*The Annals of Probability*
1980, Vol. 8, No. 4, 639–674

# BLOCK SYNCHRONIZATION, SLIDING-BLOCK CODING, INVULNERABLE SOURCES AND ZERO ERROR CODES FOR DISCRETE NOISY CHANNELS[1]

BY R. M. GRAY, D. S. ORNSTEIN AND R. L. DOBRUSHIN

*Stanford University and Institute for Problems of Information Transmission, Moscow*

Results are obtained on synchronizing block codes for discrete stationary totally ergodic $\bar{d}$-continuous noisy channels (which may have infinite memory and anticipation) and used to prove sliding-block joint source and channel coding theorems. The coding theorems are used to demonstrate the existence of invulnerable sources—ergodic sources which can be input directly to the channel without encoding and decoded at the receiver with zero error—at all entropy rates below channel capacity. Combining the invulnerable source theorem with the isomorphism theorem of ergodic theory shows that, if the source is a $B$-process with entropy below capacity, then infinite length codes with zero error exist, proving that the zero-error capacity equals the usual channel capacity.

**1. Introduction.** The vast majority of block coding theorems for noisy channels assume synchronous channels, that is, channels for which the receiver knows a priori the block location and hence how to segment the received data blocks for decoding. Exceptions are the works of Nedoma (1957, 1964) Dobrushin (1967), Vajda (1965), and Ahlswede and Wolfowitz (1971) who studied the problem of synchronizing block codes for asynchronous discrete stationary channels that are memoryless or have finite input memory and are totally ergodic (block ergodic inputs yield block ergodic outputs).

An intimately related problem is the development of source and channel coding theorems for stationary sliding-block codes (time-invariant possibly nonlinear digital filters). Gray and Ornstein (1976) proved that, for memoryless discrete noisy channels, the operational channel capacity for sliding-block coding is the same as that using block codes. The proof used good block codes to construct good sliding-block codes using a synchronization (synch) sequence to locate the blocks and occasional random spacing between blocks to make the coding operation stationary—yielding a time-invariant sliding-block code. The most difficult part of the proof was the demonstration that the synch sequence could with high probability be distinguished from an overlap of itself and a code word, and the proof strongly depended on the memoryless channel assumption.

Received January 24, 1978; revised February 1979.

[1]This research was supported by NSF through Grants ENG-76-02276 and GJ 776, the USSR Academy of Sciences, and by both organizations through their support of the 1975 IEEE/USSR Workshop on Information Theory, Moscow, USSR.

*AMS* 1970 *subject classifications.* Primary 94A15; secondary 60G10, 28A65, 94A05.

*Key words and phrases.* Coding for noisy channels, synchronization, sliding-block codes, zero error codes, information theory.

Kieffer (1977) subsequently generalized the techniques and results of Gray and Ornstein (1976) to discrete channels having zero input memory and anticipation, a class of channels introduced by Nedoma (1957). Kieffer's adaptation of these techniques also depends on the absence of channel input memory and anticipation.

In this paper, the techniques of Dobrushin (1967) and Nedoma (1964) are combined and adapted to obtain block synchronization theorems and sliding-block joint source and channel coding theorems for a class of discrete stationary channels having possibly infinite input memory and anticipation—the class of stationary totally ergodic $\bar{d}$-continuous channels (Gray and Ornstein (1979)). Roughly speaking, these are channels such that, (1) if one knows the channel input for a sufficiently long time, then the output probability measure during the same time is known within a $\bar{d}$ or average Hamming sense, and (2) if an input process has ergodic $N$-tuples, then so does the output process. Condition (1) yields the most general class of stationary discrete channels possessing synchronous block coding theorems (Gray and Ornstein (1979)). Condition (2) ensures that relative frequencies of output $N$-tuples will converge to the appropriate expectation if those of the input do. Condition (2) holds, for example, if the channel is asymptotically output memoryless in the sense of Kadota and Wyner (1972) or Pfaffelhuber (1971) or output weakly mixing as in Adler (1961).

The theoretical approach adopted here resembles the ad hoc engineering approach as described, e.g., by Stiffler (1971), Chapter 14. One prefixes each code block by a synch sequence which is rarely decoded erroneously within a code block, and one then observes several successive output code blocks to resolve possible confusion of a synch sequence with an overlap of itself and a code word. As noted by Dobrushin (1967), it is this possible confusion of a synch with an overlap of itself that causes the most difficulty. The elegant techniques of synchronizing noiseless channels (e.g., see Stiffler (1971) or Scholtz (1966)) do not suffice because the noise may not be small and the channel filtering can destroy the structure of such codes.

Since sliding-block coding a stationary ergodic source yields a stationary ergodic encoded process (unlike block codes), a slight modification of the sliding-block coding theorem proves that there exist stationary ergodic sources that can be directly connected to the channel without encoding, yet can be reliably recovered to within $\varepsilon$ by decoding the channel output—an "$\varepsilon$-invulnerable source." We develop a convergent sequence of codes and processes that yields in the limit an 0-invulnerable or, simply, invulnerable source having any specified entropy rate below channel capacity. Thus, for such channels, there exist stationary ergodic sources at all entropy rates below capacity that can be communicated across the channel with zero error using a possibly infinite-length sliding-block decoder. This result is coupled with the isomorphism theorem of ergodic theory to show that, given any source that is a $B$-process (Ornstein (1973), (1974), Shields (1973)) with entropy rate below capacity, then there exist an infinite-length sliding-block encoder and a decoder yielding zero error. Shannon (1956) observed that traditional

coding theorems promised that ever longer codes could yield ever smaller error, but that this did not guarantee the existence of codes having *exactly* zero error probability. This led him to loosely define the zero-error capacity of a channel $C_0$ as the supremum of all rates at which zero error communication is possible. Shannon then developed several properties of $C_0$ under the assumption that only finite length block codes were allowed—a natural restriction on block codes as general infinite length block codes are not well defined, e.g., how are they to be used? If we consider the weaker definition that $C_0$ is the supremum of all entropy rates of sources which can be communicated across the channel with zero error (without restrictions on the coding structure and hence allowing infinite-length sliding-block codes), then the zero error result proves that $C_0 = C$; that is, there are sources with rates arbitrarily near capacity that can be communicated with zero error, and hence the channel noise can be completely defeated by infinite codes. This means that, not only can ever longer codes be made to have ever smaller error, they can be made to converge in a precise sense to a limiting infinite code with no error.

D. Blackwell (1959) also developed a sequence of codes asymptotically yielding a zero error relative frequency on a memoryless channel. His codes are not time invariant (are nonstationary) and are effectively a sequential decision scheme for guessing the $n$th input symbol after viewing about $Rn$ output symbols. His system of producing a sequence of decoders that with probability one make only a finite number of errors is analogous to the construction used in Section 6 to obtain a perfect decoder, but the codes here are stationary.

In addition to providing a new characterization of channel capacity, the zero error result also provides a new interpretation of joint source and channel coding: An invulnerable source provides a natural "language" for a noisy channel that converts it into a perfectly noiseless channel, and hence the goal of joint source and channel coding is to map the given source into this language in an invertible or noiseless manner. This goal is achievable for infinite length codes but only approximately achievable for finite codes—yielding the usual $\varepsilon$-type coding theorems. Hence, finite codes can be viewed as the best possible approximation to ideal infinite codes.

**2. Sources, channels, and codes.** Let $G$ be a nonempty finite set called the alphabet and $G^n$ the set of all $G$-valued $n$-tuples $u^n = (u_0, \cdots, u_{n-1})$, $u_i \in G$ for $i = 0, \cdots, n-1$. Let $G^\infty$ denote the space of all doubly-infinite $G$-valued sequences $u = (\cdots, u_{-1}, u_0, u_1, \cdots)$, $u_i \in G$ all $i$. Let $\mathcal{B}_G$ denote the class of all subsets of $G$. Define a thin cylinder in $G^\infty$ as any set of the form $c_m(a^n) = \{u: u_m^n := (u_m, \cdots, u_{m+n-1}) = a^n\}$ (the subscript $m$ is omitted when $m = 0$), and let the event space $\mathcal{B}_G^\infty$ be the smallest $\sigma$-field containing all the thin cylinders. Define the coordinate functions $U_n$ on $G^\infty$ by $U_n(u) = u_n$. For any $F \subset G^n$, define the cylinder $c_m(F) = \cup_{a^n \in F} c_m(a^n)$. Given a measure $\mu$ on the measurable space $(G^\infty, \mathcal{B}_G^\infty)$, the sequence of random variables $\{U_n\}_{n=-\infty}^\infty$ defined on the probability

space $(G^\infty, \mathcal{B}_G^\infty, \mu)$ is called a (discrete) random process or a source and is denoted by $[G, \mu, U]$, $\{U_n\}$, or $\mu$, as convenient.

For $u \in G^\infty$, define $U_m^n(u) = u_m^n$; the subscript $m$ is omitted if $m = 0$, and the superscript $n$ is omitted if $n = 1$. Denote by $\mu^n$ the restriction of $\mu$ to $(G^n, \mathcal{B}_G^n)$, that is, $\mu^n(a^n) = \mu(c(a^n))$.

Let $T$ denote the shift operator on $G^\infty$ defined by $U_n(Tu) = U_{n+1}(u)$. A source $[G, \mu, U]$ is $n$-stationary if $\mu(T^n F) = \mu(F)$ for all $F \in \mathcal{B}_G^\infty$. A 1-stationary source is called stationary. A source that is $n$-stationary for some $n$ is called block stationary. A source is $n$-ergodic if $T^n F = F$, $F \in \mathcal{B}_G^\infty$, implies $\mu(F) = 0$ or 1. A 1-ergodic source is termed ergodic. If a source is $n$-ergodic for all integers $n$, it is said to be totally ergodic.

Given a block stationary source $[G, \mu, U]$, the entropy rate $H(\mu)$ or $H(U)$ is defined by

$$H(\mu) = H(U) = \lim_{n\to\infty} - n^{-1} \sum_{u^n \in G^n} \mu^n(u^n) \log \mu^n(u^n).$$

The limit is well known to exist (e.g., Jacobs (1959), (1962)).

A channel $\nu$ with input alphabet $A$ and output alphabet $B$ (both finite and nonempty) is a list of probability measures $\{\nu_x, x \in A^\infty\}$ on $(B^\infty, \mathcal{B}_B^\infty)$, such that $\nu_x(F)$ is measurable for each $F \in \mathcal{B}_B^\infty$. "Connecting" an input process $[A, \tau, X]$ to a channel $\nu$ yields a joint input/output pair process $[A \times B, \tau\nu, (X, Y)]$, where $\tau\nu$ is the measure on $(A^\infty \times B^\infty, \mathcal{B}_A^\infty \times \mathcal{B}_B^\infty)$ specified by

$$\tau\nu(D \times F) = \int_F \nu_x(D)\, d\tau(x),$$

$D \in \mathcal{B}_B^\infty$, $F \in \mathcal{B}_A^\infty$, and $(X, Y)_n(x, y) := (X_n(x), Y_n(y)) = (x_n, y_n)$. The induced output process $[B, \overline{\tau\nu}, Y]$ is described by the measure $\overline{\tau\nu}(F) = \tau\nu(F \times A^\infty)$, $F \in \mathcal{B}_B^\infty$. We employ the common abuse of notation that $X_n$ is also a coordinate function on $A^\infty \times B^\infty$, i.e., $X_n\colon A^\infty \times B^\infty \to A$ is defined by $X_n(x, y) = x_n$.

A channel $\nu$ is stationary if $\nu_{Tx}(TF) = \nu_x(F)$ for all $x \in A^\infty$, $F \in \mathcal{B}_B^\infty$. A channel $\nu$ is $n$-ergodic if, for every $n$-ergodic input process $[A, \tau, X]$, the induced input/output process $[A \times B, \tau\nu, (X, Y)]$ is $n$-ergodic. A 1-ergodic channel is called ergodic. If a channel is $n$-ergodic for all $n$, it is said to be totally ergodic.

Let $\alpha^n$ and $\beta^n$ be probability measures on $(B^n, \mathcal{B}_B^n)$ and let $\mathcal{P}(\alpha^n, \beta^n)$ denote the class of all joint probability measures on $(B^n \times B^n, \mathcal{B}_{B\times B}^n)$ having $\alpha^n$ and $\beta^n$ as marginals, that is, if $p \in \mathcal{P}(\alpha^n, \beta^n)$, then $p(B^n \times F) = \beta^n(F)$ and $p(F \times B^n) = \alpha^n(F)$, all $F \in \mathcal{B}_B^n$. Define for $i = 0, \cdots, n-1$ the coordinate functions $Y_i : B^n \times B^n \to B$ and $\hat{Y}_i : B^n \times B^n \to B$ by $Y_i(y_n, \hat{y}^n) = y_i$ and $\hat{Y}_i(y^n, \hat{y}^n) = \hat{y}_i$; $y^n, \hat{y}^n \in B^n$. Let $d_n$ denote the normalized Hamming distance

$$d_n(y^n, \hat{y}^n) = n^{-1} \sum_{i=0}^{n-1} d_1(y_i, \hat{y}_i)$$

where

$$d_1(a, b) = \begin{matrix} 1 & a \neq b \\ 0 & a = b. \end{matrix}$$

The $n$th order $\bar{d}$ distance between $\alpha^n$ and $\beta^n$ is

$$\begin{aligned}\bar{d}_n(\alpha^n, \beta^n) &= \inf_{p \in \mathcal{P}(\alpha^n, \beta^n)} E_p d_n(Y^n, \hat{Y}^n) \\ &= \inf_{p \in \mathcal{P}(\alpha^n, \beta^n)} n^{-1} \sum_{i=0}^{n-1} \Pr(Y_i \neq \hat{Y}_i).\end{aligned}$$

A channel $\nu$ is $\bar{d}$-continuous (Gray and Ornstein (1979)) if given $\varepsilon > 0$ there is an integer $n_0$ such that for $n \geqslant n_0$

$$\bar{d}_n(\nu_x^n, \nu_{x'}^n) \leqslant \varepsilon$$

whenever $x_i = x_i'$, $i = 0, \cdots, n-1$. Equivalently, a channel $\nu$ is $\bar{d}$-continuous if

$$\limsup_{n \to \infty} \max_{a^n \in A^n} \sup_{x, x' \in c(a^n)} \bar{d}_n(\nu_x^n, \nu_{x'}^n) = 0. \tag{2.1}$$

Given a stationary input/output process $[A \times B, \tau\nu, (X, Y)]$, the average mutual information rate $I(X; Y)$ or $I(\tau\nu)$ between input and output is defined by $I(\tau\nu) = H(\tau) + H(\overline{\tau\nu}) - H(\tau\nu)$. The information rate or Shannon capacity of a stationary channel is defined by $C = \sup I(\tau\nu)$, where the supremum is over all stationary input processes $[A, \tau, X]$.

A sequence encoder $\bar{f}: G^\infty \to A^\infty$ is a measurable mapping from source sequence space to channel input sequence space and a sequence decoder $\bar{g}: B^\infty \to G^\infty$ is likewise a measurable mapping from the channel output sequence space into the original source sequence space. A sequence coder, say $\bar{f}$, is $N$-stationary if $\bar{f}(T^N u) = T^N \bar{f}(u)$. A 1-stationary sequence coder is said to be stationary. A source $[G, \mu, U]$, sequence encoder $\bar{f}$, channel $[A, \nu, B]$, and sequence decoder $\bar{g}$ together yield a communications system process $[G \times A \times B \times G, q, (U, X, Y, \hat{U})]$ where the measure $q$ is specified by

$$\begin{aligned}q(F_1 \times F_2 \times F_3 \times F_4) &= \int_{F_1 \cap \bar{f}^{-1}(F_2)} \nu_{\bar{f}(u)}(F_3 \cap \bar{g}^{-1}(F_4))\, d\mu(u), \\ &\qquad F_1 \in \mathcal{B}_G^\infty, F_2 \in \mathcal{B}_A^\infty, F_3 \in \mathcal{B}_B^\infty, F_4 \in \mathcal{B}_G^\infty.\end{aligned}$$

We consider two forms of coding structures yielding sequence coders with different properties: block codes and sliding-block codes. A block code of length $n$ is a pair of mappings $\gamma_n : G^n \to A^n$ (encoder) and $\psi_n : B^n \to G^n$ (decoder) which induce sequence codes $\gamma(u) = (\cdots, \gamma_n(u_{-n}, \cdots, u_{-1}), \gamma_n(u_0, \cdots, u_{n-1}), \cdots)$ and $\bar{\psi}(y) = (\cdots, \psi_n(y_{-n}, \cdots, y_{-1}), \psi_n(y_0, \cdots, y_{n-1}), \cdots)$. Define also $\gamma_{mn}(u^{mn}) = (\gamma_n(u^n), \gamma_n(u_n^n), \cdots, \gamma_n(u_{(m-1)n}^n))$. The sequence coders induced by length $n$ block codes are $n$-stationary. Given a block code $(\gamma_n, \psi_n)$, a source $\mu$, and a channel $\nu$, the block-error probability is defined by

$$P_b(\mu, \nu, \gamma_n, \psi_n) = \Pr(\psi_n(Y^n) \neq U^n) = \int \nu_{\bar{\gamma}(u)}^n(\psi_n^{-1}(u^n))\, d\mu(u).$$

A sliding-block code is a pair of mappings $f: G^\infty \to A$ (encoder) and $g: B^\infty \to G$ (decoder) which induce sequence coders $\bar{f}(u) = (\cdots, f(T^{-1}u), f(u), f(Tu), \cdots)$ and $\bar{g}(y) = (\cdots, g(T^{-1}y), g(y), g(Ty), \cdots)$. The sliding-block encoder $f$ is said to have (finite) length $n'$, memory $m'$, and delay $n' - m' - 1$ if there is a mapping $f_{n'} : G^{n'} \to A$ such that $f = f_n(U_{-m'}^n)$, that is, $f(u) = f_n(U_{-m'}^{n'}(u)) = f_{n'}(u_{-m'}, \cdots, u_0, \cdots, u_{-m'+n'-1})$. A finite length sliding block code will be

denoted by $f$ or $f_{n'}$ as convenient, that is, it can be considered either as a mapping on $G^\infty$ or on $G^{n'}$. A sliding-block decoder $g$ is likewise said to have finite length $n$, memory $m$, and delay $n - m - 1$ if there is a mapping $g_n : B^n \to G$ such that $g = g_n(Y_{-m}^n)$. Sliding-block codes induce stationary sequence coders. Given a stationary source $[G, \mu, U]$, sliding-block code $(f, g)$ (finite or infinite length), and a stationary channel $\nu$, the symbol probability of error is defined by

$$P_s(\mu, \nu, f, g) = \Pr(U_0 \neq \hat{U}_0) = \int \nu_{\bar{f}(u)}(y : g(y) \neq u_0)\, d\mu(u).$$

If $g$ has finite length $n$ and memory $m$, then

$$P_s(\mu, \nu, f, g) = \int \nu^n_{\bar{f}(u)}(y^n : g_n(y^n) \neq u_m)\, d\mu(u).$$

A particular case of interest is the identity sliding-block encoder $i : A^\infty \to A$ defined by $i(x) = x_0$. If for a source $[A, \mu, X]$ there exists a sliding block decoder $g$ such that $P_s(\mu, \nu, i, g) \leqslant \varepsilon$, the source is said to be $\varepsilon$-invulnerable.

A source $[G, \mu, U]$ is said to be block admissible (for the channel $\nu$) if given $\varepsilon > 0$ there exists for sufficiently large $n$ a block code $(\gamma_n, \psi_n)$ for which $P_b(\mu, \nu, \gamma_n, \psi_n) \leqslant \varepsilon$. A source $[G, \mu, U]$ is sliding-block admissible if given $\varepsilon > 0$ there exists a finite length sliding-block $(f, g)$ for which $P_s(\mu, \nu, f, g) \leqslant \varepsilon$.

It is known that if $\mu$ and $\nu$ are stationary, and if $H(u) > C$, then $\mu$ is not block admissible (negative or converse coding theorem) and that if $\mu$ is stationary and ergodic and $\nu$ is stationary, ergodic, and $\bar{d}$-continuous, and $H(\mu) < C$, then $\mu$ is block admissible (positive coding theorem) (Gray and Ornstein (1979)). General converse coding theorems exist for sliding-block codes, that is, if $\mu$ and $\nu$ are stationary and $H(\mu) > C$, then $\mu$ is not sliding-block admissible (Gray and Ornstein (1976)). Positive sliding-block coding theorems exist, however, only for channels without memory and anticipation (Gray and Ornstein (1976), Kieffer (1978)). We here generalize these positive theorems to stationary, totally ergodic $\bar{d}$-continuous channels.

Both block and sliding-block codes can be constructed from codebooks. A block-length $n$ codebook $\mathcal{C} = \{\mathbf{w}_i, W_i;\ i = 1, \cdots, M\}$ is a collection of $|\mathcal{C}| =$ cardinality of $\mathcal{C} = M$ codewords $\mathbf{w}_i \in A^n$ and disjoint (but not necessarily exhaustive) decoding sets $W_i \in \mathcal{B}_B^n$. The rate of the codebook is defined as $n^{-1} \ln M$. A codebook $\mathcal{C}$ is called an $(M, n, \varepsilon)$ codebook for $\nu$ if

$$\max_{1 \leqslant j \leqslant M} \sup_{x \in c(\mathbf{w}_j)} \nu_x^n(W_j^c) \leqslant \varepsilon.$$

A codebook is said to be $\delta$-robust if the expanded decoding sets $(W_i)_\delta = \{y^n : d_n(y^n, W_i) \leqslant \delta\}$ are disjoint, where $d_n(y^n, W_i) = \min_{v^n \in W_i} d_n(y^n, v^n)$.

One can treat a channel $\nu$ as if it had no memory and anticipation by "averaging out" the effect of past and future input symbols using some channel input source measure $\tau$, that is, given $\nu$ and a channel input source $\tau$, define for each $n$ and $a^n \in A^n$ for which $\mu^n(a^n) \neq 0$ the measure $\hat{\nu}^n(\cdot\,|a^n)$ on $(B^n, \mathcal{B}_B^n)$ by

$$\hat{\nu}^n(F|a^n) = \tau^n(a^n)^{-1} \int_{x \in c(a^n)} \nu_x^n(F)\, d\tau(x). \tag{2.2}$$

A codebook $\mathcal{C} = \{\mathbf{w}_i, W_i;\ i = 1, \cdots, M\}$ is called a $(\tau, M, n, \varepsilon)$ Feinstein code for $\nu$ if $\max_{1<j<M} \hat{\nu}^n(W_i^c|\mathbf{w}_i) \leqslant \varepsilon$. Good $\delta$-robust Feinstein codes can be used to construct good codebooks and good block codes for $\bar{d}$-continuous channels. The principal result in this construction is the following lemma which is quoted for reference (Gray and Ornstein (1979), Lemma 4).

LEMMA 2.1. *If*

$$\max_{a^n \in A^n} \sup_{x, x' \in c(a^n)} \bar{d}_n(\nu_x^n, \nu_{x'}^n) \leqslant \delta^2, \tag{2.3}$$

*then if* $\tau^n(x^n) > 0$ *and* $G \in \mathcal{B}_B^n$,

$$\nu_x^n((G)_\delta) \geqslant \hat{\nu}^n(G|x^n) - \delta. \tag{2.4}$$

*Thus if* $\nu$ *is* $\bar{d}$*-continuous, given* $\delta > 0$ *there is an* $n_0$ *such that* (2.3) *and hence* (2.4) *hold for* $n \geqslant n_0$.

The lemma means, for example, that if $\{\mathbf{w}_i, W_i;\ i = 1, \cdots, M\}$ is a $\delta$-robust $(\tau, M, n, \varepsilon)$ Feinstein code and $n \geqslant n_0$, then $\{\mathbf{w}_i; (W_i)_\delta;\ i = 1, \cdots, M\}$ is an $(M, n, \varepsilon + \delta)$ codebook.

We here will construct good sliding-block codes from good robust Feinstein codes.

Sliding-block codes can be constructed from block codes via the Rohlin-Kakutani (R-K) theorem of ergodic theory. A proof of the R-K theorem may be found in Shields' (1973) and a tutorial description of its use in constructing sliding-block codes in Gray and Ornstein (1976). The R-K theorem states that, given an ergodic source $[G, \mu, U]$, an $\varepsilon > 0$, and a positive integer $n$, there exists an event $F \in \mathcal{B}_G^\infty$ (called the base) such that $F, TF, \cdots, T^{n-1}F$ are disjoint,

$$\mu\left(\cup_{k=0}^{n-1} T^k F\right) \geqslant 1 - \varepsilon \tag{2.5}$$

and

$$\begin{aligned} \mu(c(u^n)|F) &= \mu(c(u^n) \cap F)/\mu(F) = \mu(c(u^n)) \\ &= \mu^n(u^n), \qquad \text{all } u^n \in G^n. \end{aligned} \tag{2.6}$$

We note also that the above implies that

$$(1 - \varepsilon)n^{-1} \leqslant \mu(F) \leqslant n^{-1}. \tag{2.7}$$

The above structure is called a $(n, \varepsilon)$-gadget. A block code $\gamma_n : G^n \to \mathcal{C}$ is imbedded in the gadget to form an infinite-length sliding-block code $f : G^\infty \to A$ by labeling the columns $\{T^i c(u^n) \cap F;\ i = 0, \cdots, n-1\}$ by the code words $\gamma_n(u^n)$, that is, for $a^*$ an arbitrary reference letter

$$f(u) = \begin{cases} a & u \in T^i(c(\alpha^n) \cap F), \quad (\gamma_n(u^n))_i = a \\ a^* & u \notin \cup_{i=0}^{n-1} T^i F. \end{cases} \tag{2.8}$$

A useful property of this imbedding procedure is given by the following lemma. Roughly, it states that the entropy rate of the sliding-block encoded process cannot be much larger than the block code rate $n^{-1} \log|\mathcal{C}|$.

LEMMA 2.2. *If $f : G^{\infty} \to A$ is an infinite-length sliding-block code constructed by imbedding a block code $\gamma_n : G^n \to \mathcal{C}$ in an $(n, \varepsilon)$-gadget for an ergodic source $[G, \mu, U]$, then*

$$H(\mu\bar{f}^{-1}) \leqslant n^{-1} \log|\mathcal{C}| + h(1/n) \tag{2.9}$$

*where $h(a) = -a \log a - (1 - a)\log(1 - a)$ is the binary entropy function and $\bar{f}$ is the induced sequence coder.*

PROOF. Define the random variable $Z_i(u) = 1_F(T^i u)$, where $1_F$ is the indicator function for $F$ ($Z_i$ "marks" the base). Note that whenever $Z_i = 1$, then $Z_{i+1} = \cdots = Z_{i+n-1} = 0$. We have for $X_n(u) = f(T^n u)$ that

$$H(\mu\bar{f}^{-1}) = H(X) \leqslant H(X, Z) = H(X|Z) + H(Z) \tag{2.10}$$

and

$$H(Z) \leqslant H_1(Z^1) = h(\mu(F)).$$

From the R-K theorem, $\mu(F) \leqslant 1/n$, and hence

$$H(Z) \leqslant h(1/n). \tag{2.11}$$

We have also

$$H(X|Z) = \lim_{m\to\infty} m^{-1} H(X^m|Z^m), \tag{2.12}$$

where

$$H(X^m|Z^m) = \Sigma_{z^m} H(X^m|z^m) \Pr(Z^m = z^m).$$

Each time $Z_i = 1$, there are $|\mathcal{C}|$ choices of the next $n$ output symbols. Define the Hamming weight of $z^m$ as $w_H(z^m) = \sum_{i=0}^{m-1} z_i$ = number of ones in $z_0, \cdots, z_{m-1}$. The $m$-tuple $x^m$ has at least $w_H(z^m) - 1$ complete $n$-blocks from $\mathcal{C}$ (the final one may be cut off) and at most $w_H(z^m)$ $n$-blocks and their location is specified by $z^m$. Prior to the appearance of the first 1 in $z^m$, there can be at most one piece of an $n$-block and the remaining symbols must be $a^*$ and this means at most $|\mathcal{C}| \cdot n$ possibilities (since the tail end of the $n$-block starting before $t = 0$ must end in one of the first $n - 1$ positions). Thus given $z^m$ there are at most $n \cdot |\mathcal{C}|^{w_H(z^m)+1}$ possible $x^m$'s and therefore

$$\begin{aligned} H(X^m|z^m) &= -\Sigma_{a^m \in A^n} \Pr(X^m = a^m|z^m) \log \Pr(X^m = a^m|z^m) \\ &\leqslant \log|\mathcal{C}|^{w_H(z^m)+1} = (w_H(z^m) + 1)\log|\mathcal{C}| \end{aligned}$$

and hence from (2.7)

$$\begin{aligned} H(X^m|Z^m) &= \Sigma_{z^m} H(X^m|z^m) \Pr(Z^m = z^m) \\ &\leqslant \left\{E\left(\Sigma_{i=0}^{m-1} Z_i\right) + 1\right\} \log|\mathcal{C}| \\ &\leqslant (m\mu(F) + 1)\log|\mathcal{C}| \\ &\leqslant (mn^{-1} + 1)\log|\mathcal{C}| \end{aligned}$$

so that

$$H(X|Z) \leqslant n^{-1} \log|\mathcal{C}|$$

which with (2.10) and (2.11) completes the proof.

**3. Statement and discussion of results.** The results developed herein are obtained using straightforward techniques from information theory and ergodic theory. Unfortunately, however, the bookkeeping details are often long and uninformative and can obscure the basic ideas. Hence in this section we state the results and describe in an intuitive manner the method of proof. The actual proofs are presented in terse form in the next section.

*Synchronization words.* The first step in constructing a good sliding-block code from a codebook is to construct a synchronization or synch word for the codebook. A synch word serves as a "punctuation mark" or prefix to locate the beginning of a codeword. It should have a length only a small fraction of the codeword length so that only a small percentage of the time spent transmitting information is devoted to punctuation, and the synch word should rarely be falsely detected within a time frame occupied by a code word, that is, the synch word decoding set should not look like a segment of any word in a codeword decoding set. As the channel is assumed to be $\bar{d}$-continuous, good robust Feinstein codes can be used to obtain good codebooks and hence we begin with such Feinstein codes. The following lemma states that given a sequence of good robust Feinstein codes, if the code length is sufficiently large we can find a synch word such that the codebook is only slightly modified, the synch word length is a specified fraction of the codeword length, and synch decoding words never appear as a segment of codeword decoding words. The technique is due to Dobrushin (1967) and is an application of the random coding technique of Shannon (1949, 1957) and Wolfowitz (1964). One selects a short good robust Feinstein code (from which the synch word will be selected) and then performs the following "thought experiment:" A word from the short code and a word from the long code are selected independently and at random and the probability that the short decoding word appears in the long decoding word is shown to be small. Since this average is small, there must be at least one short word such that the probability of its decoding word appearing in the decoding word of a randomly selected long code word is small. This in turn implies that if all long decoding words containing the short decoding word are removed from the long code decoding sets, the decoding sets of most of the original long code words will not be changed by much. In fact, one must remove a bit more from the long word decoding sets in order to ensure the desired properties are preserved when passing from a Feinstein code to a channel codebook.

LEMMA 3.1. *Assume that* $\varepsilon \leqslant 1/4$ *and* $\{\mathcal{C}_n;\ n \geqslant n_0\}$ *is a sequence of* $\varepsilon$-robust $\{\tau, M(n), n, \varepsilon/2\}$ *Feinstein codes for a* $\bar{d}$-continuous *channel* $\nu$ *having capacity* $C > 0$. *Assume also that* $h(2\varepsilon) + 2\varepsilon \log(\|B\| - 1) < C$. *For each* $n \geqslant n_0$ *let* $\{p_n(i),$

*$i = 1, \cdots, M(n)\}$ be an arbitrary probability mass function and choose $\delta \in (0, 1/4)$. Then there exists an $n_1$, such that for each $n \geq n_1$ the following statements are true:*

(A) *If $\mathcal{C}_n = \{\mathbf{v}_i, \Gamma_i; i = 1, \cdots, M(n)\}$, then there is a modified codebook $\mathcal{W}_n = \{\mathbf{w}_i; W_i; i = 1, \cdots, K(n)\}$ and a set of $K(n)$ indices $\mathcal{K}_n = \{k_1, \cdots, k_{K(n)}\} \subset \{1, \cdots, M(n)\}$ such that $\mathbf{w}_i = \mathbf{v}_{k_i}$, $W_i \subset (\Gamma_{k_i})_{\varepsilon^2}$, $i = 1, \cdots, K(n)$, and*

$$\max_{1 \leq j \leq K(n)} \sup_{x \in c(\mathbf{w}_j)} \nu_x^n(W_j^c) \leq \varepsilon. \tag{3.1}$$

(B) *There is a* synch *word $\boldsymbol{\sigma} \in A^r$, $r = r(n) = \lceil \delta n \rceil$ = smallest integer larger than $\delta n$ and a synch decoding set $S \in \mathcal{B}_B^r$ such that*

$$\sup_{x \in c(\boldsymbol{\sigma})} \nu_x^r(S^c) \leq \varepsilon \tag{3.2}$$

*and such that no r-tuple in $S$ appears in any n-tuple in any $W_i$, that is, if $G(b^r) = \{y^n : y_i^r = b^r, \text{ some } i = 0, 1, \cdots, n - r\}$ and $G(S) = \cup_{b^r \in S} G(b^r)$, then*

$$G(S) \cap W_i = \phi_1 \qquad i = 1, \cdots, K(n). \tag{3.3}$$

(C) *We have that*

$$\Sigma_{k \notin K_n} p_n(k) \leq \varepsilon\delta. \tag{3.4}$$

The modified code $\mathcal{W}_n$ has fewer words than the original code $\mathcal{C}_n$, but (3.4) ensures that $\mathcal{W}_n$ cannot be much smaller since, for example, if $p_n(i) = 1/M(n)$, $i = 1, \cdots, M(n)$, then (3.4) becomes

$$K(n) \geq (1 - \varepsilon\delta)M(n). \tag{3.5}$$

Given a codebook $\mathcal{W}_n = \{\mathbf{w}_i, W_i; i = 1, \cdots, K(n)\}$, a synch word $\boldsymbol{\sigma} \in A^r$, and a synch decoding set $S$, we shall call the length $n + r$ codebook, $\{\boldsymbol{\sigma} \times \mathbf{w}_i, S \times W_i; i = 1, \cdots, K(n)\}$ a prefixed (or punctuated) codebook.

*Comment.* To prove the basic coding theorems we need only consider the case $p_n(i) = 1/M(n)$, $i = 1, \cdots, M(n)$. The more general result is required, however to prove the invulnerable source theorem.

By combining the preceeding lemma with the existence of robust Feinstein codes at rates less than capacity (Gray and Ornstein (1979)) we have the following.

COROLLARY 3.1. *Let $\nu$ be a stationary ergodic $\bar{d}$-continuous channel and fix $\varepsilon > 0$ and $R \in (0, C)$. Then there exists for sufficiently large blocklength $N$ a length $N$ codebook $\{\boldsymbol{\sigma} \times w_i, S \times W_i; i = 1, \cdots, M\}$, $M \geq 2^{NR}$ $\boldsymbol{\sigma} \in A^r$, $\mathbf{w}_i \in A^n$, $r + n = N$, such that*

$$\begin{aligned} \sup_{x \in c(\boldsymbol{\sigma})} \nu_x^r(S^c) &\leq \varepsilon \\ \max_{1 \leq j \leq M} \nu_x^n(W_j^c) &\leq \varepsilon \\ W_j \cap G(S) &= \phi, \qquad j = 1, \cdots, M. \end{aligned}$$

SLIDING-BLOCK CODING: TOTALLY ERGODIC SOURCES. The synch word can be used to mark the beginning of a codeword and it will rarely be falsely detected during a codeword. Unfortunately, however, an $r$-tuple consisting of a segment of a

synch and a segment of a codeword may be falsely detected as a synch with nonnegligible probability. To resolve this confusion we look at the relative frequency of synch-detects over a sequence of blocks instead of simply trying to find a single synch. The idea is that if we look at enough blocks, the relative frequency of the synch-detects in each position should be nearly the probability of occurrence in that position and these quantities taken together give a pattern that can be used to determine the true synch location. For the ergodic theorem to apply, however, we require that blocks be ergodic and hence we first consider totally ergodic sources and channels.

LEMMA 3.2. *Let $\nu$ be a totally ergodic stationary $\bar{d}$-continuous channel, fix $\varepsilon$, $\delta > 0$, and assume that $\mathcal{C}_N = \{\sigma \times \mathbf{w}_i;\ S \times W_i;\ i = 1, \cdots, K\}$ is a prefixed codebook satisfying* (3.1)–(3.3). *Let $\gamma_N : G^N \to \mathcal{C}_N$ assign an $N$-tuple in the prefixed codebook to each $N$-tuple in $G^N$ and let $[G, \mu, U]$ be an $N$-stationary, $N$-ergodic source. There exists for sufficiently large $L$ (which depends on the source) a synch locating function $\sigma : B^{LN} \to \{0, 1, \cdots, N-1\}$ and a set $\Phi \in \mathcal{B}_G^m$, $m = (L+1)N$, such that, if $u^m \in \Phi$ and $\gamma_N(u_{LN}^N) = \sigma \times \mathbf{w}_i$, then*

(3.6) $$\inf_{x \in c(\gamma_m(u^m))} \nu_x\big(y : \sigma(y^{LN}) = \theta,\ \theta = 0, \cdots, N-1; y_{LN}^N \in S \times W_i\big) \geqslant 1 - 3\varepsilon$$

*and*

(3.7) $$\mu^m(\Phi) \geqslant 1 - \varepsilon.$$

The lemma can be interpreted as follows. The source is block encoded using $\gamma_N$. The decoder observes a possible synch word and then looks "back" in time at previous channel outputs and calculates $\sigma(y^{LN})$ to obtain the exact synch location which is correct with high probability.

The synch locator function is constructed roughly as follows: since $\mu$ and $\nu$ are $N$-stationary and $N$-ergodic, if $\bar{\gamma} : A^\infty \to B^\infty$ is the sequence encoder induced by the length $N$-block code $\gamma_N$, then the encoded source $\mu\bar{\gamma}^{-1}$ and the induced channel output process $\eta$ are all $N$-stationary and $N$-ergodic. The sequence $s_j = \eta(T^j c(S))$, $j = \cdots, -1, 0, 1, \cdots$ is therefore periodic with period $N$. Furthermore, $s_j$ can have no smaller period than $N$ since from (3.1)–(3.3) $\eta(T^j c(S)) \leqslant \varepsilon$, $j = r+1, \cdots, N-r$, and $\eta(c(S)) \geqslant 1 - \varepsilon$. Thus if we define the synch pattern $\{s_j; j = 0, \cdots, N-1\}$, the synch pattern is distinct from any cyclic shift of itself of the form $\{s_k, \cdots, s_{N-1}, s_0, \cdots, s_{k-1}\}$, where $1 \leqslant k \leqslant N-1$. The synch locator computes the relative frequencies of the occurrence of $S$ at intervals of length $N$ for each of $N$ possible starting points to obtain, say, a vector $\hat{\mathbf{s}} = (\hat{s}_0, \cdots, \hat{s}_{N-1})$. The ergodic theorem implies the $\hat{s}_i$ will be near their expectation and hence with high probability $(\hat{s}_0, \cdots, \hat{s}_{N-1}) = (s_\theta, \cdots, s_{N-1}, s_0, \cdots, s_{\theta-1})$, determining $\theta$. Another way of looking at the result is to observe that the sources $\eta T^j$, $j = 0, 1, \cdots, N-1$ are each $N$-ergodic and hence any two are either identical or orthogonal in the sense that they place all of these measures on disjoint $N$-invariant sets. No two can be identical, however, since if $\eta T^i = \eta T^j$

for $i \neq j$, $0 \leqslant i, j \leqslant N-1$, then $\eta$ would be periodic with period $0 < |i-j| < N$, yielding a contradiction. Since membership in any set can be determined with high probability by observing the sequence for a long enough time, the synch locator attempts to determine which of the $N$ distinct sources $\eta T^j$ is being observed. In fact, synchronizing the output is exactly equivalent to forcing the $N$-sources $\eta T^j$, $j = 0, \cdots, N-1$, to be distinct $N$-ergodic sources. After this is accomplished, the remainder of the proof is devoted to using the properties of $\bar{d}$-continuous channels to show that synchronization of the output source when driven by $\mu$ implies that with high probability the channel output can be synchronized for all fixed input sequences in a set of high $\mu$ probability.

Lemma 3.2 is stronger (and more general) than the similar results of Nedoma (1964) and Vajda (1965), but the extra structure is required for application to sliding-block decoding.

The next lemma uses the prefixed block code and the synch locator function combined with the R-K theorem to construct a good sliding-block code for a totally ergodic source with entropy less than capacity. The encoder has infinite length and the decoder finite length. The subsequent corollary removes the requirement of infinite encoder length and thereby proves that a stationary totally ergodic source $\mu$ is sliding-block admissible for a stationary totally ergodic $\bar{d}$-continuous channel if $H(\mu) < C$. The lemma is proved by assigning prefixed code words to the set of roughly $2^{NH(\mu)}$ "typical" source sequences (from the Shannon-McMillan theorem) and then "stationarizing" the block code by imbedding it in a gadget. The gadget height is chosen large enough to ensure that the synch locator function will perform correctly most of the time.

LEMMA 3.3. *Given a $\bar{d}$-continuous totally ergodic stationary channel $\nu$ with Shannon capacity $C$, a stationary totally ergodic source $[G, \mu, U]$ with entropy rate $H(\mu) < C$, and $\delta > 0$, there exists for sufficiently large $m$ a sliding-block decoder $g_m : B^m \to G$ and an infinite length sliding-block encoder $f : G^\infty \to A$ such that $P_e(\mu, \nu, f, g_m) \leqslant \delta$.*

COROLLARY 3.2. *If $\nu$ is a stationary $\bar{d}$-continuous totally ergodic channel with Shannon capacity $C$, then any totally ergodic source $[G, \mu, U]$ with $H(\mu) < C$ is admissible.*

*Ergodic sources.* If a prefixed blocklength $N$ block code of Corollary 3.1 is used to block encode a general ergodic source $[G, \mu, U]$, then successive $N$-tuples from $\mu$ may not be ergodic, and hence the previous analysis does not apply. From the Nedoma decomposition (Nedoma (1963)), however, any ergodic source $\mu$ can be represented as a mixture of $N$-ergodic sources, all of which are simply shifted versions of each other. Given an ergodic measure $\mu$ and an integer $N$, then there exists a decomposition of $\mu$ into $M$ $N$-ergodic, $N$-stationary components where $M$

divides $N$, that is, there is a set $\Pi \in \mathcal{B}_G^\infty$ such that

$$T^M\Pi = \Pi \tag{3.8}$$

$$\mu(T^i\Pi \cap T^j\Pi) = 0, \quad 0 \leqslant i,j \leqslant M, i \neq j \tag{3.9}$$

$$\mu\Big(\bigcup_{i=0}^{M-1} T^i\Pi\Big) = 1 \tag{3.10}$$

$$\mu(\Pi) = 1/M, \tag{3.11}$$

the sources $(G, \pi_i, U)$, where $\pi_i(F) := \mu(F|T^i\Pi) = \mu(F \cap T^i\Pi)/\mu(\Pi) = M\mu(F \cap T^i\Pi)$ are $N$-ergodic and $N$-stationary and

$$\begin{aligned}\mu(F) &= M^{-1}\textstyle\sum_{i=0}^{M-1}\pi_i(F) \\ &= M^{-1}\textstyle\sum_{i=0}^{M-1}\mu(F|T^i\Pi) = \sum_{i=0}^{M-1}\mu(F \cap T^i\Pi).\end{aligned} \tag{3.12}$$

This decomposition provides a method of generalizing the results for totally ergodic sources to ergodic sources: since $\mu(\cdot|\Pi)$ is $N$-ergodic, Lemma 3.2 is valid if $\mu$ is replaced by $\mu(\cdot|\Pi)$. The infinite sliding-block code $f$ can ensure that the appropriate mode occurs at the output by testing for $T^{-i}\Pi$ at the gadget base and, if the base is in $T^{-i}\Pi$, insert $i$ dummy symbols, and then encode using the length $N$ prefixed block code. This means that the code is "lined up" with the $N$-ergodic design mode, say $\Pi$, and the relative frequencies converge to an appropriate expectation, yielding the desired code. A finite length encoder is then obtained as previously.

THEOREM 3.1. *If $\nu$ is stationary $\bar{d}$-continuous totally ergodic channel with Shannon capacity $C$, then any ergodic source $[G, \mu, U]$ with $H(\mu) < C$ is admissible.*

*Invulnerable sources.* The following lemma is a slight variation of Lemma 3.3. Roughly speaking, it is an observation that, given the assumptions and properties of Lemma 3.2 and the sliding-block encoder $f$ constructed from a block encoder $\gamma_N$, as in Lemma 3.3, then the receiver can reliably construct the $N$-ergodic channel input process instead of the original source. This means that there exist ergodic sources that can be connected directly to the noisy channel and recovered to within $\varepsilon$ by the decoder—an $\varepsilon$-invulnerable source. This provides a new characterization of channel capacity as given by the corollary.

LEMMA 3.4. *Let $\delta$, $\varepsilon$, $\nu$, $\mathcal{C}_N$, $N \geqslant 3$, $r$, $\gamma_N$, $L$, $m = (L+1)N$, $\Phi$, and $[G, \mu, U]$ be as in Lemma 3.2. Choose $K$ so large that $m \leqslant \varepsilon KN$, and let $f$ be the infinite-length sliding-block code obtained by imbedding $\gamma_N$ in a $(KN, \varepsilon)$-gadget with base $F$ as in (2.5)–(2.8). There is a length $m$ decoder $h_m$ such that, with $i(x) = x_0$,*

$$P_e(\mu\bar{f}^{-1}, \nu, i, h_m) \leqslant 3\varepsilon,$$

*that is, $(A, \mu\bar{f}^{-1}, X)$ is $3\varepsilon$-invulnerable.*

COROLLARY 3.3. *Let $\nu$ be a stationary totally ergodic $\bar{d}$-continuous channel. Given $\Delta, \delta > 0$ and $H < C$, there exists a $\delta$-invulnerable source $[A, \tau, X]$ such that*

$H(\tau) \geqslant H - \Delta$, *and*

$$\inf_{\delta} \sup_{\delta\text{-invulnerable } [A, \tau, X]} H(\tau) = C.$$

Corollary 3.3 raises a further question: do there exist 0-invulnerable or, simply, invulnerable sources with entropy rates near capacity. An affirmative answer to this question is provided by combining the preceeding result with an iteration that allows us to take an $\varepsilon$-invulnerable source of a given entropy rate and construct an $\varepsilon'$-invulnerable source with $\varepsilon' \ll \varepsilon$ such that the entropy rate of the $\varepsilon'$-invulnerable source is only slightly less than that of the $\varepsilon$-invulnerable source and such that the $\varepsilon'$-invulnerable source is close (in a $\bar{d}$-sense) to the $\varepsilon$-invulnerable source. This closeness is necessary in order to get a converging sequence of codes which in the limit give the desired invulnerable source. The basic technique used here can be described as follows: Given the $\varepsilon$-invulnerable source, say $\{X_k\}$, and the decoder, say $g$, if we let the source run for a long enough time $n$, then the ergodic theorem says that the probability that more than $2\varepsilon n$ errors occur between $\{X_k\}$ and the decoded process, say $\{\hat{X}_k\}$, will be very small, say less than $\delta^2$. This in turn implies that with probability at least $1 - \delta$ the channel output $y^n$ will yield a reproduction $\hat{x}^n$ differing from $x^n$ in fewer than $2\varepsilon n$ places (neglecting the "edge effects" due to the finite decoder length which are negligible for large enough $n$). Using $\bar{d}$-continuity, these good channel output $n$-blocks can be decoded using the old decoder to within $2\varepsilon n$ errors with high probability regardless of past or future channel input symbols. Of this collection of good decoder output $n$-blocks, we then form a reduced set of $n$-blocks by going through the list and removing all $n$-blocks having Hamming distance less than $6\varepsilon$ from a previous member of the list. This provides a collection of codewords $\mathcal{C}$ and we form a codebook by taking as a channel output decoding set all $y^n$ that decode (using $g$) into a reproduction $\hat{x}^n$ within $2\varepsilon n$ of the codeword. Note that by construction this codebook is $\varepsilon$-robust. We next synchronize the codebook with a synch of length $r$ and use it to form an infinite length sliding-block encoding $f'$ of the $\varepsilon$-invulnerable source by building a long gadget and then encoding each nonoverlapping $(n + r)K$-tuple in each gadget column into the closest (in the Hamming sense) prefixed codeword. This encoded $\varepsilon$-invulnerable source, say $\{X'_k\}$ is now an $\varepsilon'$-invulnerable source with the following decoder: As most of the new source consists of long sequences from the $\varepsilon$-invulnerable source (the remainder being a small amount of synchronization and the gadget garbage), the receiver first decodes using the old decoder $g$ and also finds the new synch word. Following a synch word $g$ will produce a tentative reproduction $\tilde{x}^n$. If $\tilde{x}^n \in \mathcal{C}$, then the new decoder prints out $\tilde{x}^n$ as the next $n$ symbols. If $\tilde{x}^n \notin \mathcal{C}$, however, the decoder knows that $g$ has produced a word not in the codebook and hence finds the closest word in $\mathcal{C}$ to $\tilde{x}^n$, say $\tilde{x}^n$, and then prints $\hat{x}^n$. The probability that $\tilde{x}^n$ will be within $2\varepsilon n$ of $x^n$ is at least $1 - \delta$ for $x^n \in \mathcal{C}$, but this means that $y^n$ is within the decoding set for $x^n$. Thus the blocks will be decoded correctly with probability roughly $1 - \delta$ which in turn implies that if $\delta$ is small enough, the new source is decoded within $\varepsilon'$. This can be depicted as below with [ ] denoting the $(n + r)K$-block synch and ( ) denoting the $n$-block synch.

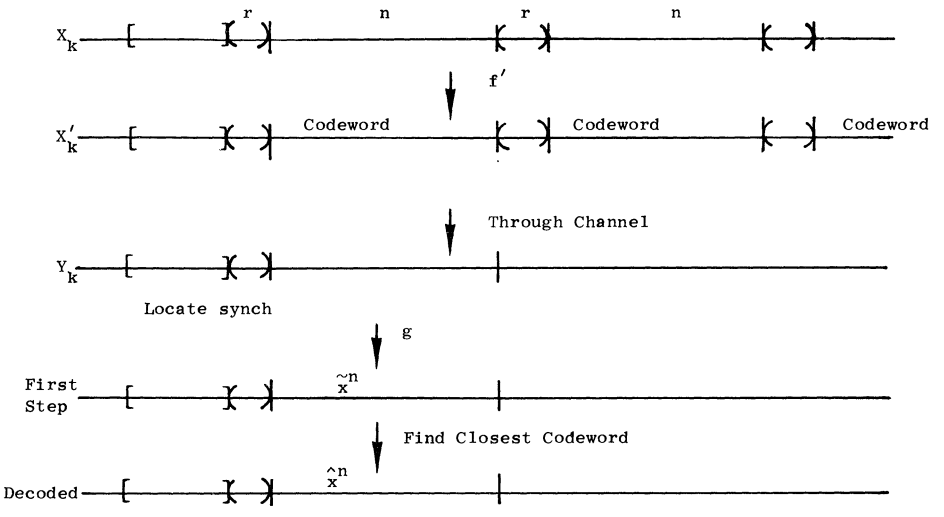


FIG. 1

Lastly, since $\{X'_K\}$ is close to $\{X_k\}$, the entropy rates are nearly equal.

To construct an invulnerable source one chooses a sequence $\varepsilon_n \to 0$ and then constructs a sequence of $\varepsilon_k$-invulnerable sources by applying the above iteration, beginning with the $\varepsilon_1$-invulnerable source of Corollary 3.3. These can all be considered as codings $f^{(n)}$ of the original $\varepsilon_1$-invulnerable source and these codes can be shown to converge to a limit code $f$ using a result of Shields (1973) and by appropriate choice of the $\varepsilon_n$ the resulting source will have the desired entropy. The source is shown to be invulnerable by using an infinite length decoder $g$ which decodes the received sequence $y$ using all of the decoders $g_k$ for all the $\varepsilon_k$-invulnerable sources and then sets $g(y) = a$ if all but a finite number of the $g_k$ decode $y$ as $a$.

Making these arguments precise yields the following results. The unfortunately tedious proofs are presented in the next section.

THEOREM 3.2. *Let $\nu$ be a stationary totally ergodic channel with Shannon capacity $C$, and let $H^* \in (0, C)$. There exists a totally ergodic invulnerable source $[A, \tau^*, X^*]$ with entropy rate $H(\tau^*) = H^*$, and hence*

$$\sup_{\text{invulnerable } [A, \tau^*, X^*]} H(\tau^*) = C.$$

Theorem 3.2 has an immediate corollary in terms of $B$-processes. A $B$-process is any process obtainable by finite or infinite length sliding-block coding an i.i.d. process (Ornstein (1973)). An alternate characterization is that $B$-processes are those processes which can be approximated arbitrarily closely in the $\bar{d}$-distance by a mixing multi-step Markov process. Since $\tau^*$ was constructed by sliding-block coding an i.i.d. source, we immediately have the following.

COROLLARY 3.4. *Given $\nu$, $C$, and $H^*$ as in Theorem 3.2 there exists an invulnerable B-process $[A, \tau^*, X^*]$ with $H(\tau^*) = H^*$.*

*Zero-error codes and zero-error capacity.* Combining Corollary 3.4 with the isomorphism theorem of ergodic theory (Ornstein (1974)) as stated in terms of sliding block codes (Gray (1975)) yields the following.

THEOREM 3.3. *If $(G, \mu, U)$ is a B-process and $\nu$ is a stationary totally ergodic $\bar{d}$-continuous channel with capacity $C > H(\mu)$, then there is an infinite-length sliding-block encoder $f: G^\infty \to A$ and an infinite-length sliding-block decoder $g: B^\infty \to A$ such that*

$$P_e(\mu, \nu, f, g) = 0,$$

*that is, the source can be communicated with zero error across the noisy channel.*

Define the weak zero-error capacity $C_0$ of a channel as the supremum of the entropy rates of all stationary processes that can be communicated across the channel with zero error using any block stationary coding (such as block codes or sliding-block codes). This is in contrast to the "strong" zero error capacity introduced by Shannon (1956) which refers to zero-error capacity in a difficult combinatorial problem and it can differ from the usual capacity. The following corollary shows, however, that under quite general conditions the weak zero error capacity is simply $C$.

COROLLARY 3.5. *Given a stationary totally ergodic $\bar{d}$-continuous channel with Shannon capacity $C$ and weak zero-error capacity $C_0$, then*

$$C = C_0.$$

DISCUSSION. The Shannon capacity defined by $C = \sup I(\tau\nu)$, where the supremum is over all stationary (or ergodic or block-stationary) input processes is often achievable, that is, the supremum is actually a maximum. For example, let $A = B = \{0, 1\}$ and let $\nu$ be a binary symmetric channel (BSC) with parameter $p < \frac{1}{2}$, that is,

$$\nu_x^n(y^n) = \Pi_{i=1}^n p^{x_i \oplus y_i}(1 - p)^{1 - x_i \oplus y_i},$$

where $\oplus$ denotes modulo two addition. For this channel, an i.i.d. equiprobable source $[A, \tau, X]$ yields $I(\tau\nu) = C = 1 - h(p)$. A natural question is whether the supremum of Theorem 3.2 or that defining $C_0$ is also a maximum, that is, does there exist an invulnerable process $[A, \tau, X]$ with entropy rate $H(\tau) = C$? We show that in general this cannot be true by showing that no invulnerable source with $H(\tau) = C$ exists for the BSC.

Assume that $[A, \tau, X]$ is invulnerable and has $H(X) = C$. The BSC can be represented by

$$Y_k = N_k \oplus X_k$$

where $\{N_k\}$ are i.i.d. with $\Pr(N_k = 1) = p$. We assumed that

$$H(X) = C = 1 - h(p) = 1 - H_1(N^1) = 1 - H(N).$$

Since $[A, \tau, X]$ is invulnerable, we have as previously that $H(X|Y) = 0$, and hence, since $H(Y|X) = H(N)$,

$$I(X; Y) = H(X) - H(X|Y) = C = H(Y) - H(Y|X) = H(Y) - H(N),$$

and hence

$$H(Y) = 1.$$

From Shields (1973), pages 51–52, however, if $H(Y) = 1$, then $\{Y_k\}$ must be i.i.d. with $\Pr(Y_i = 1) = \frac{1}{2}$. Thus, for all $n$, the probability mass function for $X^n$ must satisfy

$$\begin{aligned} p_{Y^n}(y^n) &= \left(\tfrac{1}{2}\right)^n = \Sigma_{x^n} p_{Y^n|X^n}(y^n|x^n) p_{X^n}(x^n) \\ &= \Sigma_{x^n} p_{X^n}(x^n) \Pi_{i=0}^{n-1} p^{y_i \oplus x_i} (1-p)^{1-y_i \oplus x_i}. \end{aligned} \tag{3.13}$$

This can be expressed in vector form by defining the column vector $\mathbf{p}_X^{(n)} = \{p_{X^n}(x^n),\ x^n \in \{0, 1\}^n\}$, $\mathbf{p}_Y^{(n)} = \{p_{Y^n}(y^n);\ y^n \in \{0, 1\}^n\}$, and the matrix

$$\mathbf{P}_n = \left\{p_{Y^N|X^N}(y^n|x^n);\ x^n, y^n \in \{0, 1\}^n\right\} = \mathbf{P}_1^{[n]},$$

where $\mathbf{P}_1^{[n]}$ is the $n$th Kronekker product (Bellman (1960), pages 227–229) of the matrix

$$\mathbf{P}_1 = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix},$$

so that (3.13) becomes

$$\mathbf{p}_Y^{(n)} = \mathbf{P}_n \mathbf{p}_X^{(n)},$$

and hence, if $\mathbf{P}_n$ has an inverse for all $n$,

$$\mathbf{p}_X^{(n)} = \mathbf{p}_n^{-1} \mathbf{p}_Y^{(n)}$$

uniquely defines the $p_{X^n}^n$ and hence $[A, \tau, X]$. For $p \in (0, 1)$, $\mathbf{P}_1$ has an inverse

$$\mathbf{P}_1^{-1} = \frac{1}{1-2p} \begin{pmatrix} 1-p & -p \\ -p & 1-p \end{pmatrix},$$

and hence (Bellman (1960)) $\mathbf{P}_n^{-1} = (\mathbf{P}_1^{-1})^{[n]}$, and

$$\mathbf{p}_X^{(n)} = \left(\mathbf{P}_1^{-1}\right)^{[n]} \mathbf{p}_Y^{(n)},$$

in particular, since the inverse exists, $\mathbf{p}_X^{(n)}$ has a unique solution. It is easily shown that $p_{X^n}(x^n) = \left(\frac{1}{2}\right)^n$ is a solution to (3.13), and therefore $\{X_i\}$ is an i.i.d. equiprobable source with $\Pr(X_i = 1) = \frac{1}{2}$. This implies, however, that $H(X) = 1 > C$, a contradiction.

Intuitively, an invulnerable source (or any good code) for this channel prints long sequences of symbols that look i.i.d. for a long time, but eventually the memory and dependency show up and the entropy drops, allowing the receiver to discern different input sequences by viewing the channel output. If one signals *at* capacity, however, the source must actually be i.i.d., the entropy can therefore never drop, and hence the redundancy required to distinguish sequences at the

output can never be inserted, preventing a good code. This provides a channel version of Berger and Lau's (1977) source coding result that Shannon's rate-distortion function cannot be achieved with equality using sliding-block codes.

Some open problems are (1) generalizing the output memory assumptions by removing the totally ergodic requirement; (2) generalizing the results to other distance measures, and other notions of continuity hopefully allowing results for input-constrained continuous alphabet (and continuous time?) channels, and (3) the development of a structural theory for the infinite codes yielding zero error. A simple construction for nonstationary sources yielding zero error in the limit is implicit in Cover, McEliece, and Posner (1979).

As a final comment, the results here immediately generalize the sliding-block code information transmission theorem of Gray and Ornstein (1976) from memoryless channels to those considered here by coupling sliding-block source coding with a fidelity criterion with Theorem 3.1.

## 4. Proofs.

PROOF OF LEMMA 3.1. Since $\nu$ is $\bar{d}$-continuous, we can choose $n_2$ so large that for $n \geq n_2$ we have that for

$$\max_{a^n \in A^n} \sup_{x, x' \in c(a^n)} \bar{d}_n(\nu_x^n, \nu_{x'}^n) \leq (\delta\varepsilon/2)^2. \tag{4.1}$$

From Corollaries 2–3 of Gray and Ornstein (1979) there is an $n_3$ such that for each $r \geq n_3$ there exists an $\varepsilon/2$-robust $(\tau, J, r, \varepsilon/2)$-Feinstein code $\mathcal{C}_s = \{\mathbf{s}_j, S_j; j = 1, \cdots, J\}$, $J \geq 2^{rR_s}$ where $R_s \in (0, C - h(2\varepsilon) - 2\varepsilon \log(\|B\| - 1))$. Assume that $n_1$ is large enough to ensure that $\delta n_1 \geq n_2$, $\delta n_1 \geq n_3$ and $n_1 \geq n_0$. Let $1_F$ denote the indicator function of the set $F$ and define $\lambda_n$ as

$$\begin{aligned} \lambda_n &= J^{-1}\sum_{j=1}^{J}\sum_{i=1}^{M(n)} p_n(i)\hat{\nu}^n\big(G\big((S_j)_\varepsilon\big) \cap \Gamma_i|\mathbf{v}_i\big) \\ &= J^{-1}\sum_{j=1}^{J}\sum_{i=1}^{M(n)} p_n(i)\sum_{b^r \in (S_j)_\varepsilon}\sum_{y^n \in \Gamma_i} \hat{\nu}^n(y^n|\mathbf{v}_i) 1_{G(b^r)}(y^n) \\ &= J^{-1}\sum_{j=1}^{J}\sum_{i=1}^{M(n)} p_n(i)\sum_{y^n \in \Gamma_i} \hat{\nu}^n(y^n|\mathbf{v}_i)\Big\{\sum_{j=1}^{J}\sum_{b^r \in (S_j)_\varepsilon} 1_{G(b^r)}(y^n)\Big\}. \end{aligned} \tag{4.2}$$

Since the $(S_j)_\varepsilon$ are disjoint and a fixed $y^n$ can belong to at most $n - r \leq n$ sets $G(b^r)$, the bracketed term above is bound above by $n$, whence

$$\lambda_n \leq nJ^{-1}\sum_{i=1}^{M(n)} p_n(i)\hat{\nu}^n\Gamma_i|\mathbf{v}_i) \leq nJ^{-1} \leq n2^{-rR_s} \leq n2^{-\delta nR_s} \to_{n\to\infty} 0$$

so that choosing $n_1$ also so that $n_1 2^{-\delta nR_s} \leq (\delta\varepsilon)^2$ we have that $\lambda_n \leq (\delta\varepsilon)^2$ if $n \geq n_1$. From (4.2) this implies that for $n \geq n_1$ there must exist at least one $j$ such that

$$\sum_{i=1}^{M(n)} p_n(i)\hat{\nu}^n\big(G\big((S_j)_\varepsilon\big) \cap \Gamma_i|\mathbf{v}_i\big) \leq (\delta\varepsilon)^2$$

which in turn implies there must exist a set of indices $\mathcal{K}_n \subset \{1, \cdots, M(n)\}$ such that

$$\hat{\nu}^n\big(G\big((S_j)_\varepsilon\big) \cap \Gamma_i|\mathbf{v}_i\big) \leq \delta\varepsilon, i \in \mathcal{K}_n \tag{4.3}$$

$$\sum_{i \notin \mathcal{K}_n} p_n(i) \leq \delta\varepsilon. \tag{4.4}$$

Define $\boldsymbol{\sigma} = \mathbf{s}_j$, $S = (S_j)_{\varepsilon/2}$, $\mathbf{w}_i = \mathbf{v}_{k_i}$, and $W_i = (\Gamma_{k_i} \cap G((S_j)_\varepsilon)^c)_{\varepsilon\delta}$, $i = 1, \cdots, K(n)$. We have from Lemma 2.1 and (4.1) that if $x \in c(\boldsymbol{\sigma})$, then since $\varepsilon\delta \leqslant \varepsilon/2$,

$$\nu_x^r(S) = \nu_x^r\big((S_j)_{\varepsilon/2}\big) \geqslant \hat{\nu}^r(S_j|\boldsymbol{\sigma}) - \varepsilon/2 \geqslant 1 - \varepsilon,$$

proving (3.2). If $x \subset c(\mathbf{w}_i)$, then using (4.3)

$$\nu_x^n(w_i) = \nu_x^n\big(\big(\Gamma_{k_i} \cap G\big((S_j)_\varepsilon\big)^c\big)_{\varepsilon\delta}\big) \geqslant \hat{\nu}^n\big(\Gamma_{k_i} \cap G\big((S_j)_\varepsilon\big)^c|\mathbf{v}_{k_i}\big) - \varepsilon\delta$$
$$= \hat{\nu}^n\big(\Gamma_{k_i}|\mathbf{v}_{k_i}\big) - \nu^n\big(\Gamma_{k_i} \cap G\big((S_j)_\varepsilon\big)|\mathbf{v}_{k_i}\big) - \varepsilon\delta \geqslant 1 - \varepsilon/2 - 2\varepsilon\delta \geqslant 1 - \varepsilon,$$

proving (3.1). Next note that if $y^n \in (G((S_j)_\varepsilon)^c)\varepsilon\delta$, then there is a $b^n \in G((S_j)_\varepsilon)^c$ such that $d_n(y^n, b^n) \leqslant \varepsilon\delta$ and thus for $i = 0, 1, \cdots, n - r$ we have that $d_r(y_i^r, b_i^r) \leqslant (n/r)(\varepsilon\delta/2) \leqslant \varepsilon/2$. Since $b^n \in G((S_j)_\varepsilon)^c$, it has no $r$-tuple within $\varepsilon$ of an $r$-tuple in $S_j$ and hence the $r$-tuples $y_i^r$ are at least $\varepsilon/2$ distant from $S_j$ and hence $y^n \in G((S)_{\varepsilon/2})^c$. We have therefore that $(G((S_j)_\varepsilon)^c)_{\varepsilon\delta} \subset G((S_j)_\varepsilon)^c$ and hence

$$G(S) \cap W_i = G\big((S_j)_\varepsilon\big) \cap \big(\Gamma_{k_i} \cap G\big((S_j)_\varepsilon\big)^c\big)_{\delta\varepsilon}$$
$$\subset G\big((S_j)_{\varepsilon/2}\big) \cap \big(G\big((S_j)_\varepsilon\big)^c\big)_{\delta\varepsilon} = \phi,$$

completing the proof.

PROOF OF COROLLARY 3.1. Choose $\delta \in (0, \varepsilon/2)$ so small that $C - h(2\delta) - 2\delta\log(\|B\| - 1) > (1 + \delta)R(1 - \log(1 - \delta^2))$ and chose $R' \in ((1 + \delta)R(1 - \log(1 - \delta^2)), C - h(2\delta) - 2\delta\log(\|B\| - 1)$. From Corollaries 3.1–3.2 of Gray and Ornstein (1979) there exists an $n_0$ such that for $n \geqslant n_0$ there exist $\delta$-robust $(\tau, \mu', n, \delta)$ Feinstein codes with $M' \geqslant 2^{nR'}$. From Lemma 3.1 there exists a codebook $\{\mathbf{w}_i, W_i;\ i = 1, \cdots, I(n)\}$, a synch word $\boldsymbol{\sigma} \in A^r$, and a synch decoding $S \in \mathcal{B}_B^r$, $r = \lceil \delta n \rceil$, such that

$$\max_j \sup_{x \in c(\mathbf{w}_j)} \nu_x^n(W_j^c) \leqslant 2\delta \leqslant \varepsilon,$$

$$\sup_{x \in c(\boldsymbol{\sigma})} \nu_x^r(S) \leqslant 2\delta \leqslant \varepsilon,$$

$G(S) \cap W_j = \varphi, j = 1, \cdots, K(n)$, and from (3.5),

$$M = K(n) \geqslant (1 - \delta^2)M(n).$$

We therefore have that with $N = n + r$

$$N^{-1}\log M \geqslant \big(n\lceil n\delta\rceil\big)^{-1}\log((1 - \delta^2)2^{nR'}) = (nR' + \log(1 - \delta^2))/(n + n\delta)$$
$$= (R' + n^{-1}\log(1 - \delta^2))/(1 + \delta)$$
$$\geqslant (R' + \log(1 - \delta^2))/(1 + \delta) \geqslant R,$$

completing the proof.

PROOF OF LEMMA 3.2. Choose $\xi > 0$ so that $\xi \leqslant \varepsilon/2$ and

$$\xi < \tfrac{1}{8}\min_{i,\, j:\, s_i \neq s_j}|s_i - s_j|. \tag{4.5}$$

For $\alpha > 0$ and $\theta = 0, 1, \cdots, N-1$ define the sets $\psi(\theta, \alpha) \in \mathcal{B}_B^{LN}$ and

$$\tilde{\psi}(\theta, \alpha) \in \mathcal{B}_B^m, \qquad m = (L+1)N,$$

$$\psi(\theta, \alpha) = \left\{ y^{LN} : \left| \frac{1}{L-1} \Sigma_{i=0}^{L-2} 1_S(y_{j+iN}^r) - s_{\theta+j} \right| \leqslant \alpha; \qquad j = 0, \cdots, N-1 \right\}$$

$$\tilde{\psi}(\theta, \alpha) = B^\theta \times \Psi(\theta, \alpha) \times B^{N-\theta}.$$

From the ergodic theorem $L$ can be chosen large enough so that

$$\eta\left(\cap_{\theta=0}^{N-1} T^{-\theta} c(\psi(\theta, \xi))\right) = \eta^m\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi)\right) \geqslant 1 - \xi^2. \tag{4.6}$$

Assume also that $L$ is large enough so that if $x_i = x_i'$, $i = 0, \cdots, m-1$ then

$$\bar{d}_m(\nu_x^m, \nu_{x'}^m) \leqslant (\xi/N)^2. \tag{4.7}$$

From (4.6)

$$\xi^2 \geqslant \eta^m\left(\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi)\right)^c\right) = \Sigma_{a^m \in G^m} \int_{c(a^m)} d\mu(u) \nu_{\gamma(u)}^m\left(\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi)^c\right)\right)$$

$$:= \Sigma_{a^m \in G^m} \mu^m(a^m) \hat{\nu}\left(\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi)\right)^c | \gamma_m(a^m)\right)$$

and hence there must be a set $\Phi \in \mathcal{B}_B^m$ such that

$$\begin{aligned} &\hat{\nu}^m\left(\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi)\right)^c | \gamma_m(a^m)\right) \leqslant \xi, a^m \in \Phi \\ &\mu^m(\Phi) \leqslant \xi. \end{aligned} \tag{4.8}$$

Define the synch locating function $\sigma : B^{LN} \to \{0, 1, \cdots, N-1\}$ by

$$\sigma(y^{LN}) = \begin{cases} \theta & y^{LN} \in (\psi(\theta, \xi))_{2\xi/N} := \psi(\theta) \\ 1 & \text{otherwise.} \end{cases}$$

We show that $\sigma$ is well defined by showing that $\psi(\theta) \subset \psi(\theta, 4\xi)$, which sets are disjoint for $\theta = 0, 1, \cdots, N-1$ from (4.5): if $y^{LN} \in \psi(\theta)$, there is a $b^{LN} \in \psi(\theta, \xi)$ for which $d_{LN}(y^{LN}, b^{LN}) \leqslant 2\xi/N$ and hence for any $j \in \{0, 1, \cdots, N-1\}$ at most $LN(2\xi/N) = 2\xi L$ of the consecutive nonoverlapping $N$-tuples $y_{j+iN}^N$, $i = 0, \cdots, L-2$, can differ from the corresponding $b_{j+iN}^N$ and therefore

$$\left| \frac{1}{L-1} \Sigma_{i=0}^{L-2} 1_S(y_{j+iN}^r) - s_{\theta+j} \right| \leqslant \left| \frac{1}{L-1} \Sigma_{i=0}^{L-2} 1_S(b_{j+iN}^r) - s_{\theta+j} \right| + 2\xi \leqslant 3\xi$$

and hence $y^{LN} \in \psi(\theta, 4\xi)$. If $\tilde{\psi}(\theta) = B^\theta \times \psi(\theta) \times B^{N-\theta} \in \mathcal{B}_B^m$, then we also have that $(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi))_{\xi/N} \subset \cap_{\theta=0}^{N-1} \tilde{\psi}(\theta)$ since if $y^n \in (\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi))_{\xi/N}$, then there is a $b^m$ such that $b_\theta^{LN} \in \psi(\theta, \xi)$, $\theta = 0, 1, \cdots, N-1$ and $d_m(y^m, b^m) \leqslant \xi/N$ for $\theta = 0, 1, \cdots, N-1$. This implies from Lemma 2.1, (4.7) and (4.8) that if $x \in \gamma^m(a^m)$ and $a^m \in \Phi$

$$\begin{aligned} &\nu_x^m\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta)\right) \\ &\qquad \geqslant \nu_x^m\left(\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi)\right)_{\xi/N}\right) \geqslant \hat{\nu}\left(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta, \xi) | \gamma^m(a^m)\right) - \xi/N \\ &\qquad \geqslant 1 - \xi - \xi/N \geqslant 1 - \varepsilon. \end{aligned} \tag{4.9}$$

To complete the proof we use (3.1)–(3.2) and (4.9) to obtain for $a^m \in \Phi$ and $\gamma_m(a_{NL}^N) = \sigma \times \mathbf{w}_i$,

$$\nu_x\big(y : \sigma(y_\theta^{LN}) = \theta, \theta = 0, 1, \cdots, N-1; y_{LN}^N \in S \times W_i\big) \geqslant \nu_x^m\big(\cap_{\theta=0}^{N-1} \tilde{\psi}(\theta)\big) - \nu_{T^{-NL}x}^N(S \times W_i^c) \geqslant 1 - \varepsilon - 2\varepsilon.$$

PROOF OF LEMMA 3.3. Choose $R$, $H(\mu) < R < C$, and fix $\varepsilon > 0$ so that $\varepsilon \leqslant \delta/3$ and $\varepsilon \leqslant (R - H(\mu))/2$. Choose $N \geqslant 3$ so large that the conditions and conclusions of Corollary 3.1 hold. Construct a block encoder $\gamma_N$ as follows: From the Shannon-McMillan theorem (e.g., Ash (1965), page 197), given the set

$$\mathcal{S}_N(\mu, \varepsilon) = \{u^N : |-N^{-1} \log \mu^N(u^N) - H(\mu)| \leqslant \varepsilon\}, \tag{4.10a}$$

there is an $n_0$ so large that for $N \geqslant n_0$

$$\mu^N(\mathcal{S}_N(\mu, \varepsilon)) \geqslant 1 - \varepsilon. \tag{4.10b}$$

Note also that if $M' = |\mathcal{S}_N(\mu, \varepsilon)|$, then

$$2^{N(H(\mu)-\varepsilon)} \leqslant M' \leqslant 2^{N(H(\mu)+\varepsilon)} \leqslant 2^{N(R-\varepsilon)}. \tag{4.10c}$$

Index the members of $\mathcal{S} = \mathcal{S}_N(\mu, \varepsilon)$ as $\boldsymbol{\mu}_i$, $i = 1, 2, \cdots, M'$. If $u^N = \mathbf{u}_1$, set $\gamma_N(u^N) = \sigma \times \mathbf{w}_i$, otherwise set $\gamma_N(u^N) = \sigma \times \mathbf{w}_{M'+1}$. Since, for large $N$, $2^{N(R-\varepsilon)} + 1 \leqslant 2^{NR}$, $\gamma_N$ is well defined. Define also the block decoder $\psi_N(y^N) = \mathbf{u}_i$ if $y^N \in S \times W_i$, $i = 1, \cdots, M'$, otherwise set $\psi_N(y^N) = \mathbf{u}^*$, an arbitrary reference vector. Choose $L$ so large that the conditions and conclusions of Lemma 3.2 hold for $\mathcal{C}$ and $\gamma_N$. The sliding-block decoder $g_m : B^m \to G$, $m = (L+1)N$, yielding decoded process $\hat{U}_k = g_m(Y_{k-NL}^m)$, is defined as follows: If $\sigma(y_{k-NL}, \cdots, y_{k-1}) = \theta$, form $b^N = \psi_N(y_{k-\theta}, \cdots, y_{k-\theta+N})$ and set $\hat{U}_k(y) = g_m(y_{k-NL}, \cdots, y_{k+N}) = b_\theta$, the appropriate symbol of the appropriate block. The encoder $f$ will send very long sequences of block words with random spacing (via the R-K theorem) to make the code stationary. Let $K$ be a large integer satisfying $\varepsilon K \geqslant (L+1)$ so that $m \leqslant \varepsilon KN$ and $N \geqslant 3$ and $L \geqslant 1$

$$(KN)^{-1} \leqslant (3K)^{-1} \leqslant \varepsilon/6. \tag{4.11}$$

Imbed the block code $\gamma_{KN}$ in an $(NK, \varepsilon)$-gadget to obtain an infinite length sliding-block code as in (2.5)–(2.8). We have, defining the error event $\{y : \hat{U}_0(y) \neq U_0(u)\} = \mathcal{E}_u$,

$$\begin{aligned} P_e(\mu, \nu, f, g_m) &= \Pr(U_0 \neq \hat{U}_0) = \int d\mu(u) \nu_{f(u)}(y : U_0(u) \neq \hat{U}_0(y)) \\ &\leqslant \textstyle\sum_{i=0}^{LN-1} \int_{T^iF} d\mu(u) \nu_{f(u)}(\mathcal{E}_u) + \sum_{i=LN}^{KN-1} \int_{T^iF} d\mu(u) \nu_{f(u)}(\mathcal{E}_u) \\ &\quad + \textstyle\int_{(\cup_{i=0}^{KN-1} T^iF)^c} d\mu(u) \\ &\leqslant LN\mu(F) + \textstyle\sum_{i=LN}^{KN-1} \int_{T^iF} d\mu(u) \nu_{f(u)}(\mathcal{E}_u) + \varepsilon \\ &\leqslant 2\varepsilon + \textstyle\sum_{i=LN}^{KN-1} \sum_{a^{KN} \in G^{KN}} \int_{u' \in T^i(F \cap c(a^{KN}))} d\mu(u') \\ &\quad \times \nu_{f(u')}(y' : U_0(u') \neq \hat{U}_0(y')), \end{aligned} \tag{4.12}$$

where we have used the fact that $\mu(F) \leqslant (KN)^{-1}$, and hence $LN\mu(F) \leqslant L/K \leqslant \varepsilon$. Fix $i = kN + j$, $0 \leqslant j \leqslant N - 1$, and define $u = T^{-(j+LN)}u'$, $y = T^{-(j+LN)}y'$, and the above integrals become

$$
\begin{aligned}
(4.13)\quad & \int_{u' \in T^i(F \cap c(a^{KN}))} d\mu(u') \nu_{f(u')}(y' : U_0(u') \neq g_m(Y^m_{-NL}(y'))) \\
&= \int_{T^{(k-L)N}(F \cap c(a^{KN}))} d\mu(u) \nu_{f(T^{j+LN}u)}(y : U_0(T^{j+LN}u) \neq g_m(Y^m_{-NL}(T^{j+LN}y))) \\
&= \int_{T^{(k-L)N}(F \cap c(a^{KN}))} d\mu(u) \nu_{f(T^{j+LN}u)}(y : u_{j+LN} \neq g_m(y_j^m)) \\
&\leqslant \int_{T^{(k-L)N}(F \cap c(a^{KN}))} d\mu(u) \nu_{f(T^{j+LN}u)}(y : u^N_{LN} \neq \psi_N(y^N_{LN}) \text{ or } \sigma(y_j^{LN}) \neq j).
\end{aligned}
$$

If $u^N_{LN} = \mathbf{u}_i \in \mathcal{S}_N(\mu, \varepsilon)$, then $u^N_{LN} = \psi_N(y^N_{LN})$ if $y^N_{LN} \in S \times W_i$. If $u \in T^{(k-L)N}c(a^{KN})$, then $u^m = a^m_{(k-L)N}$, and hence from Lemma 3.2 stationarity, (2.5)–(2.8) and (4.11) we have for $i = kN + j$ that

$$
\begin{aligned}
&\sum_{a^{KN} \in G^{KN}} \int_{T^i(c(a^{KN}) \cap F)} d\mu(u) \nu_{f(u)}(\mathcal{E}_u) \\
&\quad \leqslant 3\varepsilon \sum_{\substack{a^{KN} \in G^{KN} \\ a^m_{(k-L)N} \in \Phi \cap (G^{LN} \times \mathcal{S}_N(\mu,\varepsilon))}} \mu(T^{(k-L)N}(c(a^{KN}) \cap F)) \\
&\qquad + \sum_{\substack{a^{KN} \in G^{KN} \\ a^m_{(k-L)N} \notin \Phi \cap (G^{LN} \times \mathcal{S}_N(\mu,\varepsilon))}} \mu(T^{(k-L)N}(c(a^{KN}) \cap F)) \\
(4.14)\qquad &\quad = 3\varepsilon \sum_{\substack{a^{KN} \in G^{KN} \\ a^m_{(k-L)N} \in \Phi \cap (G^{LN} \times \mathcal{S}_N(\mu,\varepsilon))}} \mu^{KN}(a^{KN})\mu(F) \\
&\qquad + \sum_{\substack{a^{KN} \in GK^N \\ a^m_{(k-L)N} \notin \Phi \cap (G^{LN} \times \mathcal{S}_N(\mu,\varepsilon))}} \mu^{KN}(a^{KN})\mu(F) \\
&\quad \leqslant 3\varepsilon(KN)^{-1} + \mu^m(\Phi^c)(KN)^{-1} + \mu^N(\mathcal{S}_N(\mu, \varepsilon)^c)(KN)^{-1} \\
&\quad \leqslant 5\varepsilon(KN)^{-1} \leqslant 5\varepsilon^2/6 \leqslant \varepsilon,
\end{aligned}
$$

and hence from (4.14)

$$P_e(\mu, \nu, f, g_m) \leqslant 2\varepsilon + \varepsilon = 3\varepsilon \leqslant \delta,$$

as was to be proved.

PROOF OF COROLLARY 3.2. Construct via Lemma 3.3 an infinite length encoder $f$ and a finite length decoder $g_m$ such that $P_e(\mu, \nu, f, g_m) \leqslant \delta$. From Gray (1975), given $\varepsilon > 0$ there exists for sufficiently large $q$ a finite length sliding-block code $f_q : G^{2q+1} \to A$ such that $P_r(f \neq f_q) = \int d\mu(u) d_1(f_q(U^{2q+1}_{-q}(u)), f(u)) \leqslant \varepsilon$. By choosing $\varepsilon$ sufficiently small the corollary follows from Corollary A.1 of the Appendix.

PROOF OF THEOREM 3.1. Assume $N$ is large enough for Corollary 3.1 and (4.10) to hold. From the Nedoma decomposition, we have

$$M^{-1}\sum_{i=0}^{M-1} \mu^N(\mathcal{S} | T^i\Pi) = \mu^N(\mathcal{S}) \geqslant 1 - \varepsilon,$$

and hence there exists at least one $i$ such that

$$\mu^N(\mathcal{S} | T^i\Pi) \geqslant 1 - \varepsilon,$$

that is, at least one $N$-ergodic mode must put high probability on the set $\mathcal{S}$ of "typical" $N$-tuples of $\mu$. For convenience, relabel the indices so that the good mode

is $\mu(\cdot|\Pi)$ and call $\mu(\cdot|\Pi)$ the design mode. Since $\mu(\cdot|\Pi)$ is $N$-ergodic and $N$-stationary, Lemma 3.2 holds with $\mu$ replaced by $\mu(\cdot|\Pi)$, that is, there is a source/channel block code $(\gamma_N, \psi_N)$ and a synch locating function $\sigma: B^{LN} \to \{0, 1, \cdots, N-1\}$ such that there is a set $\Phi \in \mathcal{B}_G^m$, $m = (L+1)N$, for which (3.6) holds and

$$\mu^m(\Phi|\Pi) \geqslant 1 - \varepsilon.$$

The sliding-block decoder $g_m$ is constructed exactly as in Lemma 3.2. The infinite sliding-block encoder, however, is somewhat different. Instead of only carving up the base $F$ of the gadget according to $KN$-tuples, we also carve it up according to $T^i\Pi$. Define the partition $P$ of $G^\infty$ by $P = \{P_i;\ i = 0, \cdots, M\}$ by $P_i = T^i\Pi$, $i = 0, \cdots, M-1$, and $P_M = (\cup_{i=0}^{M-1} T^i\Pi)^C$. We have $\mu(P_M) = 0$, but $P_M$ must be included in the code construction for the code to be well defined.

Again from the R-K theorem given $KN$ and $\varepsilon > 0$, there is a base $F$ such that

(1) $F, TF, \cdots, T^{KN}F$ are disjoint,

(2) $\mu(\cup_{k=0}^{KN-1} T^k F) \geqslant 1 - \varepsilon$,

(3) $\mu(c(u^{KN}) \cap P_i|F) = \mu(c(u^{KN}) \cap P_i)$, all $u^{KN} \in G^{KN}$ and $i = 0, \cdots, M$.

In particular,

$$(4.15)\quad \mu\big(c(u^{KN}) \cap T^i\Pi|F\big) = \mu\big(c(u^{KN}) \cap T^i\Pi\big) = \mu\big(c(u^{KN})|T^i\Pi\big)\mu(T^i\Pi) = M^{-1}\pi_i^{KN}(u^{KN}).$$

The infinite sliding-block encoder $f$ is defined as follows: If $u \notin \cup_{k=0}^{NK-1} T^k F$, set $f(u) = a^*$. If $u \in \cup_{k=0}^{NK-1} T^k(F \cap P_M)$, set $f(u) = a^*$. If $u \in T^i(F \cap c(u^{KN}) \cap T^{-j}\pi)$, then (1) if $i < j$, set $f(u) = a^*$ (these are spacing symbols to force the right mode), (2) if $j \leqslant i \leqslant KN - (M-j)$, then we can write $i = j + kN + r$, where $0 \leqslant k \leqslant (K-1)N$, $0 \leqslant r \leqslant N-1$; form $\gamma_N(u_{j+kN}^N) = a^N$ and set $f(u) = a_r$; this is essentially the same as before, except that, if $u \in T^{-j}\Pi$, then we do not start block encoding until the $j$th symbol, in which position we are in $T^j(T^{-j}\Pi) = \Pi$; (3) if $KN - (M-j) \leqslant i \leqslant KN - 1$, then $f(u) = a^*$. We have, as in the proof of Lemma 3.3 that

$$(4.16)\quad P_e(\mu, \nu, f, g_m) = \int d\mu(u)\nu_{f(u)}\big(y: U_0(u) \neq g_m(Y_{-LN}^m(y))\big) \leqslant 2\varepsilon + \sum_{i=LN}^{KN-1}\int_{u \in T^iF} d\mu(u)\nu_{f(u)}\big(y: U_0(u) = \hat{U}_0(y)\big).$$

Since $\mu(P_M) = 0$, the sum is equal to

$$\sum_{i=LN}^{KN-1}\sum_{j=0}^{M-1}\sum_{a^{KN} \in G^{KN}}\int_{u \in T^i(c(a^{KN}) \cap F \cap T^{-j}\Pi)} d\mu(u)\nu_{f(u)}\big(y: U_0(u) \neq \hat{U}_0(y)\big)$$

$$\leqslant \sum_{j=0}^{M-1}\sum_{i=LN+j}^{KN-(M-j)}\sum_{a^{KN} \in G^{KN}}\int_{u \in T^i(c(a^{KN}) \cap F \cap T^{-j}\Pi)} d\mu(u)\nu_{f(u)}\big(y: U_0(u) \neq \hat{U}_0(y)\big) + \sum_{j=0}^{M-1} M\mu(F \cap T^{-j}\Pi),$$

where the second term is

$$M\sum_{j=0}^{M-1}\mu(F)\mu(T^{-j}\Pi) \leqslant M/(KN) \leqslant 1/K \leqslant \varepsilon$$

whence from (4.16)

$$P_e(\mu,\nu,f,g_m) \leqslant 3\varepsilon + \sum_{j=0}^{M-1}\sum_{i=LN+j}^{KN-(M-j)}\sum_{a^{KN}\in G^{KN}}\int_{u\in T^i(c(a^{KN})\cap F\cap T^{-j}\Pi)} d\mu(u) \times \nu_{f(u)}(y : U_0(u) \neq \hat{U}(y)). \tag{4.17}$$

Analogous to (4.13) (except here $i = j + kN + r$, $u = T^{-(LN+r)}u'$)

$$\int_{u'\in T^i(c(a^{KN})\cap F\cap T^{-j}\Pi)} d\mu(u')\nu_{f(u')}(y' : U_0(u') = g_m(Y_{-LN}^m(y'))) \leqslant \int_{T^{j+(k-L)N}(c(a^{KN})\cap F\cap T^{-j}\Pi)} d\mu(u)\nu_{f(T^{r+LN}u)}(y : u_{LN}^N \neq \psi_N(y_{LN}^N) \quad \text{or} \quad \sigma(y_r^{LN}) \neq r)$$

and hence since $u \in T^{j+(k-L)N}(c(a^{KN}) \cap F \cap T^{-j}\Pi)$ implies $u^m = a_{j+(k-L)N}^m$, we have analogous to (4.14) that for $i = j + kN + r$

$$\begin{aligned}
&\sum_{a^{KN}\in G^{KN}}\int_{T^i(c(a^{KN})\cap F\cap T^{-j}\Pi)} d\mu(u)\nu_{f(u)}(y : U_0(u) = g_m(Y_{-LN}^m(y))) \\
&\quad = \varepsilon\sum_{a^{KN}: a_{j+(k-L)N}^m\in\Phi}\mu\big(T^{j+(k-L)N}(c(a^{KN}) \cap F \cap T^{-j}\Pi)\big) \\
&\qquad + \sum_{a^{KN}a_{j+(k-L)N}^m\notin\Phi}\mu\big(T^{j+(k-L)N}(c(a^{KN}) \cap F \cap T^{-j}\Pi)\big) \\
&\quad = \varepsilon\sum_{a^{KN}: a_{j+(k-L)N}^m\in\Phi}\mu\big(c(a^{KN}) \cap F \cap T^{-j}\Pi\big) \\
&\qquad + \sum_{a^{KN}: a_{j+(k-L)N}^m\notin\Phi}\mu\big(c(a^{KN}) \cap F \cap T^{-j}\Pi\big).
\end{aligned} \tag{4.18}$$

From the R-K theorem, we have $\mu(c(a^{KN}) \cap F \cap T^{-j}\Pi) = \mu(c(a^{KN}) \cap T^{-j}\Pi) \times \mu(F) = \mu(c(a^{KN})|T^{-j}\Pi)\mu(\Pi)\mu(F)$, and hence the above becomes

$$\begin{aligned}
&\varepsilon\mu\big(T^{-(j+(k-L)N)}c(\Phi)|T^{-j}\Pi\big)\mu(\Pi)\mu(F) \\
&\qquad + \mu\big(T^{-(j+(k-L)N)}c(\Phi^c)|T^{-j}\Pi\big)\mu(\pi)\mu(F) \\
&\quad = \varepsilon\mu(c(\Phi)|\Pi)\mu(\Pi)\mu(F) + \mu(c(\Phi^c)|\Pi)\mu(\Pi)\mu(F) \leqslant 2\varepsilon M^{-1}(KN)^{-1},
\end{aligned}$$

which with (4.17) and (4.18) yields

$$P_e(\mu,\nu,f,g_m) \leqslant 3\varepsilon + MKN2\varepsilon M^{-1}(KN)^{-1} \leqslant 5\varepsilon. \tag{4.19}$$

The theorem follows from (4.19) as in the proof of Corollary 3.2.

PROOF OF LEMMA 3.4. Define a decoder $h_m$ yielding decoded process $\hat{X}_n(y) = h_m(Y_{n-LN}^m(y))$ as follows: if $\sigma(y_{-NL}, \cdots, y_{-1}) = \theta$, and $y_\theta^N \in S \times W_i$, set $a^N = \sigma \times \mathbf{w}_i$ and assign $h_m(y_{-NL}^n) = a_\theta$, otherwise set $h_m(y_{-NL}^m) = a^*$, an arbitrary reference symbol. In other words, $h_m$ operates like $g_m$ in Lemma 3.3 except that it prints the code symbol, not the original source symbol. Define the block decoder $\lambda_N : B^N \to A^N$ by $\lambda_N(y^N) = \sigma \times \mathbf{w}_i$ if $y^N \in S \times W_i$ and, say, $\lambda_N(y^N) = a^{*N}$ otherwise. Analogous to (4.12) and (4.13),

$$\begin{aligned}
P_e(\mu\tilde{f}^{-1}, \nu, i, h_m) &= \Pr(X_0 \neq \hat{X}_0) \\
&\leqslant 2\varepsilon + \sum_{k=L}^{K}\sum_{j=0}^{N-1}\sum_{a^{KN}\in G^{KN}}\int_{T^{(k-L)}(F\cap c(a^{KN}))} d\mu(u)\nu_{f(T^{j+LN}u)} \\
&\qquad \cdot(y : \gamma_N(u_{LN}^N) \neq \lambda_N(y_{LN}^N) \quad \text{or} \quad \sigma(y_j^{LN}) \neq j).
\end{aligned} \tag{4.20}$$

If $\gamma_N(u_{LN}^N) = \sigma \times w_i$, then $\gamma_N(u_{LN}^N) = \lambda_N(y_{LN}^N)$ if $y_{LN}^N \in S \times W_i$. If $u \in T^{(k-L)N}c(a^{KN})$, then $u^m = a_{(k-L)N}^m$. Thus, analogous to (4.14)

$$\Sigma_{a^{KN} \in G^{KN}} \int_{T^{(k-L)}(F \cap c(a^{KN}))} d\mu(u) \nu_{f(T^{jN+LN}u)}$$

$$\cdot \left(y : \gamma_N(u_{LN}^N) \neq \lambda_N(y_{LN}^N) \quad \text{or} \quad \sigma(y_j^{LN}) \neq j\right) \tag{4.21}$$

$$\leqslant 3\varepsilon\mu(F) + \mu^m(\Phi^c)(KN)^{-1} \leqslant 4\varepsilon(KN)^{-1} \leqslant 4\varepsilon^2/m \leqslant \varepsilon,$$

and hence

$$P_e(\mu\bar{f}^{-1}, \nu, i, h_m) \leqslant 3\varepsilon.$$

PROOF OF COROLLARY 3.3. Let $[A, \mu, U]$ be an independent identically distributed (i.i.d.) process with $H(\mu) = H < C$ (such a process always exists), and assume that $\delta$ is so small that

$$h(\delta) + \delta \log|A| < \Delta.$$

This entails no loss of generality, as a $\delta'$-invulnerable source with $\delta' \leqslant \delta$ is also $\delta$-invulnerable. Repeat the proof of Lemma 3.4 using the $\gamma_N$ defined in the proof of Lemma 3.3 and choosing $\varepsilon \leqslant \delta/3$ as in the proof of Lemma 3.3. Construct both the decoder $g_m$ for the original source and the decoder $h_m$ for the encoded source. From Lemma 3.4

$$P_e(\mu\bar{f}^{-1}, \nu, i, h_m) \leqslant 3\varepsilon \leqslant \delta,$$

and from Lemma 3.3

$$P_e(\mu, \nu, f, g_m) \leqslant \delta.$$

From the data processing theorem (Billingsley (1965), Theorem 17.3) if $\tau = \mu\bar{f}^{-1}$,

$$H(\tau) = H(X) \geqslant I(X, \hat{U}) \geqslant I(U; \hat{U}),$$

and from Shields' (1973) Lemma 8.2

$$H(U|\hat{U}) := H(U) - I(U; \hat{U}) \leqslant h(P_e(\mu, \nu, f, g_m)) + P_e(\mu, \nu, f, g_m) \log|A|$$
$$\leqslant h(\delta) + \delta \log|A| \leqslant \Delta,$$

and hence

$$H(\tau) \geqslant H(\mu) - \Delta = H - \Delta.$$

This implies that

$$\inf_{\delta > 0} \sup_{\delta\text{-invulnerable } [A, \tau, X]} H(\tau) \geqslant C.$$

Again invoking Shields' (1973) Lemma 8.2 and Billingsley's (1965), Theorem 17.3, we have

$$H(X) - I(X; \hat{X}) := H(X|\hat{X}) \leqslant h(\delta) + \delta \log|A|$$
$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) \leqslant I(X; Y) \leqslant C$$

and therefore

$$H(X) = H(\tau) \leqslant C + h(\delta) + \delta \log|A|.$$

Since

$$\sup_{\delta\text{-invulnerable } [A, \tau, X]} H(\tau)$$

is nonincreasing as $\delta \to 0$,

$$\begin{aligned}\inf_{\delta>0} \sup_{\delta\text{-invulnerable } [A, \tau, X]} H(\tau) &= \lim_{\delta\to 0} \sup_{\delta\text{-invulnerable } [A, \tau, X]} H(\tau) \\ &\leqslant \lim_{\delta\to 0}\{C + h(\delta) + \delta \log|A|\} = C,\end{aligned}$$

which completes the proof.

The proof of Theorem 3.2 is based on the following iteration lemma.

LEMMA 4.1. *Let $\nu$ be a stationary totally ergodic $\bar{d}$-continuous channel and $[G, \mu, U]$ an* i.i.d. *source. Assume we have for $k \geqslant 1$ an infinite-length sliding-block encoder $f^{(k)} : G^\infty \to A$ and a length $m_k$ sliding-block decoder $g^{(k)} : B^{m_k} \to A$ such that*

$$P_e\left(\mu \overline{f^{(k)}}^{-1}, \nu, i, g^{(k)}\right) \leqslant \varepsilon_k,$$

*that is, the process $[A, \tau^{(k)}, X^{(k)}]$ defined by $\tau^{(k)} = \mu\overline{f^{(k)}} - 1$ (or, equivalently, $X_i^{(k)}(u) = f^{(k)}(T^i u)$ is $\varepsilon_k$-invulnerable). Assume also we have for an integer $n_k$ a collection $\mathcal{G}_{n_k} \in A^{n_k}$ called "good $n_k$-blocks" such that the relative frequency of nonoverlapping good $n_k$-blocks in $\tau^{(k)}$ is greater than $(1 - \varepsilon_k)n_k^{-1}$, that is, if for any integer $M$, $Z_M^{(k)}(x^M) =$ the number of nonoverlapping good $n_k$-blocks in $x^M$, then with $\tau^{(k)}$-probability one*

$$\lim_{M\to\infty} M^{-1} Z_M^{(k)}(X^{(k)M}) \geqslant (1 - \varepsilon_k)/n_k.$$

The lemma states that, given $\varepsilon_{k+1} > 0$, there is an encoder $f^{(k+1)} : G^\infty \to A$ and for sufficiently large $m_{k+1}$ a decoder $g^{(k+1)} : B^{m_{k+1}} \to A$ such that, if $\tau^{(k+1)} = \mu\overline{f^{(k+1)}}^{-1}$ (or $X_n^{(k+1)}(u) = f^{(k+1)}(T^n u)$), then

$$P_e\left(\tau^{(k+1)}, \nu, i, g^{(k+1)}\right) \leqslant \varepsilon_{k+1}, \tag{4.22}$$

and hence $[A, \tau^{(k+1)}, X^{(k+1)}]$ is $\varepsilon_{k+1}$-invulnerable, and

$$\Pr(X^{(k)} \neq X^{(k+1)}) := |f^{(k+1)} - f^{(k)}|_\mu \leqslant 7\varepsilon_k, \tag{4.23}$$

and defining $H_k = H(\tau^{(k)})$,

$$H_k(1 - \varepsilon_k) \geqslant H_{k+1} \geqslant H_k - h(7\varepsilon_k) - 7\varepsilon_k \log|A|, \tag{4.24}$$

and with $\tau^{(k+1)}$-probability one

$$\lim_{M\to\infty} M^{-1} Z_M^{(k)}(X^{(k+1)M}) \geqslant (1 - \varepsilon_k - \varepsilon_{k+1})/n_k. \tag{4.25}$$

In the proof a set $\mathcal{G}_{n_{k+1}}$ of good $n_{k+1}$-tuples is constructed such that with $\tau^{(k+1)}$-probability one

$$\lim_{M\to\infty} M^{-1} Z_M^{(k+1)}(X^{(k+1)M}) \geqslant (1 - \varepsilon_{k+1})/n_{k+1}, \tag{4.26}$$

which sets up the induction used in the proof of Theorem 3.2.

*Comment.* Equation (4.22) alone is not sufficient to obtain a limiting code with the required properties. Both (4.23) and (4.25) are to ensure that the new code will

produce a new process looking much like the old process (but more invulnerable), and (4.24) is necessary to obtain a limit process with entropy exactly $H^*$ rather than simply close to $H$.

PROOF. Assume for simplicity that $g^{(k)}$ is symmetric and $m_k = 2q + 1$. Choose $\varepsilon$ so small that

$$\begin{aligned} \varepsilon &\leqslant \varepsilon_k/m_k \leqslant \varepsilon_k/3 \\ \varepsilon &\leqslant \varepsilon_{k+1}/7 \\ \varepsilon &\leqslant \varepsilon_k H_k/1 + \log|A| + 6\varepsilon_k \end{aligned} \tag{4.27}$$

where $H_k := H(\tau^{(k)})$. Given $[A, \tau^{(k)}, X^{(k)}]$, we begin by constructing a set of "good" input blocks called good $n_{k+1}$-blocks to be synchronized and imbedded in a gadget. Define $\mathcal{S}_t(\tau^{(k)}, \varepsilon)$, as in (4.10a) and choose $t$ so large that

$$\tau^{(k)t}\big(\mathcal{S}_t(\tau^{(k)}, \varepsilon)\big) \geqslant 1 - \varepsilon/2,$$

and recall

$$2^{t(H_k-\varepsilon)} \leqslant |\mathcal{S}_t(\tau^{(k)}, \varepsilon)| \leqslant 2^{t(H_k+\varepsilon)}. \tag{4.28}$$

Since the process is totally ergodic, we can choose $n$ and hence $s$ so large that, for $n = n_{k+1} = st \geqslant \bar{n}$ and

$$\mathcal{G}_n^{(1)} = \Big\{x^n : s^{-1}\textstyle\sum_{i=0}^{s-1} \chi_{\mathcal{S}_t}(\tau^{(k)}, \varepsilon)\big(x_{it}^t\big) \geqslant 1 - \varepsilon\Big\}, \tag{4.29}$$

we have

$$\tau^{(k)n}\big(\mathcal{G}_n^{(1)}\big) \geqslant 1 - \varepsilon/9. \tag{4.30}$$

Since $\tau^{(k)}\nu$ is ergodic, we can also choose $\bar{n}$ so large that, for $n = n_{k+1} \geqslant n$,

$$(\tau^{(k)}\nu)^n\left(x^n, y^n : \frac{1}{n-2q}\textstyle\sum_{i=q}^{n-q} d_1\big(x_i, g^{(k)}\big(y_{i-q}^{m_k}\big)\big) > 2\varepsilon_k\right) \leqslant (\varepsilon/9)^2, \tag{4.31}$$

and hence defining $\hat{\nu}(y^n|x^n) = (\tau^{(k)}\nu)^n(x^n, y^n)/\tau^{(k)}(x^n)$ there must be a set $\mathcal{G}_n^{(2)} \in \mathcal{B}_B^n$ such that, for $x^n \in \mathcal{G}_n^{(2)}$,

$$\hat{\nu}\left(y^n : \frac{1}{n-2q}\textstyle\sum_{i=q}^{n-q} d\big(x_i, g^{(k)}\big(y_{i-q}^{m_k}\big)\big) > 2\varepsilon_k | x^n\right) \leqslant \varepsilon/9, \tag{4.32}$$

where

$$\tau^{(k)n}\big(\mathcal{G}_n^{(2)}\big) \geqslant 1 - \varepsilon/9. \tag{4.33}$$

Finally, by assumption, we can choose $\bar{n}$ so large that, if $n = n_{k+1} \geqslant \bar{n}$ and if

$$\mathcal{G}_n^{(3)} := \big\{x^n : n^{-1}Z_n^{(k)}(x^n) \geqslant n_k^{-1}(1 - \varepsilon_k)\big\}, \tag{4.34}$$

then

$$\tau^{(k)n}\big(\mathcal{G}_n^{(3)}\big) \geqslant 1 - \varepsilon/9,$$

for any $n = n_{k+1} \geqslant \bar{n}$. The set $\mathcal{G}_n$ defined by

$$\mathcal{G}_{n_{k+1}} = \cap_{i=1}^{3} \mathcal{G}_{n_{k+1}}^{(i)}$$

is called the set of good $n_{k+1}$-blocks, and from (4.30), (4.33), and (4.35), for $n = n_{k+1} \geq \bar{n}$,

$$\tau^{(k)n}(\mathcal{G}_n) \geq 1 - \varepsilon/3, \tag{4.36}$$

and, if $x^n \in \mathcal{G}_n$, then (4.29), (4.31), and (4.34) all hold. Assume also that $\bar{n}$ is so large that $n_{k+1} \geq \bar{n}$ implies

$$m_k/n_{k+1} \leq \varepsilon. \tag{4.37}$$

A set $\mathcal{C} = \{\mathbf{v}_i,\ i = 1, \cdots, |\mathcal{C}|\}$ is called a full $\alpha$-separating subset of $\mathcal{G}_n$ if $\mathbf{v}_i \in \mathcal{G}_n$, all $i$, if

$$d_n(\mathbf{v}_i, \mathbf{v}_j) > \alpha, \qquad i \neq j,$$

and, if $x^n \in \mathcal{G}_n$, then there is a $\mathbf{v}_i$ for which $d_n(\mathbf{v}_i, x^n) \leq \alpha$ (such a set can always be constructed by choosing and eliminating). Let $\mathcal{C}$ be a full $6\varepsilon_k$-separating subset of $\mathcal{G}_n$. For $x^n = \mathbf{v}_i \in \mathcal{C}$, define for $\beta > 0$ the set

$$\Gamma_i(\beta) = \left\{ y^n : \frac{1}{n-2q} \Sigma_{j=q}^{n-q} d(x_j, \hat{x}_j) \leq \beta \right\},$$

where $\hat{x}_j = g^{(k)}(y_{j-q}^{m_k})$, and note that $d_n(x^n, \hat{x}^n) \leq n^{-1}[(n-2q)\beta + 2q] \leq \beta + \varepsilon$. If $\hat{x}^n \in (\Gamma_i(2\varepsilon_k))_\varepsilon$ and $\tilde{x}^n \in (\Gamma_j(2\varepsilon_k))_\varepsilon$, then $d_n(\mathbf{v}_i, \hat{x}^n) \leq 2\varepsilon_k + \varepsilon$ and $d_n(\mathbf{v}_j, \tilde{x}^n) \leq 2\varepsilon_k + \varepsilon$, since $d_n(\mathbf{v}_i, \mathbf{v}_j) > 6\varepsilon_k > 2(2\varepsilon_k + \varepsilon)$, $(\Gamma_i(2\varepsilon_k))_\varepsilon \cap (\Gamma_j(2\varepsilon_k))_\varepsilon = \varphi$, and hence, for $n_{k+1} \geq \bar{n}$, $\{\mathbf{v}_i, \Gamma_i(2\varepsilon_k);\ i = 1, \cdots, |\mathcal{C}|\}$ is an $\varepsilon$-robust $(\tau^{(k)}, |\mathcal{C}|, n_{k+1}, \varepsilon/3)$-Feinstein code.

For $\mathbf{v}_i \in \mathcal{C}$, we have from (4.32) and Lemma 2.1 that

$$\begin{aligned} \inf_{x \in c(\mathbf{v}_i)} \nu_x(\Gamma_i(2\varepsilon_k)) &\geq \inf_{x \in c(\mathbf{v}_i)} \nu_x((\Gamma_i(2\varepsilon_k))_\varepsilon) \\ &\geq \hat{\nu}(\Gamma_i(2\varepsilon_k)|\mathbf{v}_i) - \varepsilon \geq 1 - 10\varepsilon/9, \end{aligned} \tag{4.38}$$

and hence for any $x \in c(\mathbf{v}_i)$, $\mathbf{v}_i \in \mathcal{C}$,

$$\nu_x^n\left( y^n : \frac{1}{n-2q} \Sigma_{i=q}^{n-q} d\left(x_i, g^{(k)}(y_{i-q}^{m_k})\right) > 3\varepsilon_k \right) \leq 10\varepsilon/9,$$

that is, $g^{(k)}$ will do a good job during a good $n_{k+1}$-block regardless of past or future inputs. This means that, for any $x \in A^\infty$,

$$\begin{aligned} E\nu_x\left\{ M^{-1} \Sigma_{i=0}^{M-1} d\left(x_i, g^{(k)}(Y_{i-q}^{m_k})\right)\right\} &\leq M^{-1}\left\{ Z_M^{(k+1)}(x^M) n_{k+1}(3\varepsilon_k + 10\varepsilon/9) \right. \\ &\quad \left. + \left(M - Z_M^{(k+1)}(x^M) n_{k+1}\right)\right\}. \end{aligned}$$

Since by definition $Z_M^{(k+1)}(x^M) \leq M/n_{k+1}$,

$$\begin{aligned} E_{\nu_x}\left\{ M^{-1} \Sigma_{i=0}^{M-1} d\left(x_i, g^{(k)}(Y_{i-q}^{m_k})\right)\right\} \\ \leq 3\varepsilon_k + 2\varepsilon_{k+1}/7 + 1 - M^{-1} n_{k+1} Z_M^{(k+1)}(x^M). \end{aligned} \tag{4.39}$$

Let $n = n_{k+1}$, and let $\alpha_n : A^n \to \mathcal{C}$ map $x^n$ into the $\mathbf{v}_i \in \mathcal{C}$, minimizing $d_n(x^n, \mathbf{v}_i)$ and hence by construction $d_n(x^n, \alpha_n(x^n)) \leq 6\varepsilon_k$. Define $p_i^{(n)} = \tau^{(k)n}(x^n : \alpha_n(x^n) = \mathbf{v}_i) = \tau^{(k)n}(\alpha_n^{-1}(\mathbf{v}_i))$. From Lemma 3.1, given the previous and any $\delta$, we can take $n = n_{k+1}$ large enough so that there is a set of indices $\mathcal{K}$ such that there is an $(\varepsilon, n, \delta, r)$-prefixed codebook $\{\sigma \times \mathbf{w}_i;\ S \times W_i;\ i = 1, \cdots, |\mathcal{W}|\}$ for which $\mathbf{w}_i =$

$\mathbf{v}_{k_i}$ and $W_i \in (\Gamma_i(2\varepsilon_k))_\varepsilon$, and

$$\Sigma_{i\notin\mathcal{K}} p_i^{(n)} = \Sigma_{i\notin\mathcal{K}} \tau^{(k)n}\big(\alpha_n^{-1}(\mathbf{v}_i)\big) \leqslant \varepsilon. \tag{4.40}$$

Assume also that $\delta \leqslant \varepsilon$ is so small and $n_{k+1}$ so large that for $r_{k+1} = \lceil n_{k+1}\delta \rceil$ we have

$$N_{k+1}^{-1} = (n_{k+1} + r_{k+1})^{-1} \geqslant (1 - \varepsilon_{k+1}) n_{k+1}^{-1}. \tag{4.41}$$

Define the reduced code word set $\{\mathbf{w}_i;\ i = 1, \cdots, |\mathcal{J}|\} = \bar{\mathcal{C}} \subset \mathcal{C} \subset \mathcal{G}_n$, and let $\bar{\alpha}_n : A^n \to \bar{\mathcal{C}}$ map $x^n$ into the $\mathbf{w}_i$, minimizing $d_n(x^n, \mathbf{w}_i)$. Note that, if $\alpha_n(x^n) = \mathbf{v}_i$ for $i \in \mathcal{K}$ then $\bar{\alpha}_n(x^n) = \alpha_n(x^n)$, and hence, if $x^n \in \mathcal{G}_n$, then $d_n(x^n, \bar{\alpha}_n(x^n)) \leqslant 6\varepsilon_k$. Define the block encoder $\gamma_N : A^N \to \boldsymbol{\sigma} \times \bar{\mathcal{C}}$ by $\gamma_N(x^N) = \boldsymbol{\sigma} \times \bar{\alpha}_n(x_r^n)$, that is, a synchronized good $n_{k+1}$-block from $\bar{\mathcal{C}} \subset \mathcal{C} \subset \mathcal{G}_n$. We have that

$$\begin{aligned} E_{\tau^{(k)}} d_N\big(X^N, \gamma_N(X^N)\big) &\leqslant r/N + E_{\tau^{(k)}} d_n\big(X^n, \bar{\alpha}_n(X^n)\big) \\ &\leqslant r/N + \tau^{(k)}(\mathcal{G}_n^c) + \Sigma_{i\in\mathcal{K}} \Sigma_{x^n \in \mathcal{G}_n x^n \in \alpha_n^{-1}(\mathbf{v}_i)} d_n(x^n, \mathbf{v}_i)\tau^{(k)n}(x^n) \\ &\quad + \Sigma_{i\notin\mathcal{K}} \tau^{(k)n}\big(\alpha_n^{-1}(\mathbf{v}_i)\big) \leqslant \delta + \varepsilon/3 + 6\varepsilon_k \leqslant \varepsilon_{k+1}/4 + 6\varepsilon_k. \end{aligned} \tag{4.42}$$

From Lemma 3.2 we can select $L = L_{k+1}$ sufficiently large so that there exists a synch function $\sigma : B^{LN} \to \{0, \cdots, N-1\}$ and a set $\Phi \in \mathcal{B}_G^m$, $m = m_{k+1} = (L+1)N$, satisfying (4.1) and (4.2) for the above $\gamma_N$. Choose $K = K_{k+1}$ so large that

$$\begin{aligned} h\big((KN)^{-1}\big) &\leqslant \varepsilon_k H_k \\ m_{k+1} &\leqslant \varepsilon K_{k+1} N \end{aligned} \tag{4.43}$$

and imbed $\gamma_{NK}$ in a $(KN, \varepsilon)$-gadget $[A, \tau^{(k)}, X^{(k)}]$ to obtain a sliding-block code $\rho : A^\infty \to A$ of $[A, \tau^{(k)}, X^{(k)}]$ and hence a sliding-block code $f^{(k+1)} : A^\infty \to A$ of $[G, \mu, U]$ defined as the cascade $f^{(k+1)}(u) = \rho(\overline{f^{(k)}}(u))$ of $\overline{f^{(k)}}$ and $\rho$. Define the encoded process $[A, \tau^{(k+1)}, X^{(k+1)}]$ by $\tau^{(k+1)} = \tau^{(k)}\bar{\rho}^{-1} = \mu\overline{f^{(k+1)}}^{-1}$. The proof of Lemma 3.4 applies (with $[G, \mu, U]$ replaced by $[A, \tau^{(k)}, X^{(k)}]$), and hence

$$P_e(\tau^{(k+1)}, \nu, i, g) \leqslant 3\varepsilon \leqslant \varepsilon_{k+1},$$

and hence $[A, \tau^{(k+1)}, X^{(k+1)}]$ is $\varepsilon_{k+1}$-invulnerable, proving (4.22).

By construction and (2.7), (4.42) and (4.27),

$$\begin{aligned} |f^{(k+1)} - f^{(k)}|_\mu &= |\rho - i|_{\tau^{(k)}} = \int_{x:\, x_0 \neq \rho(x)} d\tau^{(k)}(x) \\ &\leqslant \Sigma_{i=0}^{KN-1} \int_{x:\, x_0 \neq \rho(x) x \in T^i F} d\tau^{(k)}(x) + \varepsilon \\ &= \Sigma_{i=0}^{K-1} N E_{\tau^{(k)}}\big\{ d_N\big(X_{iN}^N, \gamma_N(X_{iN}^N)\big) | x \in F\big\}\tau^{(k)}(F) + \varepsilon \leqslant 7\varepsilon_k \end{aligned}$$

which proves (4.23) and hence with Lemma A.1 of the Appendix proves the right-hand inequality of (4.24).

To prove the left-hand inequality of (4.24), we derive an upper bound for the code size $\bar{\mathcal{C}}$ and use it in conjunction with Lemma 2.2. Recall that $\bar{\mathcal{C}} \subset \mathcal{C}$ and $\mathcal{C}$ is a full $6\varepsilon_k$-separating subset of $\mathcal{G}_n \subset \mathcal{G}_n^{(1)}$. Thus, if $\mathcal{C}' \supset \mathcal{C}$ is a full $6\varepsilon_k$-separating subset of the larger set $\mathcal{G}_n^{(1)}$, we must have that

$$|\bar{\mathcal{C}}| \leqslant |\mathcal{C}| \leqslant |\mathcal{C}'|. \tag{4.44}$$

To upper bound $|\mathcal{C}'|$, note that, since $\mathcal{C}' = \{\mathbf{a}_i; i = 1, \cdots, |\mathcal{C}'|\}$ $6\varepsilon_k$-separates $\mathcal{G}_n^{(1)}$, the spheres

$$V_{3\varepsilon_k}(\mathbf{a}_i) = \left\{x^n : x^n \in \mathcal{G}_n^{(1)}, d_n(x^n, \mathbf{a}_i) \leqslant 3\varepsilon_k\right\}$$

are disjoint and hence

$$|\mathcal{C}'| \leqslant \frac{|\mathcal{G}_n^{(1)}|}{\min_i |V_{3\varepsilon_k}(\mathbf{a}_i)|}, \tag{4.45}$$

that is, $\bar{\mathcal{C}}$ can have no more elements than $\mathcal{C}$ which can have no more elements than the total number of disjoint spheres of radius $3\varepsilon_k$ that can be packed into $\mathcal{G}_n^{(1)}$ (which is larger than $\mathcal{G}_n$). Fix $i$. If $x^n = \mathbf{a}_i$, then from (4.29) at least $(1 - \varepsilon)s$ of the $t$-tuples $x_{it}^t$, $i = 0, \cdots, s - 1$, are in $\mathcal{S}_t(\tau^{(k)}, \varepsilon)$, and hence we can obtain a distinct $n$-tuple in $\mathcal{G}_n^{(1)}$ if we change any of these $t$-tuples to a different word in $\mathcal{S}_t(\tau^{(k)}, \varepsilon)$. Furthermore, we may change up to $3\varepsilon_k s$ of these $t$-tuples and still remain inside $V_{3\varepsilon_k}(x^n)$ since, if $y^n$ is so obtained, $d_n(x^n, y^n) \leqslant n^{-1}3\varepsilon_k st = 3\varepsilon_k$. Thus, for a fixed $i$ and a sufficiently large $t$

$$\begin{aligned} |V_{3\varepsilon_k}(\mathbf{a}_i)| &\geqslant \left(|\mathcal{S}_t(\tau^{(k)}, \varepsilon)| - 1\right)^{3\varepsilon_k s} \geqslant \left(2^{t(H_k - \varepsilon)} - 1\right)^{3\varepsilon_k s} \\ &\geqslant 2^{t(H_k - 2\varepsilon)3\varepsilon_k s}. \end{aligned} \tag{4.46}$$

We also have by construction that

$$\begin{aligned} |\mathcal{G}_n^{(1)}| &\leqslant |\mathcal{S}_t(\tau^{(k)}, \varepsilon)|^{s(1-\varepsilon)} |A|^{set} \leqslant 2^{t(H_k+\varepsilon)s(1-\varepsilon)+ste\log|A|} \\ &\leqslant 2^{n(H_k+\varepsilon+\varepsilon\log|A|)}, \end{aligned} \tag{4.47}$$

and hence from (4.45)–(4.47) and (4.27)

$$|\mathcal{C}'| \leqslant 2^{n(H_k+\varepsilon+\varepsilon\log|A|-3\varepsilon_k H_k+6\varepsilon\varepsilon_k} \leqslant 2^{nH_k(1-2\varepsilon_k)},$$

and hence from Lemma 2.2, (4.27), (4.43) and (4.44)

$$\begin{aligned} H\left(\mu\overline{f^{(k+1)}}^{-1}\right) &= H\left(\tau^{(k)}\bar{\rho}^{-1}\right) \leqslant (KN)^{-1}\log|\bar{\mathcal{C}}|^K + h\left((KN)^{-1}\right) \\ &\leqslant H_k(1 - 2\varepsilon_k) + \varepsilon_k H_k = H_k(1 - \varepsilon_k), \end{aligned}$$

proving the left-hand inequality of (4.24). By construction, each good $n_{k+1}$-block has at least $n_{k+1}n_k^{-1}(1 - \varepsilon_k)$ good $n_k$-tuples, and hence for any $x^M$

$$Z_M^{(k)}(x^M) \geqslant Z_M^{(k+1)}(x^M)n_k^{-1}(1 - \varepsilon_k)n_{k+1}. \tag{4.48}$$

Let $\bar{x}$ be a string produced by $\tau^{(k)}$ and let $x$ be $\bar{\rho}(\bar{x})$, the resulting $\tau^{(k+1)}$ string. If $\bar{x} \in F$, the following $K = K_{k+1}$ $N_{k+1}$-tuples of $x$ will be synchronized good $n_{k+1}$-tuples, and hence

$$M^{-1}Z_M^{(k+1)}(x^M) \geqslant M^{-1}K_{k+1}\left(\Sigma_{i=0}^{M-1}\chi_F(T^i\bar{x}) - 1\right).$$

Thus, from (4.41) with $\tau^{(k)}$ probability one,

$$\begin{aligned} \lim_{M\to\infty} M^{-1}Z_M^{(k+1)}\left(\left(\bar{\rho}(\bar{x})^M\right)\right. &\geqslant K_{k+1}\tau^{(k)}(F) \geqslant N_{k+1}^{-1} \\ &\geqslant (1 - \varepsilon_{k+1})n_{k+1}^{-1}, \end{aligned} \tag{4.49}$$

proving (4.26). This, in turn, implies that with $\tau^{(k+1)}$ probability one

$$\lim_{M\to\infty} M^{-1}Z_M^{(k)}(X^{(k+1)M}) \geqslant \lim_{M\to\infty} M^{-1}Z_M^{(k+1)}(X^{(k+1)M})n_k^{-1}(1\varepsilon_k)n_{k+1}$$
$$\geqslant n_k^{-1}(1-\varepsilon_k-\varepsilon_{k+1}),$$

proving (4.25).

PROOF OF THEOREM 3.2. To begin the induction, choose $\varepsilon$ so small that

$$\min\left(\tfrac{1}{2}, C-H^*\right) > 2(h(7\varepsilon_1)+7\varepsilon_1\log|A|)+h(\varepsilon_1)+\varepsilon_1\log|A|.$$

Let $[G,\mu,U]$ be an i.i.d. source with entropy $H(\mu)=H^*+2(h(7\varepsilon_1)+7\varepsilon_1\log|A|)+h(\varepsilon_1)+\varepsilon_1\log|A|<C$. From Corollary 3.3 and its proof there exist an encoder $f^{(1)}$ and decoder $g^{(1)}$ such that

$$P_e\left(\mu\overline{f^{(1)}}^{-1},\nu,i,g^{(1)}\right)\leqslant\varepsilon_1,$$
$$H(\mu)\geqslant H_1=H(\tau^{(1)})\geqslant H(\mu)-h(\varepsilon_1)-\varepsilon_1\log|A|$$
$$=H^*+2(h(7\varepsilon_1)+7\varepsilon_1\log|A|).$$

The good $n_1$ blocks are the block code words used to construct $f^{(1)}$ via Corollary 3.3. As in (4.49), with $\tau^{(1)}$ probability one,

$$\lim_{M\to\infty} M^{-1}Z_M^{(1)}(x)\geqslant(1-\varepsilon_1)n_1^{-1},$$

setting up an induction using Lemma 4.1. Define $\delta(\varepsilon)=h(7\varepsilon)+7\varepsilon\log|A|$ and define $\varepsilon_{k+1}$ as the solution to

$$2\delta(\varepsilon_{k+1})=H_k-H^*-\delta(\varepsilon_k). \tag{4.50}$$

This assignment ensures that at the next stage $H_{k+1}>H^*$ and, as we shall see, that $H_k\to H^*$. There will always be such a solution to (4.50) since $\delta(\varepsilon)$ is continuous and the right-hand side is in $(0,\frac{1}{2})$ for all $k\geqslant 1$. Iterate Lemma 4.1 to construct $f^{(k+1)}$ and $g^{(k+1)}$ such that

$$P_e\left(\mu\overline{f^{(k+1)}}^{-1},\nu,i,g^{(k+1)}\right)\leqslant\varepsilon_{k+1},$$
$$|f^{(n)}-f^{(n+1)}|_\mu\leqslant 7\varepsilon_k, \tag{4.51}$$
$$H_k(1-\varepsilon_k)\geqslant H_{k+1}\geqslant H_k-\delta(\varepsilon_k),$$

and so that each good $n_{k+1}$-block used to construct $f^{(k+1)}$ has at least $(1-\varepsilon_k/2)n_k^{-1}n_{k+1}$ good $n_k$ blocks.

By construction and the triangle inequality, if $m\geqslant n$,

$$|f^{(n)}-f^{(m)}|_\mu\leqslant\Sigma_{i=n}^m|f^{(i)}-f^{(i+1)}|_\mu\leqslant 7\Sigma_{i=n}^m\varepsilon_i. \tag{4.52}$$

We also have by construction and (4.51) that

$$H_n-H_m=\Sigma_{i=n}^{m-1}(H_i-H_{i+1})\geqslant\Sigma_{i=n}^{m-1}\varepsilon_iH_i\geqslant H^*\Sigma_{i=n}^{m-1}\varepsilon_i,$$

and hence since $H_1\geqslant H_n\geqslant H_m\geqslant H^*$ for all $m\geqslant n$,

$$\Sigma_{i=n}^{m-1}\varepsilon_i\leqslant(H_1/H^*)-1,$$

and therefore

$$\Sigma_{i=1}^{\infty}\varepsilon_i \leqslant (H_1/H^*) - 1, \tag{4.53}$$

so that from (4.52) and (4.53) we have

$$\lim_{m,\, n\to\infty} |f^{(n)} - f^{(m)}|_{\mu} = 0, \tag{4.54}$$

so that $f^{(n)}$ is a Cauchy sequence. From Shields (1973), page 38, there is a limit code $f^*$ such that

$$\lim_{n\to\infty} |f^{(n)} - f^*|_{\mu} = 0. \tag{4.55}$$

We also have from (4.53) that

$$\lim_{n\to\infty} \varepsilon_n = 0, \tag{4.56}$$

and hence from (4.55), Lemma A.1, and the continuity of $\delta(\varepsilon)$,

$$H(\mu\overline{f^*}^{-1}) = \lim_{k\to\infty} H_k = \lim_{k\to\infty}(H^* + \delta(\varepsilon_k) + 2\delta(\varepsilon_k)) = H^*.$$

Define $\tau^* = \mu\overline{f^*}^{-1}$. We now prove that $\tau^*$ (which has entropy $H(\tau^*) = H^*$) is indeed invulnerable. Toward this end, first consider the performance of the $k$th decoder $g^{(k)}$ on the $m$th source $\tau^{(m)}$ for $m \geqslant k$. Recall that $g^{(k)}$ performed well during good $n_{k+1}$-blocks (by definition of good $n_{k+1}$-blocks), and hence the performance of $g^{(k)}$ on $\tau^{(m)}$ can be measured by the percentage of time $\tau^{(m)}$ spends producing good $n_{k+1}$-blocks. From (4.39) we have that, for any $M$ and the $m_k = 2q_{k+1}$ length decoder $g^{(k)}$,

$$\begin{aligned} E_{\tau^{(m)}} d_1\big(X_0^{(m)}, g^{(k)}(Y_{-q}^{m_k})\big) &= M^{-1}\Sigma_{i=0}^{M-1} E_{\tau^{(m)}} d_1\big(X_i^{(m)}, g^{(k)}(Y_{-q}^{m_k})\big) \\ &\leqslant 3\varepsilon_k + 2\varepsilon_{k+1}/7 + 1 - M^{-1}n_{k+1}E_{\tau^{(m)}} Z_M^{(k+1)}(X^{(m)M}). \end{aligned} \tag{4.57}$$

From iteration on (4.48),

$$\begin{aligned} Z_M^{(k+1)}(x^M) &\geqslant Z_M^{(m)}(x^M) n_m n_{k+1}^{-1} \textstyle\prod_{i=k+1}^{m-1}(1 - \varepsilon_i) \\ &\geqslant Z_M^{(m)}(x^M) n_m n_{k+1}^{-1}\big(1 - \textstyle\sum_{i=k+1}^{m-1}\varepsilon_i\big), \end{aligned}$$

and hence

$$M^{-1}n_{k+1}E_{\tau^{(m)}} Z_M^{(k+1)}(X^{(m)M}) \geqslant M^{-1}n_m\big(1 - \textstyle\sum_{i=k+1}^{m-1}\varepsilon_i\big)E_{\tau^{(m)}} Z_M^{(m)}(X^{(m)M}).$$

From (4.49) however,

$$\begin{aligned} \lim_{M\to\infty} M^{-1}n_m\big(1 - \textstyle\sum_{i=k+1}^{m-1}\varepsilon_i\big)E_{\tau^{(m)}} Z_M^{(m)}(X^{(m)M}) \\ \geqslant (1 - \varepsilon_m)\big(1 - \textstyle\sum_{i=k+1}^{m-1}\varepsilon_i\big) \geqslant 1 - \textstyle\sum_{i=k+1}^{m}\varepsilon_i, \end{aligned}$$

and hence from (4.57)

$$E_{\tau^{(m)}} d_1\big(X_0^{(m)}, g^{(k)}(Y_{-q}^{m_k})\big) \leqslant 3\varepsilon_k + 2\varepsilon_{k+1}/7 + \textstyle\sum_{i=k+1}^{m}\varepsilon_i \leqslant 3\sum_{i=k}^{m}\varepsilon_i, \tag{4.58}$$

and hence from Corollary A.1 of the Appendix

$$E_{\tau^*} d_1\big(X_0, g^{(k)}(Y_{-q}^{m_k})\big) \leqslant 3\textstyle\sum_{i=k}^{\infty}\varepsilon_i, \tag{4.59}$$

which by (4.53) goes to zero as $k \to \infty$. In words, the $g^{(k)}$ all work fairly well for the limit process $\tau^*$. Denote the error event $\mathcal{E}_k = \{(x, y);\ g^{(k)}(y) \neq x_0\}$. From

(4.59) and (4.53),

$$\tau^*\nu(\mathcal{E}_k) \leqslant 3\Sigma_{i=k}^{\infty}\varepsilon_i \to 0 \text{ as } k \to \infty$$

and hence we can choose a subsequence $k_i$, $i = 1, 2, \cdots$ such that

$$\Sigma_{i=1}^{\infty}\tau^*\nu(\mathcal{E}_{k_i}) < \infty. \tag{4.60}$$

Define the decoder $g^* : B^{\infty} \to A$ by $g^*(y) = a$ if, for all but a finite number of $i$, $g^{(k_i)}(Y_{-q_{k_i}}^{m_{k_i}}(y)) = a$, otherwise set $g^*(y) = a^*$. In other words, $g^*$ sees if all but a finite number of good decoders yield the same symbol and, if so, prints that symbol. Define $\mathcal{E} = \{x, y : g^*(y) \neq x_0\}$, and we have for all $n$, $\mathcal{E} \subset \cup_{i=n}^{\infty}\mathcal{E}_{k_i}$, and hence from (4.60)

$$\tau^*\nu(\mathcal{E}) \leqslant \Sigma_{i=n}^{\infty}\tau^*\nu(\mathcal{E}_{k_i}) \to 0 \text{ as } n \to \infty$$

and hence

$$P_e(\mu\overline{f^*}^{-1}, \nu, i, g^*) = 0,$$

completing the proof.

PROOF OF THEOREM 3.3. Given $\nu$ as above and $H(\mu)$, there is from Corollary 3.4 an invulnerable $B$-process $[A, \tau^*, X^*]$ with decoder $g^* : B^{\infty} \to A$. Since $[A, \tau^*, H^*]$ and $[G, \mu, U]$ are $B$-processes with equal entropy, they are isomorphic, and hence there exists an infinite length sliding-block encoder $f : G^{\infty} \to A$ and decoder $\gamma : A^{\infty} \to G$ such that $\tau^* = \mu\bar{f}^{-1}$ and

$$\mu(u : \gamma(\bar{f}(u)) \neq u_0) = 0.$$

Define the decoder $g : B^{\infty} \to G$ by $g(y) = \gamma(\overline{g^*}(y))$, the cascade of $\gamma$ and $g^*$. Since $\tau^* = \mu\bar{f}^{-1}$ is invulnerable, from the union bound,

$$\begin{aligned}\tau^*\nu(x, y : \overline{g^*}(y) \neq x) &\leqslant \Sigma_{i=-\infty}^{\infty}\tau^*\nu(x, y : g^*(T^i y) \neq x_i) \\ &= \Sigma_{i=-\infty}^{\infty}\tau^*\nu(x, y : g^*(y) \neq x_0) = 0,\end{aligned}$$

and hence

$$\begin{aligned}P_e(\mu, \nu, f, g) &= \int d\mu(u)\nu_{\bar{f}(u)}(y : \gamma(\overline{g^*}(y)) \neq u_0) \\ &\leqslant \int d\mu(u)\nu_{\bar{f}(u)}(y : \overline{g^*}(y) \neq \bar{f}(u)) \\ &\quad + \int d\mu(u)\nu_{\bar{f}(u)}(y : \gamma(\bar{f}(u)) \neq u_0) = 0.\end{aligned}$$

PROOF OF COROLLARY 3.5. If a source $[A, \tau, X]$ can be communicated with zero error then it follows from the Fano inequality and data processing theorem (Shields' (1973) Lemma 8.2 and Billingsley (1965), Theorem 17.3) that $H(X|Y) = 0$ and hence

$$H(X) = I(X; Y) \leqslant \sup_{\text{block stationary } [A, \tau, X]} I(\tau\nu),$$

from Gray and Davisson (1977) page 102, the right-hand side is $C$, and hence $C_0 \leqslant C$. From Theorem 3.3, $C_0 \geqslant C$, proving the corollary.

## APPENDIX

Let $\mathcal{L}_A$ denote the class of all infinite (and hence also finite) length sliding-block codes $f: G^\infty \to A$. Given a stationary source $[G, \mu, U]$, define a metric on $\mathcal{L}_A$ by

$$|f - f'|_\mu = \int_{u:\, f(u)\neq f'(u)} d\mu(u) = \int d\mu(u) d_1(f(u), f'(u)),$$

that is, the average Hamming distance between the two differently encoded versions of $\mu$.

LEMMA A.1. (Follows from Shields' (1973) Lemma 8.2). *For $f, \varphi \in \mathcal{F}_A$,*

$$|H(\mu f^{-1}) - H(\mu \varphi^{-1})| \leqslant h(|f - \varphi|_\mu) + |f - \varphi|_\mu \log |A|.$$

LEMMA A.2. *Given a stationary channel $\nu$, a stationary source $[G, \mu, U]$, a length $m$ sliding-block decoder, and two encoders $f$, $f'$, then for any $r$,*

(A.1) $$|P_e(\mu, \nu, f, g) - P_e(\mu, \nu, f', g)| \leqslant m/r + r|f - f'|_\mu + m \max_{a^r \in A^r} \sup_{x,\, x' \in c(a^r)} \bar{d}_r(\nu_x^r, \nu_{x'}^r).$$

*Furthermore, if $i: A^\infty \to A$ is the identity sliding-block code $i(x) = i_1(x_0) = x_0$, then*

(A.2) $$|P_e(\mu f^{-1}, \nu, i, g) - P_e(\mu (f')^{-1}, \nu, i, g)| \leqslant m/r + r|f - f'|_\mu + m \max_{a^r \in A^r} \sup_{x,\, x' \in c(a^r)} \bar{d}_r(\nu_x^r, \nu_{x'}^r).$$

PROOF. Define $\Lambda = \{u : f(u) = f'(u)\}$ and $\Lambda_r = \{u : f(T^i u) = f'(T^i u);\ i = 0, \cdots, r-1\} = \cap_{i=0}^{r-1} T^i \Lambda$. From the union bound,

(A.3) $$\mu(\Lambda_r^c) \leqslant r\mu(\Lambda^c) = r|f - f'|_\mu.$$

We have from stationarity that if $g = g_m(Y_{-q}^m)$ then

(A.4)
$$\begin{aligned} P_e(\mu, \nu, f, g) &= \int d\mu(u) \nu_{f(u)}(y : g_m(y_{-q}^m) \neq u_0) \\ &= r^{-1} \Sigma_{i=0}^{r-1} \int d\mu(u) \nu_{f(u)}(y : g_m(y_{i-q}^m) \neq u_i) \\ &\leqslant m/r + r^{-1} \Sigma_{i=q}^{r-q} \int_{\Lambda_r} d\mu(u) \nu_{f(u)}^r(y^r : g_m(y_{i-q}^m) \neq u_i) + \mu(\Lambda_r^c). \end{aligned}$$

Fix $u \in \Lambda_r$ and let $p_u$ yield $\bar{d}_r(\nu_{f(u)}^r, \nu_{f'(u)}^r)$, that is, $\Sigma_{y'^r} p_u(y^r, y'^r) = \nu_{f(u)}^r(y^r)$, $\Sigma_{y^r} p_u(y^r, y'^r) = \nu_{f'(u)}^r(y'^r)$, and

(A.5) $$r^{-1} \Sigma_{i=0}^{r-1} p_u(y^r, y'^r : y_i \neq y_i') = \bar{d}_r(\nu_{f(u)}^r, \nu_{f'(u)}^r).$$

We have that

$$\begin{aligned} r^{-1} \Sigma_{i=q}^{r-q} \nu_{f(u)}^r(y'^r : g_m(y'^m_{i-q}) \neq u_i) &= r^{-1} \Sigma_{i=q}^{r-q} p_u(y^r, y'^r : g_m(y'^m_{i-q}) \neq u_i) \\ &\leqslant r^{-1} \Sigma_{i=q}^{r-q} p_u(y^r, y'^r : g_m(y'^m_{i-q}) \neq g_m(y_{i-q}^m)) \\ &\quad + r^{-1} \Sigma_{i=q}^{r-q} p_u(y^r, y'^r : g_m(y_{i-q}^m) \neq u_i) \\ &\leqslant r^{-1} \Sigma_{i=q}^{r-q} p_u(y^r, y'^r : y_{i-q}^m \neq y'^m_{i-q}) + P_e(\mu, \nu, f', g) \\ &\leqslant r^{-1} \Sigma_{i=q}^{r-q} \Sigma_{j=i-q}^{i-q+m} p_u(y^r, y'^r : y_j \neq y_j') + P_e(\mu, \nu, f', g) \\ &\leqslant m \bar{d}_r(\nu_{f(u)}^r, \nu_{f'(u)}^r) + P_e(\mu, \nu, f', g), \end{aligned}$$

which with (A.3)-(A.5) completes the proof of (A.1). Equation (A.2) follows almost identically:

$$\begin{aligned} P_e(\mu, \bar{f}^{-1}, \nu, i, g) &= \int d\mu(u)\nu_{\bar{f}(u)}(y : g_m(y_{-q}^m) \neq f(u)) \\ &\leqslant m/r + \mu(\Lambda_r^C) + r^{-1}\textstyle\sum_{i=q}^{r-q}\int_{\Lambda_r} d\mu(u)\nu_{\bar{f}(u)}^r(y^r : g_m(y_{i-q}^r) \neq f(T^i u)) \\ &\leqslant m/r + r|f - f'|_\mu + r^{-1}\textstyle\sum_{i=q}^{r-q} p_u(y^r, y'^r : g_m(y'^m_{i-q}) \neq g_m(y_{i-q}^m)) \\ &\quad + r^{-1}\textstyle\sum_{i=q}^{r-q} p_u(y^r, y'^r : g_m(y_{i-q}^m) \neq f(T^i u)) \\ &\leqslant m/r + r|f - f'|_\mu + m\bar{d}_r(\nu_{\bar{f}(u)}^r, \nu_{\bar{f}'(u)}^r) + P_e(\mu(\bar{f}')^{-1}, \nu, i, g). \end{aligned}$$

COROLLARY A.1. *Given a stationary $\bar{d}$-continuous channel $\nu$ and a finite decoder $g_m : B^m \to A$, then, if $f^{(n)} \in \mathcal{L}_A, f \in \mathcal{L}_A$ are such that*

$$\lim_{n\to\infty}|f^{(n)} - f|_\mu = 0,$$

*then*

$$\lim_{n\to\infty} P_e(\mu, \nu, f^{(n)}, g_m) = P_e(\mu, \nu, f, g_m)$$

*and*

$$\lim_{n\to\infty} P_e(\mu\overline{f^{(n)}}^{-1}, \nu, i, g_m) = P_e(\mu\bar{f}^{-1}, \nu, i, g_m)$$

*where $i : A^\infty \to A$ is the identity sliding-block encoder $i(x) = x_0$.*

Both results state that probability of error over a $\bar{d}$-continuous channel is a continuous function of the encoder as measured by the coder metric. In one case the original source is reproduced in the other the channel input is reproduced.

PROOF. Fix $\varepsilon > 0$ and choose $r$ so large that

$$\max_{a^r} \sup_{x,\, x' \in c(a^r)} \bar{d}_r(\nu_x^r, \nu_{x'}^r) \leqslant \varepsilon/(3m)$$

$$m/r \leqslant \varepsilon/3$$

and $n_0$ so great that $n \geqslant n_0$ implies

$$|f^{(n)} - f|_\mu \leqslant \varepsilon/(3r).$$

From Lemma A.1,

$$|P_e(\mu, \nu, f^{(n)}, g) - P_e(\mu, \nu, f, g)| \leqslant \varepsilon,$$

$$|P_e(\mu\overline{f^{(n)}}^{-1}, \nu, i, g) - P_e(\mu\bar{f}^{-1}, \nu, i, g)| \leqslant \varepsilon,$$

completing the proof.

## REFERENCES


ADLER, R. L. (1961). Ergodic and mixing properties of infinite memory channels. *Proc. Amer. Math. Soc.* **12** 924–930.

AHLSWEDE, R., and WOLFOWITZ, J. (1971). Channels without synchronization. *Adv. in Appl. Probability* **3** 383–403.

BELLMAN, R. (1960). *Introduction to Matrix Analysis*. Chapter 12. McGraw-Hill, New York.

BERGER, T., and LAU, J. K. (1977). On binary sliding-block codes. *IEEE Trans. Information Theory* **23** 343–353.

BILLINGSLEY, P. (1965). *Ergodic Theory and Information*. Wiley, New York.

BLACKWELL, D. (1959). Infinite codes for memoryless channels. *Ann. Math. Statist.* **30** 1242–1244.

COVER, T. M., MCELIECE, R. J. and POSNER, E. C. (1980). Asynchronous multiple access channel capacity. *IEEE Trans. Information Theory* (To appear.)

DOBRUSHIN, R. L. (1967). Shannon's theorems for channels with synchronization errors. *Problemy Peredači Informatsii*. **3** 18–36.

GRAY, R. M. (1975). Sliding-block source coding. *IEEE Trans. Information Theory* **IT-21** 357–368.

GRAY, R. M. and DAVISSON, L. D., editors (1977). *Ergodic and Information Theory*. Dowden, Hutchinson, and Ross, Stroudsburg, Pennsylvania.

GRAY, R. M. and ORNSTEIN, D. S. (1976). Sliding-block joint source/noisy-channel coding theorems. *IEEE Trans. Information Theory* **IT-22** 682–690.

GRAY, R. M. and ORNSTEIN, D. S. (1979). Block coding for discrete stationary $\bar{d}$-continuous channels. *IEEE Trans. Information Theory* **IT-25** 292–306.

JACOBS, K. (1959). Die übertragung diskreter informationen durch periodishe und fastperiodische kaväle. *Math. Ann.* **137** 125–135.

JACOBS, K. (1962). Über die struktur der mittleren entropie. *Math. Z.* **78** 33–43.

KADOTA, T. T. and WYNER, A. D. (1972). Coding theorems for stationary asymptotically memoryless, continuous-time channels. *Ann. Math. Statist.* **43** 1603–1611.

KIEFFER, J. C. (1977). On sliding-block coding for transmission of a source over a stationary nonanticipatory channel. *Information and Control*. **35** 1–19.

NEDOMA, J. (1957). The capacity of a discrete channel. *Trans. First Prague Conf. Information Theory* 143–181 Prague.

NEDOMA, J. (1963). Über die ergodizität und r-ergodizität stationärer wahrscheinlichkeitsmasse. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*. **2** 90–97.

NEDOMA, J. (1964). The synchronization for ergodic channels. *Trans. Third Prague Conf. Information Theory., etc.*. 529–539, Prague.

ORNSTEIN, D. W. (1973). An application of ergodic theory to probability theory. *Ann. Probability* **1** 43–58.

ORNSTEIN, D. S. (1974). *Ergodic Theory, Randomness, and Dynamical Systems*. Yale University Press, New Haven.

PFAFFELHUBER, E. (1971). Channels with asymptotically decreasing memory and anticipation. *IEEE Trans. Information Theory* **IT-17** 379–385.

SCHOLTZ, R. A. (1966). Codes with synchronization capability. *IEEE Trans. Information Theory* **IT-12** 135–140.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656.

SHANNON, C. E. (1956). The zero error capacity of a noisy channel. *IRE Trans. Information Theory* **IT-2** 8–19.

SHANNON, C. E. (1957). Certain results in coding theory for noisy channels. *Information and Control*. **1** 6–25.

SHIELDS, P. C. (1973). *The Theory of Bernoulli Shifts*. Univ. Chicago Press.

STIFFLER, J. J. (1971). *Theory of Synchronous Communication*. Prentice-Hall, Englewood Cliffs, New Jersey.

VAJDA, I. (1965). A synchronization method for totally ergodic channels. *Trans. of the Fourth Prague Conf. Information Theory* 611–625, Prague.

WOLFOWITZ, J. (1964). *Coding Theorems of Information Theory*. Second ed. Springer, Berlin.

R. M. GRAY
DEPARTMENT OF ELECTRICAL ENGINEERING
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

R. L. DOBRUSHIN
INSTITUTE FOR PROBLEMS OF
INFORMATION TRANSMISSION
U.S.S.R. ACADEMY OF SCIENCES
19 ERMOLOVA ST.
MOSCOW K-51
U.S.S.R. 103051

D. S. ORNSTEIN
MATHEMATICS DEPARTMENT
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305