The Annals of Probability
1982, Vol. 10, No. 1, 188–198

CHANNEL ENTROPY AND PRIMITIVE APPROXIMATION[1]

DAVID L. NEUHOFF AND PAUL C. SHIELDS

University of Michigan and University of Toledo

For stationary channels, channel entropy is defined as the supremum of the conditional output entropy over all stationary input sources. If the channel is $\bar{d}$-continuous and conditionally almost block independent, channel entropy gives an upper bound to the output rate of the channel. Furthermore, the channel can be approximated arbitrarily well in the $\bar{d}$-metric by any primitive channel whose noise source entropy exceeds channel entropy and cannot be so approximated if the noise source entropy is less than channel entropy.

I. Introduction. In this paper we strengthen our results on simulation by primitive channels [5]. Our earlier proof required that the noise source be continuously distributed. Here we remove this restriction and obtain exact bounds on how much noise entropy is needed for accurate simulation.

A primitive channel operates as follows. An independent, identically distributed (i.i.d.) noise sequence is generated which is independent of the input sequence. The output is obtained by passing the noise sequence together with the input sequence through a finite sliding-block encoder (that is, a nonlinear time-invariant filter). The channels which can be simulated arbitrarily well by primitive channels are precisely those channels which satisfy conditions for input and output memory decay called $\bar{d}$-continuity and conditional almost block independence (CABI).

Our earlier results required that the noise source be continuously distributed. This allowed us to construct partitions with arbitrary distribution, a necessary technique in our proofs. We show in this paper how we can build sufficiently good approximations of the desired distributions if the atoms of the noise source space are sufficiently small.

Our technique raises the question of exactly how much entropy is needed in the noise source for accurate simulation. Our answer to this question uses a concept we call channel entropy. This is the supremum of the conditional output entropies attainable from stationary input sources. We show that if the noise entropy exceeds channel entropy, then accurate simulation is possible, while accurate simulation is not possible if the noise entropy is less than channel entropy.

2. Notation and terminology. We make use of the notation and terminology of our earlier paper [5] which we summarize here for ease of reference. A process with alphabet X is a Borel probability measure μ on the doubly infinite sequence space $X_{-\infty}^{\infty}$. The symbol X or the sequence of coordinate functions $\{X_n\}$ may also be used to denote this process. A process is stationary if $\mu T = \mu T^{-1} = \mu$ where T is the shift on $X_{-\infty}^{\infty}$. We will use x_m^n to denote the sequence $\{x_m, x_{m+1}, \cdots, x_n\}$, X_m^n to denote the set of all such sequences and μ_m^n to denote the measure on X_m^n induced by μ. If $m = 1$ we will write μ^n rather than μ_1^n, while if n is understood we will write μ instead of μ^n.

If μ and $\bar{\mu}$ are processes with the same alphabet X then $\mu_m^n \vee \bar{\mu}_m^n$ denotes the collection of all measures λ on $(X \times X)_m^n$ which have μ_m^n and $\bar{\mu}_m^n$ as respective marginals. If x and $\bar{x}$ are two alphabet symbols we define $d(x, \bar{x})$ to be 1 if $x \neq \bar{x}$ and 0 if $x = \bar{x}$. We then define

$$d_n(x_1^n, \bar{x}_1^n) = \frac{1}{n} \sum_{i=1}^n d(x_i, \bar{x}_i).$$

Received August 1980.
[1] Work supported in part by AFOSR grant AFOSR-80-0054 and NSF grant MCS7807739-A02.
AMS 1970 *subject classifications.* Primary 94A20, secondary 28A70.
Key words and phrases. Channel, entropy, $\bar{d}$-distance.

The $\bar{d}$-distance is defined by

$$\bar{d}_n(\mu, \bar{\mu}) = \inf_{\lambda \in \mu^n \vee \bar{\mu}^n} E_\lambda(d_n(X_1^n, \bar{X}_1^n))$$

and

$$\bar{d}(\mu, \bar{\mu}) = \limsup d_n(\mu, \bar{\mu}).$$

We will also write $\bar{d}_n(X, \bar{X})$ for $\bar{d}_n(\mu, \bar{\mu})$.

Entropy is defined by

$$H(X) = \limsup H_n(X)$$

where

$$H_n(X) = -\frac{1}{n} \sum_{x_1^n} \mu(x_1^n) \log \mu(X_1^n).$$

The process $\bar{\mu}$ is called an independent N-blocking of μ if for each integer i

(i) $$\bar{d}_N(\mu T^{iN}, \bar{\mu} T^{iN}) = 0$$

(ii) $$\bar{X}_{iN+1}^{(i+1)N} \text{ is independent of } \bar{X}_{-\infty}^{iN}.$$

A channel with input alphabet X and output alphabet Y is a measurable family ν_x, $x \in X_\infty^\infty$, of Borel probability measures on Y_∞^∞. If x is an input sequence we denote the output process by Y/x, while if μ is an input process we denote the input/output process by $\mu\nu$ or (X, Y) and the conditional output process by ν/μ or Y/X. We assume throughout this paper that a channel is stationary, that is

$$\nu_{Tx}(TE) = \nu_x(E)$$

for all Borel sets E.

A channel is $\bar{d}$-continuous if for each $\varepsilon > 0$ there is an integer N such that

$$\bar{d}_n(Y/x, Y/\bar{x}) < \varepsilon \quad \text{if} \quad n \geq N \quad \text{and} \quad x_1^n = \bar{x}_1^n.$$

A channel is conditionally almost block independent (CABI) if, given $\varepsilon > 0$, there is an N such that, if $n \geq N$, then there is an M_n such that if x is an input sequence and $\bar{Y}$ is the independent n-blocking of Y/x then

$$\bar{d}_m(Y/x, \bar{Y}) < \varepsilon \quad \text{if} \quad m \geq M_n.$$

The $\bar{d}$-distance between two channels ν and $\hat{\nu}$ with the same input and output alphabets is defined by

$$\bar{d}(\nu, \hat{\nu}) = \limsup_n \sup_x \bar{d}_n(\nu_x, \hat{\nu}_x).$$

A channel is primitive with noise source Z if there is an i.i.d. process $\{Z_n\}$, a nonnegative integer ω and a function f: $(X \times Z)_{-\omega}^{\omega} \to Y$ such that for all n

$$Y_n = f(x_{n-\omega}^{n+\omega}, Z_{n-\omega}^{n+\omega}).$$

3. Statements of Results. In our earlier paper [5], we proved the following result.

THEOREM 1. *A channel ν is $\bar{d}$-continuous and* CABI *if and only if it is the $\bar{d}$-limit of a sequence of primitive channels with an independent uniformly distributed noise source $\{Z_n\}$.*

In this paper we show that $\{Z_n\}$ can be any i.i.d. process with sufficiently large entropy and we obtain a precise bound for the allowable noise entropy, which we call the channel entropy. If ν is a channel, we define its *channel entropy* $H(\nu)$ by

$$H(\nu) = \sup_\mu H(\nu/\mu)$$

where the supremum is over all stationary input processes μ and $H(\nu/\mu)$ is the conditional output entropy $H(\mu\nu) - H(\mu)$.

Our principle results are the following

THEOREM 2. *Given $\varepsilon > 0$ and finite input and output alphabets there is a $\delta > 0$ such that if*

$$H(Z) < H(\nu) - \delta$$

then $$\bar{d}(\nu, \nu) \geq \varepsilon$$

for any primitive channel $\bar{\nu}$ with noise source Z.

THEOREM 3. *If ν is $\bar{d}$-continuous and* CABI *and $H(Z) > H(\nu)$, then ν is the $\bar{d}$-limit of primitive channels with noise source $\{Z_n\}$.*

These two theorems will be proved in Section 5 after we show in Section 4 how entropy is related to information rate.

4. Channel entropy and rate. Our goal in this section is to show how channel entropy is related to the rate at which a $\bar{d}$-continuous, CABI channel produces output information. If $\{Y_n\}$ is a process, we defined its *ε-N rate* by

$$R_N(Y, \varepsilon) = \frac{1}{N} \log \inf\{|G|: G \subseteq Y_1^N, \quad \text{Prob}(Y_1^N \varepsilon G) > 1 - \varepsilon\}$$

where $|G|$ denotes cardinality. Our principle result is

THEOREM 4. *If ν is $\bar{d}$-continuous and* CABI *and $H(\nu) < \gamma$, then for each $\varepsilon > 0$ there is an integer N such that for all input sequences x*

$$R_n(Y/x, \varepsilon) \leq \gamma \qquad \text{if} \quad n \geq N.$$

This theorem will be a consequence of a series of lemmas which connect rate to the $\bar{d}$-distance (Lemma 1), entropy to the $\bar{d}$-distance (Lemma 2), rate and entropy for independent processes (Lemma 3), then connect rate and entropy for the conditional output Y/x (Lemma 4), and finally connect entropy for Y/x to channel entropy (Lemma 5). Theorem 4 then follows immediately from lemmas 4 and 5.

LEMMA 1. *If K is the alphabet size and $\varepsilon > 0$, there is a $\bar{\delta} > 0$ and an N such that if $n \geq N$ and $\bar{d}_n(Y, \bar{Y}) < \delta^2$, where $0 < \delta \leq \bar{\delta}$, then then*

$$R_n(Y, 2\delta) \leq R_n(\bar{Y}, \delta) + \varepsilon.$$

PROOF. We first choose $\bar{\delta} > 0$ so that

(i) $$\bar{\delta} \log(k - 1) + h(\bar{\delta}) < \varepsilon/2$$

where $h(\bar{\delta}) = -\bar{\delta} \log \bar{\delta} - (1 - \bar{\delta}) \log (1 - \bar{\delta})$, then choose N so that if $n \geq N$ then

(ii) $$\textstyle\sum_{k \leq \bar{\delta}_n} \binom{n}{k} \leq 2^{n(h(\bar{\delta}) + \varepsilon/2)}.$$

Now we suppose $n \geq N$ and $\bar{d}_n(Y, \bar{Y}) < \delta^2$, where $0 < \delta < \bar{\delta}$ and choose $\lambda \varepsilon \mu^n \vee \bar{\mu}^n$ such that

$$E_\lambda(d_n(y, \bar{y})) < \delta^2.$$

Next we let A_n be the set of all pairs $(y_1^n, \bar{y}_1^n)$ such that $d_n(y_1^n, \bar{y}_1^n) < \delta$ and note that

(iii) $$\lambda(A_n) \geq 1 - \delta.$$

Finally we let $\bar{G}_n$ be a set of sequences $\bar{y}_1^n$ such that

(iv) $$\bar{\mu}(\bar{G}_n) > 1 - \delta \text{ and } \frac{\log|\bar{G}_n|}{n} = R_n(\bar{Y}, \delta)$$

then set

$$G_n = \{y_1^n: \exists \bar{y}_1^n \in \bar{G}_n \text{ for which } d_n(y_1^n, \bar{y}_1^n) < \delta\}.$$

Conditions (iii) and (iv) guarantee that

$$\mu(G_n) \geq 1 - 2\delta$$

so that

(v) $$R_n(Y, 2\delta) \leq \frac{\log |G_n|}{n}.$$

Note, however, that a sequence in G_n is obtained by changing a sequence in $\bar{G}_n$ in no more than δn places and there are $k - 1$ choices for each change.

Therefore, $$|G_n| \leq |\bar{G}_n| \cdot (k-1)^{\delta n} \cdot \sum_{k \leq \delta n} \binom{n}{k}.$$

We take logarithms, divide by n and use (i) and (ii) and the assumption that $\delta \leq \bar{\delta}$ to obtain

$$\frac{\log |G_n|}{n} \leq \frac{\log |\bar{G}_n|}{n} + \varepsilon.$$

The lemma follows from this and (v).

LEMMA 2. *If K is the alphabet size and $\varepsilon > 0$, there is a $\delta > 0$ so that for all $n \geq 1$, if $\bar{d}_n(Y, \bar{Y}) < \delta$ then $|H_n(Y) - H_n(\bar{Y})| < \varepsilon$.*

PROOF. We choose $\delta > 0$ so that

(i) $$\delta \log(K-1) + h(\delta) < \varepsilon/2$$

where $h(\delta) = -\delta \log \delta - (1-\delta) \log (1-\delta)$. If $\bar{d}(Y, \bar{Y}) < \delta$ we choose $\lambda \in \mu_1^n \vee \bar{\mu}_1^n$ so that

$$E_\lambda(d_n(y_1^n, \bar{y}_1^n)) < \delta$$

then let $(Y, \bar{Y})$ denote some process with measure $\tilde{\lambda}$ such that $\tilde{\lambda}^n = \lambda$. As shown in Gallager [2, page 79], we then have

$$\frac{1}{n} H(Y^n | \bar{Y}^n) \leq h(\delta) + \delta \log(K-1) < \varepsilon.$$

Since $$H(Y^n) + H(\bar{Y}^n | Y^n) = H(\bar{Y}^n) + H(Y^n | \bar{Y}^n)$$

we have $$\frac{1}{n} [H(Y^n) - H(\bar{Y}^n)] \leq \frac{1}{n} H(Y^n | \bar{Y}^n) < \varepsilon.$$

Reversing the roles of Y^n and $\bar{Y}^n$ completes the proof of this lemma.

LEMMA 3. *Let Δ be a finite set of distributions defined on the finite set Y. Given $\varepsilon > 0$ there is an N such that, if $\{Y_m\}$ is a sequence of independent random variables each with distribution in Δ, then*

$$R_n(Y, \varepsilon) \leq H_n(Y) + \varepsilon \qquad \text{for} \quad n \geq N.$$

PROOF. Without loss of generality, we can suppose that Δ is the collection of probability vectors $\pi(i)$, $1 \leq i \leq D$. Let α be the least positive number contained in the set $\{\pi^{(i)}(b): 1 \leq i \leq D, b \in Y\}$.

We first use the asymptotic equipartion property to choose N_o so large that if, for some i, $\{Z_n^{(i)}\}$ is an i.i.d. sequence with distribution $\pi^{(i)}$, then for $n \geq N_o$

(i) $$\text{Prob}(Z_j^{(i)} = z_j,\ i \le j \le n) \ge 2^{-n(H(Z^{(i)})+\varepsilon/2)}$$

except for a set $B_n^{(i)}$ of sequences z_1^n of total probability at most ε/D.

If $\{Y_n\}$ is a sequence of independent random variables each with distribution in Δ, then for each integer N we can rearrange the sequence Y_1^N into blocks

$$Z_1^{(1)}, Z_2^{(1)}, \cdots, Z_{N_1}^{(1)}, Z_1^{(2)}, \cdots, Z_{N_2}^{(2)}, \cdots, Z_1^{(D)}, Z_2^{(D)}, \cdots, Z_{N_D}^{(D)}$$

so that $Z_j^{(i)}$ has distribution $\pi^{(i)}$. Thus there is a nonnegative integer K and a labeling of Δ so that

$$N_i \ge N_o \quad \text{if} \quad i \le K$$

$$N_i < N_o \quad \text{if} \quad i > K.$$

Let B_N be the set of all sequences y_1^N such that

$$(Z_1^{(i)}, Z_2^{(i)}, \cdots, Z_{N_i}^{(i)}) \in B_{N_i}^{(i)}$$

for some $i \le K$. We then have

$$\text{Prob}\ (Y_1^N \in B_N) \le \textstyle\sum_{i\le K} \text{Prob}\ (B_{N_i}^{(i)}) \le \varepsilon.$$

Furthermore if $y_1^N \notin B_N$ then (i) gives

(ii) $$\log \text{Prob}\ (Y_1^N = y_1^N) \ge -\textstyle\sum_{i\le K} N_i(H(Z^{(i)}) + \varepsilon/2) + \log \alpha \sum_{i>K} N_i$$

while the independence of the Y process gives

$$H_N(Y) = \frac{1}{N}\textstyle\sum_{i\le K} N_i H(Z_1^{(i)}) + \frac{1}{N}\sum_{i>K} N_i H(Z_1^{(i)}).$$

Note that the definition of K gives

$$\textstyle\sum_{i>K} N_i \le DN_o$$

so that, if N is large enough, then for $y_1^N \notin B_N$ we must have

$$\text{Prob}(Y_1^N = y_1^N) \ge 2^{-N(H_N(Y)+\varepsilon)}.$$

Since $\text{Prob}(B_N) \le \varepsilon$ it then follows that

$$R_N(Y, \varepsilon) \le H_N(Y) + \varepsilon$$

which proves the lemma.

LEMMA 4. *If ν is a $\bar{d}$-continuous,* CABI *channel and $\varepsilon > 0$ then there is a $\delta > 0$ and an N such that for all x*

$$R_N(Y/x, \delta) \le H_n(Y/x) + \varepsilon \text{ if } n \ge N.$$

PROOF. We will show that a small $\bar{d}$ change in Y/x will produce a process $\bar{Y}$ whose rate is close to its entropy. This, together with Lemmas 1 and 2 will prove Lemma 4. Towards this end, we fix $\alpha > 0$ and use the $\bar{d}$-continuity and CABI properties to choose N and M so that the following hold

(i) $$\bar{d}(Y/x, Y/\bar{x}) < \frac{\alpha}{2} \quad \text{if} \quad x_1^N = \bar{x}_1^N.$$

(ii) $\bar{d}_m(Y/x, \bar{Y}) < \alpha/2$ if $m \ge M$ and $\bar{Y}$ is an independent N-blocking of Y/x.

Next we choose one sequence $\bar{x} = \bar{x}(a_1^N)$ for each $a_1^N \in X^N$ so that $\bar{x}_1^N = a_1^N$ and let Δ be the resulting family of output distributions $\nu(Y^N \mid \bar{x}(a_1^N))$. We then fix x and let $\bar{Y}$ be the process defined by the two properties

(iii) $\bar{Y}_{iN+1}^{i(N+1)}$ has the distribution of $Y^N/\bar{x}(a_1^N)$ if $x_{iN+1}^{i(N+1)} = a_1^N$.

(iv) $\bar{Y}_{iN+1}^{i(N+1)}$ is dependent of $\bar{Y}_{-\infty}^{iN}$.

Conditions (i) and (ii) then guarantee that

$$\bar{d}_m(Y/x, \bar{Y}) < \alpha \quad \text{if} \quad m \geq M.$$

Finally we apply Lemma 3 to choose $N_1 \geq M$ so that if $n \geq N_1$ then

$$R_n(\bar{Y}, \alpha) \leq H_n(\bar{Y}) + \alpha.$$

In summary, we have shown that, given $\alpha > 0$, there is an N_1 such that for every x there is a process $\bar{Y}$ such that for $n \geq N$, the following hold.

(v) $$\bar{d}_n(Y/x, \bar{Y}) < \alpha.$$

(vi) $$R_n(Y, \alpha) \leq H_n(\bar{Y}) + \alpha.$$

Suitable choice of α combined with Lemmas 1 and 2 then imply Lemma 4.

LEMMA 5. *Let ν be $\bar{d}$-continuous and* CABI. *Given $\varepsilon > 0$, there is an N such that for all x*

$$H_n(Y \mid x) \leq H(\nu) + \varepsilon \quad \text{if} \quad n \geq N.$$

PROOF. We first use Lemma 2 to choose $\delta > 0$ so that if $\bar{d}_n(Y, \bar{Y}) < \delta$ then $H_n(Y) \leq H_n(\bar{Y}) + \varepsilon/2$. We then use the $\bar{d}$-continuity and CABI properties to choose N so that if $n \geq N$, then

(i) $$\bar{d}_n(Y/x, Y/\bar{x}) < \delta \quad \text{if} \quad x_1^n = \bar{x}_1^n$$

and there is an M_n so that if $m \geq M_n$, then

(ii) $$\bar{d}_m(Y/x, \bar{Y}) < \delta/2 \quad \text{if } \bar{Y} \text{ is an independent } n\text{-blocking of } Y/x.$$

Next we fix $n \geq N$ and an input sequence x and define $\bar{x}$ by

$$\bar{x}_{in+j} = x_j \quad \text{for} \quad 1 \leq j \leq n \quad \text{for all} \quad i.$$

Our choice of δ and the $\bar{d}$-continuity property (i) guarantee that

$$H_n(Y/x) \leq H_n(Y/\bar{x}) + \varepsilon/2.$$

Our goal now is to show that

(iii) $$H_n(Y/\bar{x}) \leq H(\nu) + \varepsilon/2$$

which combines with the preceding inequality to establish the lemma. Towards this end, we let $\bar{Y}$ be the independent n-blocking of $Y/\bar{x}$, and for $1 \leq i \leq n$ put $Y^{(i)} = Y/T^i\bar{x}$ and $\bar{Y}^{(i)} = T^i\bar{Y}$. If $m \geq M_n + 2n$ we can fit $Y_j^{(i)}$ and $\bar{Y}_j^{(i)}$ for $1 \leq j \leq m$ by first using the CABI property (ii) to fit for $n - i \leq j \leq m - i$. Thus we must have

$$\bar{d}_m(Y^{(i)}, \bar{Y}^{(i)}) < \frac{\delta}{2} + \frac{2n}{m}$$

so that if m also satisfies $m \geq 4n/\delta$ we then have

$$\bar{d}_m(Y^{(i)}, \bar{Y}^{(i)}) < \delta$$

and hence our choice of δ gives

(iv) $$| H_m(Y^{(i)}) - H_m(\bar{Y}^{(i)})| \leq \varepsilon/2.$$

Let μ be the measure which gives mass $1/n$ to each of the sequences $T\bar{x}, T^2\bar{x}, \cdots, T^n\bar{x}$

and let X be the process defined by μ. We then have

$$H_m(Y \mid X) = \frac{1}{n} \sum_{i=1}^{n} H_m(Y^{(i)})$$

so that (iv) gives

$$| H_m(Y \mid X) - \frac{1}{n} \sum_{i=1}^{n} H_m(\bar{Y}^{(i)})| \leq \varepsilon/2$$

for m sufficiently large. Note, however, that block independence gives

$$\lim_{m\to\infty} H_m(\bar{Y}^{(i)}) = H_n(\bar{Y}) = H_n(Y \mid \bar{x})$$

so that

$$H_n(Y \mid \bar{x}) \leq H(Y \mid X) + \varepsilon/2$$

which proves (iii) and completes our proof of the lemma.

5. Approximation by primitive channels. In this section we shall prove the two basic theorems (Theorems 2 and 3) about approximation by primitive channels. We first discuss the simpler result, which is the fact that accurate simulation is not possible if the noise source entropy is too low (Theorem 2). We will need the following lemma.

LEMMA 6. *Given $\varepsilon > 0$ and the input and output alphabet sizes, there is a $\delta > 0$ such that*

$$\text{if } \bar{d}(\nu, \bar{\nu}) < \delta \quad \text{then} \quad | H(\nu) - H(\bar{\nu})| < \varepsilon.$$

PROOF. Let K be the product of the input and output alphabet sizes, then use Lemma 2 to choose $\delta > 0$ so that for all n and all proceses Y, $\bar{Y}$ of alphabet size K

(i) $$| H_n(Y) - H_n(\bar{Y})| \leq \varepsilon/2 \quad \text{if} \quad \bar{d}_n(Y, \bar{Y}) \leq \delta.$$

If $\bar{d}(\nu, \bar{\nu}) < \delta$ we can choose N so that for all input sequences x

$$\bar{d}_n(\nu_x, \bar{\nu}_x) \leq \delta \quad \text{if} \quad n \geq N.$$

If μ is any stationary input process we can apply Proposition A.1 of our earlier paper [4] to obtain

$$\bar{d}_n(\mu\nu, \mu\bar{\nu}) \leq \sup_x \bar{d}_n(\nu_x, \bar{\nu}_x) \leq \delta \quad \text{if} \quad n \geq N.$$

This combines with (i) to give

$$| H_n(\mu\nu) - \mathrm{H}_n(\mu\bar{\nu})| \leq \varepsilon/2$$

if $n \geq N$. We then let $n \to \infty$ to obtain

$$| H(\mu\nu) - H(\mu\bar{\nu})| \leq \varepsilon/2$$

and hence $$| H(\nu/\mu) - H(\bar{\nu}/\mu)| \leq \varepsilon/2$$

since $H(\nu/\mu) = H(\mu\nu) - H(\mu)$. Finally we take the supremum over μ to obtain the desired result.

We now prove Theorem 2, which for ease of reference we restate here.

THEOREM 2. *Given $\varepsilon > 0$ and the input and output alphabet sizes, there is a $\delta > 0$ such that if*

$$H(Z) < H(\nu) - \delta$$

then $$\bar{d}(\nu, \bar{\nu}) \geq \varepsilon$$

for any primitive channel $\bar{\nu}$ with noise source Z.

PROOF. If $\bar{\nu}$ is a primitive channel with noise source Z let X be a stationary input source. The output Y is then a coding of the pair process (X, Z) so we must have

$$H(X, Y) \leq H(X, Z)$$

and hence, since Z is independent of X, we have

$$H(Y \mid X) = H(X, Y) - H(X) \leq H(X, Z) - H(X) = H(Z).$$

We then take the supremum over all X to obtain

$$H(\bar{\nu}) \leq H(Z).$$

Theorem 2 now follows easily from Lemma 6.

Now we turn to the proof of Theorem 3 which we restate here.

THEOREM 3. *If ν is $\bar{d}$-continuous and* CABI *and $H(Z) > H(\nu)$ then ν is the d-limit of primitive channels with noise source $\{Z_n\}$.*

The basic idea in the proof of this theorem is as follows. Theorem 4 guarantees that the rate $R_n(Y/x, \varepsilon)$ will be dominated by $H(Z)$ for sufficiently large m and hence there will be many more Z_1^n sequences than $(Y/x)_1^n$ sequences. We can therefore partition Z_1^n to approximate the distribution of $(Y/x)_1^n$ (Lemma 8). This provides a block coding of (x, Z) which approximates Y/x and we can then apply the technique of our earlier paper to make this code stationary. We first discuss two lemmas.

The distribution distance $\operatorname{dist}_n(Y, \bar{Y})$ is defined as follows

$$\operatorname{dist}_n(Y, \bar{Y}) = \textstyle\sum_{y_1^n} \mid \operatorname{Prob}(Y_1^n = y_1^n) - \operatorname{Prob}(\bar{Y}_1^n = y_1^n)|.$$

The following was proved by Gray and Ornstein [3].

LEMMA 7. $\bar{d}_n(Y, \bar{Y}) \leq \frac{1}{2} \operatorname{dist}_n(Y, \bar{Y})$.

Our next lemma provides our basic block code.

LEMMA 8. *Suppose $\{Z_n\}$ is* i.i.d., *$\alpha < H(Z)$ and $\varepsilon > 0$. There is an integer N such that if $n \geq N$ and $R_n(Y, \varepsilon) \leq \alpha$ then there is a mapping $\phi : Z^n \to Y^n$ such that*

$$\bar{d}_n(Y, \bar{Y}) \leq 2\varepsilon$$

where $$\bar{Y}_1^n = \phi(Y_1^n).$$

PROOF. The theorem is trivial if $\varepsilon \geq \frac{1}{2}$ if $\alpha \leq 0$ or if the Y alphabet size K is less than 2. Thus we can assume that $0 < \varepsilon < \frac{1}{2}$, $\alpha > 0$ and $K \geq 2$.

Let us put $H \frac{1}{2}(Z)$, $\delta = (H - \alpha)/2$ and

$$G_n(Z) = \{z_1^n : \operatorname{Prob}(Z_1^n = z_1^n) < 2^{-n(H-\delta)}\}.$$

We use the Shannon-McMillan-Breiman theorem to choose N so large that if $n \geq N$ then

(i) $$\operatorname{Prob}(Z_1^n \,\varepsilon\, G_n(Z)) \geq 1 - \varepsilon.$$

We also require that

(ii) $$2^{-N\delta} < \varepsilon \quad \text{and} \quad 2^{-2N} < \varepsilon.$$

Let us fix $n \geq N$ and suppose $R = R_n(Y, \varepsilon) \leq \alpha$. We can then find a set $G_n(Y)$ of sequences y_1^n such that

(iii) $$| G_n(Y)| = 2^{Rn} \quad \text{and} \quad \text{Prob}(Y_1^n \, \varepsilon \, G_n(Y)) \geq 1 - \varepsilon.$$

To define our block code ϕ we first list the members of $G_n(Y)$ in some order, say $y^{(1)}$, $y^{(2)}, \cdots y^{(M)}$, where $M = 2^{nR}$. We also choose a sequence y_1^n which is not in $G_n(Y)$ and call it $y^{(0)}$. The existence of $y^{(0)}$ is guaranteed since $K \geq 2$ and $2^{-2n} < \varepsilon$.

We now assign as many sequences of $G_n(Z)$ to $y^{(1)}$ as we possibly can without exceeding the probability of $y^{(1)}$, then assign as many as possible to $y^{(2)}$ without exceeding the probability of $y^{(2)}$, etc. In this manner we obtain disjoint subsets $C_1, C_2, \cdots, C_M$ of $G_n(Z)$ such that for $1 \leq i \leq M$

(a) $$\text{Prob}(Z_1^n \in C_i) \leq \text{Prob}(Y_1^n = y^{(i)})$$

and the inequality is violated if we add to C_i any member of $G_n(Z) - \cup_{i=1}^M C_i$. We also let $C_0 = Z^n - \cup_{i=1}^M C_i$ and define ϕ by

$$\phi(z_1^n) = y^{(i)} \quad \text{if} \quad z_1^n \in C_i$$

then define $\bar{Y}_1^n = \phi(Z_1^n)$.

Let us put

$$p_i = \text{Prob}(Z_1^n \in C_i), \qquad q_i = \text{Prob}(Y_1^n = y^{(i)})$$

so that (a) now reads as

(a) $$p_i \leq q_i, \qquad 1 \leq i \leq M.$$

If $G_n(Z) = \cup_{i=1}^M C_i$, then we must have

$$\sum q_i \geq \sum p_i \geq 1 - \varepsilon$$

and $$\sum_{i=1}^M | p_i - q_i | = \sum_{i=1}^M q_i - \sum_{i=1}^M p_i \leq 1 - (1 - \varepsilon) = \varepsilon.$$

Since at most ε of the Z space and ε of the Y space are left over, we must therefore have

$$\text{dist}_n(Y, \bar{Y}) \leq 3\varepsilon.$$

If $G_n(z) \neq \cup_{i=1}^M C_i$, then since we have filled each C_i as much as possible and each Z_1^n in $G_n(Z)$ has probability less than $2^{-n(H+\delta)}$, we must have

$$p_i \geq q_i - 2^{-n(H+\delta)}.$$

Thus $$\sum_{i=1}^M | p_i - q_i | \leq M 2^{-n(H+\delta)} < \varepsilon.$$

Furthermore

$$\sum_{i=1}^M p_i \geq \sum_{i=1}^M q_i - M 2^{-n(H+\delta)} \geq 1 - 2\varepsilon,$$

so at most ε of the Y-space and 2ε of the z-space are left over, hence

$$\text{dist}_n(Y, \bar{Y}) \leq 4\varepsilon.$$

This proves the lemma, for in any case we have $\text{dist}_n(Y, \bar{Y}) \leq 4\varepsilon$ so that Lemma 7 gives

$$d_n(Y, \bar{Y}) \leq 2\varepsilon.$$

Our next lemma shows how to construct primitive approximations when the noise source Z is a direct product $(\bar{Z}, \bar{R})$ of two i.i.d. sources.

LEMMA 9. *Suppose $Z_n = (\bar{Z}, \bar{R}_n)$ is an* i.i.d. *process with $\{\bar{Z}_n\}$ independent of $\bar{R}_n$ and ν is a $\bar{d}$-continuous,* CABI *channel for which $H(\nu) < H(\bar{Z})$. Given $\varepsilon > 0$ there is a primitive channel $\bar{\nu}$ with noise source Z such that $\bar{d}(\nu, \bar{\nu}) \leq \varepsilon$.*

PROOF. Our proof is essentially the same as the proof in our earlier paper for the case when $\bar{Z}_n$ is uniformly distributed on [0, 1], except that here we use Lemma 8 to provide the desired block coding. We sketch the ideas here and refer to [4] for the details.

The steps in our proof are as follows.

Step 1. For each $x_1^N \in X_1^N$, choose an infinite sequence $\tilde{x}$ such that $\tilde{x}_1^N = x_1^N$ and put $\nu_{x_1^N} = \nu_{\tilde{x}}^N$. If $\alpha > 0$ and N is sufficiently large we can use $\bar{d}$-continuity to guarantee that

$$\bar{d}_N(\nu_{\bar{x}}, \nu_{x_1^N}) \leq \alpha \quad \text{if} \quad \bar{x}_1^N = x_1^N.$$

Step 2. Put $\delta = (H(\bar{Z}) = H(\nu))/2$. If β is a positive number and N is sufficiently large, then we can use Theorem 4 to guarantee that for all x_1^N

$$R_N(\nu_{x_1^N}, \beta) \leq H(\bar{Z}) - \delta,$$

then use Lemma 8 to guarantee that there is a block code $\phi_{x_1^N}$ of $\bar{Z}^N$ into Y^N such that

$$\bar{d}_N(\nu_{x_1^N}, \tilde{\nu}_{x_1^N}) \leq 2\beta$$

where $\tilde{\nu}_{x_1^N}$ in the distribution of $\phi_{x_1^N}(Z_1^N)$.

Step 3. Construct a finite sliding block coding $\{\tilde{R}_n\}$ of $\{R_n\}$ for which

$$\text{Prob}(\tilde{R}_1^N = 0, 1, 1, 1, \cdots 1)$$

is very close to $1/N$. The sequence 0, 1, 1, $\cdots$, 1 of length N is called an N-cell. The random blocking process $\{\tilde{R}_n\}$ will tell us when to apply the N-block coding of $\{Z_n\}$. $\{R_n\}$ can be constructed by taking an event A of very small probability, then inserting 0's every Nth place in the long blocks of 1's which occur in $\{1 - X_A(T^n x)\}$.

Step 4. For a given input sequence x, the channel $\bar{\nu}$ constructs the output process $\bar{Y}$ as follows. If $\tilde{R}_n$ is not contained in an N-cell, we put $\bar{Y}_n = b$, where b is some fixed output letter. If $\tilde{R}_n$ is the zero of an N-cell we put

$$\bar{Y}_n^{n+N-1} = \phi_{x_n^{n+N-1}}(\bar{Z}_n^{n+N-1}).$$

We are then guaranteed that if n is the first letter of an N-cell, then

$$\bar{d}_N((\nu_x)_n^{n+N-1}, \quad (\bar{\nu}_x)_n^{n+N-1}) \leq \alpha + 2\beta.$$

If α and β are small enough and N is sufficiently large the CABI property guarantees that for all m sufficiently large and all x

$$\bar{d}_m(\nu_n, \bar{\nu}_x) \leq \varepsilon$$

which proves Lemma 9.

We are now ready to complete the proof of our fundamental result, Theorem 3. Let ν be a $\bar{d}$-continuous, CABI channel and let $\{Z_n\}$ be an i.i.d. process for which $H(Z) > H(\nu)$. We can use the Friedman-Ornstein isomorphism theorem [1] to find i.i.d. processes $\{\bar{Z}_n\}$ and $\{\bar{R}\}$ such that $H(Z) > H(\nu)$ and $\{Z_n\}$ is isomorphic to the direct product $\{(Z_n, R_n)\}$. This isomorphism is implemented by an infinite sliding block code $F: Z_{-\infty}^{\infty} \to (\bar{Z}, \bar{R})$.

To construct our channel $\bar{\nu}$, we first use Lemma 9 to construct a primitive channel $\hat{\nu}$ with noise source $(\bar{Z}, \bar{R})$ such that

(i) $$\bar{d}(\nu, \hat{\nu}) < \varepsilon/2.$$

Let us suppose that the sliding block codes for $\hat{\nu}$ has width $2\omega + 1$, that is, there is a function $f: (X \times Z \times R)_{-\omega}^{\omega} \to Y$ such that for each x, the output $\hat{Y}/x$ of $\hat{\nu}$ is given by

$$(\bar{Y}/x)_0 = f(x_{-\omega}^{\omega}, \bar{Z}_{-\omega}^{\omega}, \bar{R}_{-\omega}^{\omega}).$$

As shown in [3, Theorem 3.1], we can approximate F arbitrarily well by a finite code, hence we can find an integer M and a function $\bar{F}: Z_{-M}^{M} \to (\bar{Z}, \bar{R})$ such that

(ii) $$\text{Prob}(F(\{Z_n^{(i)}\}) = \bar{F}(\{Z_n^{(i)}\}) : -\omega \leq i \leq \omega) > 1 - \varepsilon/2 \quad \text{where} \quad Z_n^{(i)} = Z_{n+i}.$$

Our primitive channel $\bar{\nu}$ uses the noise source $\{Z_n\}$ and the encoder

$$\bar{f}(x_{-M-\omega}^{M+\omega}, Z_{-M-\omega}^{M+\omega}) = f(x_{-\omega}^{\omega}, \bar{Z}_{-\omega}^{\omega}, \bar{R}_{-\omega}^{\omega})$$

where

$$(\bar{Z}_i, \bar{R}_i) = \bar{F}(\{Z_n^{(i)}\}).$$

Property (ii) guarantees that $\bar{d}(\bar{\nu}, \hat{\nu}) < \varepsilon/2$ so that property (i) gives the desired result $\bar{d}(\nu, \bar{\nu}) < \varepsilon$.

This completes our proof of Theorem 3. We note in closing that it is possible to modify the above arguments to nest the encodings so as to create an infinite sliding code from $(X \times Z)_{-\infty}^{\infty}$ to Y, which produces a channel $\bar{\nu}$ such that $\bar{d}(\nu, \bar{\nu}) = 0$.

REFERENCES

[1] Friedman, N. A. and Orstein, D. S. (1970). On isomorphism of weak Bernoulli transformations. *Advances in Math.* **4** 337–352.
[2] Gallager, R. (1968). *Information theory and reliable communication.* Wiley, New York.
[3] Gray, R. M. (1975). Sliding-block source coding. *IEEE Trans. Inform. Theory* **IT-21** 357–368.
[4] Gray, R. M. and Ornstein, D. S. (1979). Block coding for discrete stationary $\bar{d}$-continuous noisy channels. *IEEE Trans. Inform. Theory* **IT-25** 292–306.
[5] Neuhoff, D. L. and Shields, P. C. (1979). Channels with almost finite memory. *IEEE Trans. Inform. Theory.* **IT-25** 440–447.

Department of Electrical and
Computer Engineering
University of Michigan
Ann Arbor, Michigan 48109

Department of Mathematics
University of Toledo
Toledo, Ohio 43606