EXPONENTIAL ERROR BOUNDS FOR FINITE STATE CHANNELS

DAVID BLACKWELL
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction and summary

A *finite state channel* is defined by (1) a finite nonempty set A, the set of *inputs*, (2) a finite nonempty set B, the set of *outputs*, (3) a finite nonempty set T, the set of (channel) states, (4) a *transition law* $p = p(t'|t, a)$, specifying the probability that, if the channel is in state t and is given input a, the resulting state is t', and (5) a function ψ from T to B, specifying the output $b = \psi(t)$ of the channel when it is in state t.

For any sequence $\{a_n, n = 1, 2, \cdots\}$ of random variables with values in A, we may consider the process $\{a_n\}$ as supplying the inputs for the channel, as follows: an initial channel state t_0 is selected with a uniform distribution over T. The input a_1 is then given the channel. The channel then selects a state t_1, with

$$P\{t_1 = t|t_0, a_1\} = p(t|t_0, a_1) \tag{1}$$

and produces output $b_1 = \psi(t_1)$. The channel is then given input a_2 and selects state t_2, with

$$P\{t_2 = t|t_0, t_1, a_1, a_2, b_1\} = p(t|t_1, a_2), \tag{2}$$

and so on. In general, for $n \geqq 0$,

$$\begin{aligned} P\{a_{n+1} = a, t_{n+1} = t, b_{n+1} = b|a_i, 1 \leqq i \leqq n, t_i, 0 \leqq i \leqq n, b_i, 1 \leqq i \leqq n\} \\ = P\{a_{n+1} = a|a_i, i \leqq n\}p(t|t_n, a)\chi(t, b), \end{aligned} \tag{3}$$

where $\chi(t, b) = 1$ if $\psi(t) = b$ and 0 otherwise.

For any random variable x with a finite set of values and any random variable y, the (nonnegative) random variable whose value when $x = x_0$ and $y = y_0$ is

$$-\log P\{x = x_0|y = y_0\} \tag{4}$$

(all logs are base 2) is called the (conditional) entropy of x given y and will be denoted by $i(x|y)$. Its expected value, which cannot exceed the log of the number of values of x, will be denoted by $I(x|y)$. For y a constant, $i(x|y)$ and $I(x|y)$ will be denoted by $i(x)$, $I(x)$ respectively. If each of x, y has only finitely many values, the random variable

$$j(x, y) = i(x) + i(y) - i(x, y) = i(x) - i(x|y) = i(y) - i(y|x) \tag{5}$$

This paper was prepared with the partial support of the Office of Naval Research (Nonr-222-53).

is called the mutual information between x and y. Its expected value will be denoted by $J(x, y)$. For any stationary process $\{x_n, -\infty < n < \infty\}$ whose variables have only finitely many values, we write $I^*(x)$ for $I(x_0|x_{-1}, x_{-2}, \cdots)$.

An inequality of Shannon [8] and Feinstein [3] relates the existence of codes to the distribution of $j\{(a_1, \cdots, a_N), (b_1, \cdots, b_N)\} = j_N$, as follows.

Shannon-Feinstein inequality. For any integer D, and any number γ there are two functions f, g, where f maps $(1, \cdots, D)$ into the set U of sequences of length N of elements of A, g maps the set V of sequences of length N of elements of B into $(1, \cdots, D)$, for which

$$P\{g(b_1, \cdots, b_N) \neq d|(a_1, \cdots, a_N) = f(d)\} \leqq P\{j_N < \gamma\} + \frac{D}{2\gamma} \tag{6}$$

for $d = 1, \cdots, D$.

Note that the left side of (6) is independent of the distribution of $\{a_n\}$; it is simply the probability that, when an initial state for the channel is selected with a uniform distribution and the channel is then given the input sequence $f(d)$, the resulting output sequence $b_1, \cdots, b_N$ will be one for which $g(b_1, \cdots, b_N) \neq d$. The pair (f, g) can be considered as a code, with which we can transmit any of D messages over our channel in N transmission periods; when message d is presented to the sender, he gives the channel input sequence $f(d)$; the receiver then observes some output sequence v and decides that message $g(v)$ is intended.

For a given number $c \geqq 0$, let $D = [2^{Nc}]$ and write

$$\theta_N(c) = \min_{f,g} \max_{1 \leqq d \leqq D} P\{g(b_1, \cdots, b_N) \neq d|(a_1, \cdots, a_N) = f(d)\}. \tag{7}$$

Thus $\theta_N(c)$ is small if and only if we can transmit any binary sequence of length Nc, by using the channel for N periods, with small error probability.

Shannon [7] associated with each channel a number C, called the capacity of the channel, and proved that, for certain channels, $\theta_N(c) \rightarrow 0$ as $N \rightarrow \infty$ for every $c < C$ but not for any $c > C$. His original work has been considerably simplified and extended by several writers, including Shannon himself [8], McMillan [6], Feinstein [2], [3], Khinchin [5], and Wolfowitz [9], [10]. In particular, for certain channels, Wolfowitz has shown that $\theta_N(c) \rightarrow 1$ as $N \rightarrow \infty$ for every $c > C$.

For a certain class of finite state channels, the indecomposable channels defined below, the fact that $\theta_N(c) \rightarrow 0$ as $N \rightarrow \infty$ for $c < C$ was first proved by Breiman, Thomasian, and the writer [1]. We present in this paper a simpler proof of the slightly stronger fact that for these channels $\theta_N(c) \rightarrow 0$ exponentially: for any $c < C$ there are constants $\alpha > 0$, $\beta < 1$ for which, for all N,

$$\theta_N(c) < \alpha\beta^N. \tag{8}$$

The Shannon-Feinstein inequality reduces (8) at once to the study of the distribution of j_N for large N, as follows: if for a given c we can find an input sequence $\{a_n\}$ for which, for some $\alpha_1 > 0$, $\beta_1 < 1$,

$$P\{j_N \leqq Nc\} \leqq \alpha_1\beta_1^N \tag{9}$$

for all N, the Shannon-Feinstein inequality yields, for every $\epsilon > 0$ and all N,

$$\theta_N(c - \epsilon) \leqq \alpha_1\beta_1^N + 2^{-N\epsilon} \leqq \alpha_2\beta_2^N, \tag{10}$$

where $\alpha_2 = \alpha_1 + 1$, $\beta_2 = \max(\beta_1, 2^{-\epsilon})$. Thus our result (8) is implied by: for every $c < C$, there is an input sequence $\{a_N\}$ for which, for some $\alpha_1 > 0$, $\beta_1 < 1$, (9) holds.

We now define indecomposable channels and the number C. Let

$$\{x_n, n = 1, 2, \cdots\}$$

be any Markov process with a finite number R of states $r = 1, 2, \cdots, R$ and indecomposable transition matrix $\pi = \pi(r'|r) = P\{x_{n+1} = r'|x_n = r\}$. Let ϕ be any function from $(1, \cdots, R)$ to A, and let $a_n = \phi(x_n)$. We consider the source process $\{a_n\}$ as driving the channel, as described above. The process $\{z_n = (x_n, t_n)\}$ is then a Markov process, with transition matrix

$$m = m(r', t'|r, t) = \pi(r'|r)p[t'|t, \phi(r')].$$

If for every indecomposable π and every ϕ, the matrix m is also indecomposable, the finite state channel (A, B, T, p, ψ) is called *indecomposable.* There is then, for each m, a unique stationary Markov process $\{z_n^* = (x_n^*, t_n^*), -\infty < n < \infty\}$ with transition matrix m. Define $a_n^* = \phi(x_n^*)$, $b_n^* = \psi(t_n^*)$, $-\infty < n < \infty$, and let $J^*(\pi, \phi) = I^*(a) + I^*(b) - I^*(a, b)$. The number

$$C = \sup_{\pi,\phi} J^*(\pi, \phi), \tag{11}$$

where the sup is over all indecomposable π and all ϕ, is called the capacity of the channel. The main result of this paper is

THEOREM 1. *Let (A, B, T, p, ψ) be an indecomposable channel of capacity C. For every $c < C$ there is an input sequence $\{a_n, n = 1, 2, \cdots\}$ and there are numbers $\alpha > 0$, $\beta < 1$ for which, for all N,*

$$P\{j[(a_1, \cdots, a_N), (b_1, \cdots, b_N)] \leqq Nc\} \leqq \alpha\beta^N. \tag{12}$$

2. Preliminary reduction of theorem 1

To prove theorem 1, we choose π, ϕ for which $J^* = J^*(\pi, \phi) > c$. Let $z_n = (x_n, t_n)$, with $n = 0, 1, 2, \cdots$ be a Markov process with the transition matrix m and some initial distribution for which the initial distribution of t_0 is uniform. Let $a_n = \phi(x_n)$, $b_n = \psi(t_n)$, with $n = 1, 2, \cdots$. We shall show that the input sequence $\{a_n\}$ has the property specified by theorem 1. Let us write $u_N = (a_1, \cdots, a_N)$, $v_N = (b_1, \cdots, b_N)$. Since $j(u_N, v_N) = i(u_N) + i(v_N) - i(u_N, v_N)$ and $J^* = I^*(a) + I^*(b) - I^*(a, b)$, theorem 1 would be proved if we could bound the probability of each of the events

$$\begin{gathered} \{i(u_N) \leqq N(I^*(a) - \delta)\}, \\ \{i(v_N) \leqq N[I^*(b) - \delta]\}, \\ \{i(u_N, v_N) \geqq N[I^*(a, b) + \delta]\} \end{gathered} \tag{13}$$

above by $\alpha\beta^N$ for some $\alpha > 0$, $\beta < 1$, where $J^* - c = 3\delta$. That we can do this is the assertion of

THEOREM 2. *There are functions* $\alpha = \alpha(R, w, \epsilon)$, $\beta = \beta(R, w, \epsilon)$, *defined for* $R = 2, 3, \cdots, w > 0, \epsilon > 0$, *continuous in* w, ϵ, *increasing in* R *and decreasing in* w, ϵ *with* $a > 0$ *and* $0 < \beta < 1$ *such that, for any Markov process*

$$\{z_n, n = 1, 2, \cdots\}$$

with R *states* $r = 1, 2, \cdots, R$, *indecomposable transition matrix* $\pi = \pi(r'|r) = P\{z_{n+1} = r'|z_n = r\}$ *with smallest positive element* $\geqq w$ (π *may have some elements* 0) *and any function* ϕ *from* $1, \cdots, R$ *into a finite set* A,

$$P\{|i(a_1, \cdots, a_N) - NI^*(a)| \geqq N\epsilon\} \leqq \alpha(R, w, \epsilon)\beta^N(R, w, \epsilon) \tag{14}$$

for all N, *where* $a_n = \phi(z_n)$, *and* $I^*(a)$ *is as defined in section* 1, *namely if* $\{z_n^*, -\infty < n < \infty\}$ *is a stationary Markov process with transition matrix* π *and* $a_n^* = \phi(z_n^*)$, *then* $I^*(a) = I(a_0^*|a_{-1}^*, a_{-2}^*, \cdots)$.

Theorem 2 is a form of the equipartition theorem (Shannon [7], McMillan [6]) for "finitary" processes, with an exponential bound on the probability of exceptional sequences.

3. Proof of theorem 2 for ϕ the identity

For ϕ the identity function, so that $a_n(= z_n)$ is itself a Markov process, we have

$$i_N = i(z_1, \cdots, z_N) = -\log \lambda(z_1) - \sum_{n=1}^{N-1} \log \pi(z_{n+1}|z_n). \tag{15}$$

We use the following inequality of Katz and Thomasian [4].

Katz-Thomasian inequality. For $\{z_n\}$, w *as in theorem* 2 *and* ϕ *real-valued,* $P|\{|\phi(z_1) + \cdots + \phi(z_N) - N\mu| \geqq N\epsilon\} \leqq \alpha_1\beta_1^N$ *where*

$$\begin{aligned} \beta_1 &= \beta_1(R, w, \epsilon, M) = \exp - \left(\frac{w^{3R}\epsilon^2}{2^8M^2r^2}\right), \\ \alpha_1 &= \alpha_1(R, w, \epsilon, M) = \frac{8R}{w^R}\frac{1}{1 - \beta_1}, \\ M &= \max_{r'} \phi(r') - \min_r \phi(r), \end{aligned} \tag{16}$$

and $\mu = \sum \lambda(r)\phi(r)$, *where* λ *is the stationary distribution for* π.

We apply the Katz-Thomasian inequality to $z_n' = (z_n, z_{n+1})$, with $\phi' = -\log \pi(r'|r)$, so that $\mu = -\sum_{r,r'} \lambda(r)\pi(r'|r) \log \pi(r'|r) = I^*(z)$, and $M \leqq -\log w$ [we may exclude from z_n' the pairs r, r' with $\pi(r'|r) = 0$], obtaining

$$\begin{aligned} P\left\{\left|\sum_{n=1}^{N-1} \log \pi(z_{n+1}|z_n) + (N - 1)I^*(z)\right| \geqq (N - 1)\epsilon\right\} \\ \leqq \alpha_1(R^2, w, \epsilon, -\log w)\beta_1^{N-1}(R^2, w, \epsilon, -\log w) \\ = \alpha_2(R, w, \epsilon)\beta_2^N(R, w, \epsilon), \end{aligned} \tag{17}$$

say. Thus

$$P\{|i_N - NI^*(z) + \log \lambda(z_1) + I^*(z)| \geqq N\epsilon\} \leqq \alpha_2\beta_2^N. \tag{18}$$

Since $0 \leqq I^*(z) \leqq \log R$ and, for every $\delta > 0$,

$$P\{|\log \lambda(z_1)| \geqq N\delta\} = \sum_{r:\lambda(r) \leqq 2^{-N\delta}} \lambda(r) \leqq R2^{-N\delta} \tag{19}$$

we easily obtain $\alpha_3(R, w, \epsilon)$, $\beta_3(R, w, \epsilon)$ for which

$$P\{|i_N - NI^*(z)| \geqq N\epsilon\} \leqq \alpha_3\beta_3^N. \tag{20}$$

4. Proof of theorem 2, general case

We prove the general case by approximating the process $\{a_n\}$, in blocks, by a suitable Markov process, and using the fact that we have already proved the theorem for Markov processes. The idea is this: if, in addition to observing the $a_n = \phi(z_n)$ process we observe periodically, say every k trials, the current state of the underlying z_n process, the process now observed, with observations grouped in blocks of k, is a Markov process, so that all long sequences, except a set of exponentially small probability, have about the correct probability. We can choose k so large that (1) this correct probability is nearly the correct probability for the corresponding a sequence and (2) except with exponentially small probability, the probability of the actual a sequence will be near the probability of the actual observed sequence.

Thus choose a positive integer k, and let $x_1 = (a_1, \cdots, a_{k-1}, z_k)$, $x_2 = (a_{k+1}, \cdots, a_{2k-1}, z_{2k})$, $\cdots$, $x_n = (a_{(n-1)k+1}, \cdots, a_{nk-1}, z_{nk})$, $\cdots$. The $\{x_n\}$ process is Markov and, for k relatively prime to the period of $\{z_n\}$, is indecomposable. It has at most R^k states, and the smallest positive element in its transition matrix is at least w^k.

Thus, from the preceding section,

$$P\{|i(x_1, \cdots, x_N) - NI_k| \geqq \epsilon N\} \leqq \alpha_4\beta_4^N, \tag{21}$$

where

$$\begin{aligned} \alpha_4 &= \alpha_4(R, w, \epsilon, k) = \alpha_3(R^k, w^k, \epsilon), \\ \beta_4 &= \beta_4(R, w, \epsilon, k) = \beta_3(R^k, w^k, \epsilon), \end{aligned} \tag{22}$$

and $I_k = I^*(x)$. Now $kI^*(a) \leqq I_k \leqq kI^*(a) + \log R$ and $i(x_1, \cdots, x_N) = i(a_1, \cdots, a_{Nk}) + i(z_k, \cdots, z_{Nk}|a_1, \cdots, a_{Nk})$, which we write $i_N(x) = i_{Nk}(a) + i_N(z|a)$. Then

$$\begin{aligned} P\{i_{Nk}(a) \geqq Nk[I^*(a) + \epsilon]\} &\leqq P\left\{i_N(x) \geqq Nk\left(\frac{I_k - \log R}{k} + \epsilon\right)\right\} \\ &= P\{i_N(x) \geqq N[I_k + (k\epsilon - \log R)]\} \leqq \alpha_4\beta_4^N, \end{aligned} \tag{23}$$

provided $k\epsilon - \log R \geqq \epsilon$, that is, $k \geqq 1 + (1/\epsilon) \log R$. Similarly,

$$P\{i_{Nk}(a) \leqq Nk[I^*(a) - 4\epsilon]\} \leqq P\left\{i_N(x) - i_N(z|a) \leqq Nk\left(\frac{I_k}{k} - 4\epsilon\right)\right\} \tag{24}$$
$$\leqq P\{i_N(x) \leqq N(I_k - k\epsilon)\} + P\{i_N(z|a) \geqq 3Nk\epsilon\} = P_1 + P_2.$$

As above, $P_1 \leqq \alpha_4\beta_4^N$. To bound P_2, write $i_N(z) = i(z_k, z_{2k}, \cdots, z_{Nk})$. Then

$$P_2 \leqq P\{i_N(z) \geqq Nk\epsilon\} + P\{i_N(z|a) \geqq i_N(z) + Nk\epsilon\} = P_3 + P_4. \tag{25}$$

The process $\{z_{nk}, n = 1, 2, \cdots, k \text{ fixed}\}$ is a Markov process with R states and (k is relatively prime to the period of $\{z_n\}$) indecomposable transition matrix. For k so large that $k\epsilon \geqq \log R + \epsilon$, that is, $k \geqq 1 + (1/\epsilon) \log R$, we have $P_3 \leqq \alpha_3\beta_3^N$.

For P_4 we use

LEMMA 1. *For any two random variables a, z each with a finite set of values, and any $\delta \geqq 0$,*

$$P\{i(z|a) \geqq i(z) + \delta\} \leqq 2^{-\delta}. \tag{26}$$

PROOF. A pair (z_0, a_0) of values of z, a for which $i(z_0|a_0) \geqq (z_0) + \delta$ is one for which

$$\frac{P\{z = z_0|a = a_0\}}{P\{z = z_0\}} \leqq 2^{-\delta}, \tag{27}$$

that is, $P\{z = z_0, a = a_0\} \leqq 2^{-\delta}P\{z = z_0\}P\{a = a_0\}$. Summing over all pairs (z_0, a_0) for which the inequality is satisfied yields the lemma.

From the lemma, we obtain $P_4 \leqq 2^{-Nk\epsilon}$. Thus

$$P\{i_{Nk}(a) \leqq NkI^*(a) - 4\epsilon\} \leqq \alpha_5\beta_5^N, \tag{28}$$

where $\alpha_5 = \alpha_5(R, w, \epsilon, k) = \alpha_4 + \alpha_3 + 1$ and $\beta_5 = \max(\beta_4, \beta_3, 2^{-k\epsilon})$.

Combining (23) and (28) we obtain $\alpha_6(R, w, \epsilon, k)$, $\beta_6(R, w, \epsilon, k)$ for which

$$P\{|i_{Nk} - NkI^*(a)| \geqq Nk\epsilon\} \leqq \alpha_6\beta_6^N. \tag{29}$$

The block size k is still at our disposal, subject to $k \geqq 1 + (1/\epsilon) \log R$ and relatively prime to the period of $\{z_n\}$. We can find such a

$$k \leqq k^* = [R + 1 + (1/\epsilon) \log R]$$

and obtain, for this k,

$$P\{|i_{Nk} - NkI^*(a)| \geqq Nk\epsilon\} \leqq \alpha_7\beta_7^N, \tag{30}$$

where

$$\begin{aligned} \alpha_7 &= \alpha_7(R, w, \epsilon) = \alpha_6(R, w, \epsilon, k^*), \\ \beta_7 &= \beta_7(R, w, \epsilon) = \beta_6(R, w, \epsilon, k^*). \end{aligned} \tag{31}$$

Finally, for any n, say, $n = Nk + d$, with $0 \leqq d \leqq k - 1$, we have

$$i_n - n[I^*(a) + \epsilon] \leqq i_{(N+1)k} - (N + 1)k\left\{I^*(a) + \epsilon - \frac{I^*(a) + \epsilon}{N + 1}\right\} \tag{32}$$
$$\leqq i_{(N+1)k} - (N + 1)k\left\{I^*(a) + \frac{\epsilon}{2}\right\}$$

for

$$\frac{I^*(a) + \epsilon}{N + 1} \leqq \frac{\epsilon}{2}, \tag{33}$$

which, since $I^*(a) \leqq \log R$, will certainly hold for

$$N \geqq 2\left(\frac{\log R}{\epsilon} + 1\right) = N_0, \tag{34}$$

say, and similarly $i_n - n\{I^*(a) - \epsilon\} \geqq i_{Nk} - Nk\{I^*(a) - \epsilon/2\}$ for

$$N \geqq 2\left(\frac{\log R}{\epsilon} - 1\right). \tag{35}$$

Thus

$$P\{|i_n - nI^*(a)| \geqq n\epsilon\} \leqq \alpha_8\beta_8^{N-N_0}, \tag{36}$$

where

$$\begin{aligned} \alpha_8 &= \alpha_8(R, w, \epsilon) = 2\alpha_7\left(R, w, \frac{\epsilon}{2}\right) \\ \beta_8 &= \beta_8(R, w, \epsilon) = \beta_7\left(R, w, \frac{\epsilon}{2}\right). \end{aligned} \tag{37}$$

Finally, with

$$\alpha_9 = \alpha_8\beta_8^{-N_0}, \qquad \beta_9 = \beta_8^{1/k^*}, \tag{38}$$

we obtain

$$P\{|i_n - nI^*(a)| \geqq n\epsilon\} \leqq \alpha_9\beta_9^n$$

for all n, completing the proof.

REFERENCES

[1] D. Blackwell, L. Breiman, and A. J. Thomasian, "Proof of Shannon's transmission theorem for finite state indecomposable channels," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 1209–1220.

[2] A. Feinstein, "A new basic theorem of information theory," *IRE Transactions P.G.I.T.*, 1954, pp. 2–22.

[3] ———, *Foundations of Information Theory*, New York, McGraw-Hill, 1958.

[4] M. Katz, Jr., and A. J. Thomasian, "An exponential bound for functions of a Markov chain," *Ann. Math. Statist.*, Vol. 31 (1960), pp. 470–474.

[5] A. I. Khinchin, *Mathematical Foundations of Information Theory*, New York, Dover, 1957.

[6] B. McMillan, "The basic theorems of information theory," *Ann. Math. Statist.*, Vol. 24 (1953), pp. 196–219.

[7] C. E. Shannon, "Mathematical theory of communication," *Bell System Tech. J.*, Vol. 27 (1948), pp. 379–423.

[8] ———, "Certain results in coding theory for noisy channels," *Information and Control*, Vol. 1 (1957), pp. 6–25.

[9] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math.*, Vol. 1 (1957), pp. 591–606.

[10] ———, "Strong converse of the coding theorem for semicontinuous channels," *Illinois J. Math.*, Vol. 3 (1959), pp. 477–489.