.161628
p_1^0	0.501723
p_2^0	0.0
p_3^0	0.498277

the capacity $C(P) = 0.161628$. It should be noted that in this case the straightforward method due to Muroga [1] cannot be applied, since it leads to a negative probability $p_2^0 < 0$.

References

[1] S. Muroga, "On the capacity of a discrete channel I," *J. Phys. Soc. Jap.*, vol. 8, 1953, pp. 484–494.
[2] B. Meister and W. Oetti, "On the capacity of a discrete, constant channel," *Inform. Contr.*, vol. 11, 1967, pp. 341–351.
[3] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 1948, pp. 379–423, 623–656.
[4] S. Arimoto, "The basis of sequential estimation process from the viewpoint of information theory," *Inform. Process. Jap.*, vol. 9, 1969, pp. 51–55.
[5] F. Jelinek, *Probabilistic Information Theory.* New York: McGraw-Hill, 1968.
[6] S. Kullback, *Information Theory and Statistics.* New York: Dover, 1968.
[7] C. Shannon, "Geometrische Deutung Einiger Ergebnisse bei der Berechnung der Kanalkapazität," *Nachrichtentech. Z.*, vol. 1, 1957, pp. 1–4.
[8] H. J. Helgert, "A comparison of arbitrary and symmetric channels on the basis of capacity," *IEEE Trans. Inform. Theory*, vol. IT-13, July 1967, pp. 467–470.
[9] ——, "On a bound for channel capacity," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-13, Jan. 1967, pp. 124–126.

Narrow-Band Systems and Gaussianity

ATHANASIOS PAPOULIS, FELLOW, IEEE

Abstract—**The approach to Gaussianity of the output $y(t)$ of a narrow-band system $h(t)$ is investigated. It is assumed that the input $x(t)$ is an a-dependent process, in the sense that the random variables $x(t)$ and $x(t+u)$ are independent for $u > a$. With $F(y)$ and $G(y)$ the distribution functions of $y(t)$ and of a suitable normal process, a realistic bound B on the difference $F(y) - G(y)$ is determined, and it is shown that $B \to 0$ as the bandwidth ω_0 of the system tends to zero. In the special case of the shot noise process**

$$y(t) = \sum_i h(t - t_i)$$

it is shown that

$$|F(y) - G(y)| < (\omega_0/\lambda)^{\frac{1}{2}}$$

where λ_i is the average density of the Poisson points t_i.

I. Introduction

In THE engineering applications of random signals it is often assumed that if a stationary process $x(t)$ is the input to a linear system, then the resulting response

$$y(t) = \int_{-\infty}^{\infty} x(t - \alpha)h(\alpha)\, d\alpha \tag{1}$$

tends to a normal process as the bandwidth ω_0 of the system tends to zero. This theorem is not always true, as one can show with a trivial counterexample. However, it holds under fairly general conditions. To apply it meaningfully, we need to establish not only conditions for its asymptotic validity, but also realistic bounds for the deviation of $y(t)$ from Gaussianity for a given $\omega_0 \neq 0$.

As one might expect from the central limit theorem, $y(t)$ will approach Gaussianity if the past $x(u)$, $u \leq t$, of the process $x(t)$ is "almost" independent of its future $x(u)$, $u \geq t + \tau$, for sufficiently large τ. This loose requirement is precisely formulated in Rosenblatt's classic paper [1] as follows.

Let B_t and F_τ be the Borel fields generated by the random variables $x(u)$ for $u \leq t$ and $u \geq t + \tau$, respectively. We say that the process $x(t)$ satisfies the *strong mixing condition* if there is a function $g(\alpha)$ with

$$0 \leq g(\alpha) \downarrow 0 \qquad \text{as} \qquad \alpha \to \infty \tag{2}$$

such that for any pair of events $B \in B_t$, $F \in F_\tau$

$$|P(BF) - P(B)P(F)| < g(\tau). \tag{3}$$

Assuming further that the moments of $x(t)$ of order up to four exist and satisfy certain conditions and that its power spectrum $S(\omega)$ is such that

$$S(\omega) > 0, \tag{4}$$

he shows that the output $y(t)$ of a certain class of filters tends to Gaussianity.

In applying Rosenblatt's result, we are faced with the problem of testing the mixing condition (3). Furthermore, the problem of establishing realistic bounds for the distance of $y(t)$ from Gaussianity remains. In this paper we shall overcome these difficulties but only at the sacrifice of generality. We shall base our analysis on the assumption that the process $x(t)$ is a-dependent, i.e., that the events B and F are independent for $\tau > a$. This assumption is equivalent to the condition

$$g(\alpha) = 0, \qquad \alpha > a. \tag{5}$$

The above is, of course, more restrictive than (2); however, it holds in many applications and can often be readily established. Suppose, for example, that $x(t)$ is the output

Manuscript received April 7, 1971; revised June 28, 1971. This work was supported by the Rome Air Development Center, Air Force Systems Command, Griffiss Air Force Base, Rome, N.Y., and the Air Force Office of Aerospace Research under Contract F 30602-69-C-0331.
The author is with the Department of Electrical Engineering, Polytechnic Institute of Brooklyn, Brooklyn, N.Y.