Identification Without Randomization
Rudolf Ahlswede and Ning Cai

Abstract—In the theory of identification via noisy channels randomization in the encoding has a dramatic effect on the optimal code size, namely, it grows double-exponentially in the blocklength, whereas in the theory of transmission it has the familiar exponential growth.

We consider now instead of the discrete memoryless channel (DMC) more robust channels such as the familiar compound (C) and arbitrarily varying channels (AVC). They can be viewed as models for jamming situations. We make the pessimistic assumption that the jammer knows the input sequence before he acts. This forces communicators to use the maximal error concept (see [1]) and also makes randomization in the encoding superfluous. Now, for a DMC W by a simple observation, made in [2], in the absence of randomization the identification capacity, say C_{NIR}(W), equals the logarithm of the number of different row-vectors in W. We generalize this to compound channels.

A formidable problem arises if the DMC W is replaced by the AVC W. In fact, for 0-1 matrices only in W we are—exactly as for transmission—led to the equivalent zero-error capacity of Shannon (see [3]). But for general W the identification capacity C_{NIR}(W) is quite different from the transmission capacity C(W). An observation is that the separation codes of [1] are also relevant here. We present a lower bound on C_{NIR}(W). It implies for instance

\[ W = \left\{ \frac{1}{2}\delta \right\} \]

that C_{NIR}(W) = 1, which is obviously tight. It exceeds C(W), which is known [1]) to exceed 1 - h(\delta), where h is the binary entropy function.

We observe that a separation code with worst case average list size L (which we call an NRA-code) can be partitioned into \( 2^L \) transmission codes. This gives a meaningful characterization of the capacity of AVC with maximal probability of error in terms of the capacity of codes with list decoding.

We also prove that randomization in the decoding does not increase C(C(W)) and C_{NIR}(W).

Finally, we draw attention to related work on source coding [4], [5].

Index Terms—Arbitrarily varying channels, identification, list code capacity, separation codes.

I. INTRODUCTION AND RESULTS

Let \( \mathcal{X}, \mathcal{Y} \) be the finite input and output alphabets of the channel considered, namely, the discrete memoryless (DMC) W, the arbitrarily varying channel (AVC) W specified by a set of \( |\mathcal{X}| \times |\mathcal{Y}| \)-stochastic matrices and also written in the form

\[ W = \{ \cdot, s \} \in \mathcal{S} \text{ finite}, \]

and the compound channel (CC) C(W) associated with W.

We study here primarily identification codes without randomization (NRI-codes) for W. An \((n, M, \lambda_1, \lambda_2)\) NRI-code for W is a system of pairs \((u, D_u)\) \( u \in \mathcal{U} \) such that \( \mathcal{U} \subseteq \mathcal{X}^n \), \( D_u \subseteq \mathcal{Y}^n \) (for \( u \in \mathcal{U} \)), \( |\mathcal{U}| = M \), and for all \( u, u' \in \mathcal{U} \) (\( u \neq u' \)), \( s^u \in S^n \)

\[ W^n(D_u[u, s^u]) > 1 - \lambda_1 \]

(1.1)

and

\[ W^n(D_u[u', s^u]) < \lambda_2 \]

(1.2)

Manuscript received August 1, 1998; revised May 1, 1999.

1 The authors are with Fachbereich Mathematik, Universität Bielefeld, 33501 Bielefeld, Germany.

Communicated by I. Csiszar, Associate Editor for Shannon Theory.

Publisher Item Identifier S 0018-9448/99$10.00 © 1999 IEEE
Here for \( s^n = (s_1, \ldots, s_n) \), \( y^n = (y_1, \ldots, y_n) \), and \( u = (u_1, \ldots, u_n) \):

\[
W^n(y|u, s^n) = \prod_{i=1}^{n} W(y_i|u_i, s_i).
\]

(Recall that in the definition of ID-codes in [2] instead of \( \mathcal{U} \subset X^n \) we used more generally \( \mathcal{U} \subset \mathcal{P}(X^n) \), the set of all probability distributions (PD) on \( X^n \).

We also point out that already in [1] it had been shown that for the DMC \( W \) with distinct row vectors the capacity of NRI-codes is \( \log |X| \) even before the concept of introduction was introduced in [2].

A related concept, used already in [1], are \( (n, M, \lambda) \)-nonrandom separation codes (SP-codes), which we abbreviate as NRS-codes. They are defined as a system of quadruples

\[
\{ (u, u', D(u, u'), D(u', u')) : u, u' \in \mathcal{U}, u \neq u' \}
\]

where \( \mathcal{U} \subset X^n \), \( \mathcal{U} = M, D(u, u') \subset Y^n \),

\[
D(u, u') \cap D(u', u) = \emptyset (u \neq u'), \tag{1.3}
\]

and

\[
W^n(D(u, u')|u, s^n) \geq 1 - \lambda(s^n \in S^m). \tag{1.4}
\]

Notice that with the choice

\[
D(u, u') = D_u \backslash D_{u'} \tag{1.5}
\]

we can associate with every \( (n, M, \lambda_1, \lambda_2) \) ID-code (respectively, NRI-code) an \( (n, M, \lambda) \) SP-code (respectively, NRS-code), where, by (1.1) and (1.3) \( \lambda = \lambda_1 + \lambda_2 \). This fact has been used in [2] in the proof of the (so)-converse (an exponential weak converse in the sense of [4]), because for the DMC (in case of randomization and no randomization as well) both code concepts lead to the same capacities.

Next we present a third kind of codes, called NRA-codes, which were discovered in [1]. Their properties are stronger than those of NRS-codes, but weaker than those of NRI-codes.

These codes can be viewed as list codes with an additional separation property (like (1.3) and (1.4)). They are essential for our analysis and described below.

Analogously, we speak of \( A \), if in the definition \( \mathcal{U} \subset \mathcal{P}(X^n) \). For the CCC only sequences \( s^n = (s_1, \ldots, s_n) \in S^n \) are considered and the code constants are modified accordingly.

For a system \( \{ (u, D_u) : u \in \mathcal{U} \} \subset X^n \) satisfying (1.1) we define the worst case average list size

\[
\overline{L}_U = \max_{u \in \mathcal{U}, \nu \in \mathcal{D}_U} \overline{L}(u, s^n) \tag{1.6}
\]

where

\[
\overline{L}(u, s^n) = \sum_{y \in D_u} L(y^n)W^n(y|u, s^n) \tag{1.7}
\]

and

\[
L(y^n) = \sum_{y' \in \mathcal{U}, u' \in \mathcal{D}_{u'}} \delta_{uu'} \beta_{u'y^n} \tag{1.8}
\]

Now we say that \( \{(u, D_u) : u \in \mathcal{U}\} \) is an \( (n, M, \lambda_1, \lambda, \mathcal{U}) \) NRA-code, if

\[
\overline{L}_U \leq \mathcal{L} \tag{1.9}
\]

and for all \( u, u' \in \mathcal{U}, u \neq u' \), there is a partition of \( D_u \bigcup D_{u'} \), say \( \{A(u, u'), A(u', u')\} \), such that

\[
W^n(A(u, u')|u, s^n) < \lambda, \quad \text{for all } s^n. \tag{1.2'}
\]

Obviously, (1.2') holds for any partition of \( D_u \bigcup D_{u'} \), \( \lambda = \lambda_2 \), if (1.2) is true and for \( D(u, u') = \mathcal{D}_u \bigcup \mathcal{D}_{u'} \), \( \lambda = \lambda_1 + \lambda_2 \), (1.4) holds whenever (1.2') holds. On the other hand, for \( A(u', u') = D_u \bigcup D_{u'} \bigcup D(u', u) \), (1.4) implies (1.2').

A. Partitioning NRA—Codes into (Nonrandom) Transmission Codes

We start now with a first basic result.

**Theorem 1:** Consider an \( (n, M, \lambda_1, \lambda_2, \mathcal{U}) \) NRA-code for the AVC \( W \) defined above. For every \( \epsilon > 0, \delta < \epsilon \), there exists a \( \lambda' > 0 \) such that for all \( \lambda \leq \lambda' \) and sufficiently large \( n \) the NRA-code can be partitioned into \( K \) transmission subcodes for \( W \) with maximal error probability of error \( \lambda \), if for \( \ell = (1/n) \log K \)

\[
\frac{1}{n} \log K > \ell + \epsilon.
\]

Moreover, clearly this partition contains a subcode of size at least \( K/M \).

B. A Formula for \( C(W) \)

From Theorem 1 we get a nonsingle letter characterization for \( C(W) \) involving NRS-codes for the AVC. Those codes were known (also for the DMC) already in [1, Secs. IV, V, and Lemma 3, respectively]. So they were known much earlier than ID-codes and NRI-codes (see [2]). An elegant description was used in [6]. Namely, for an integer \( m \), we associate with \( W \) a graph \( G_m(W) = \{ x^n, y^n \} \) such that \( (x^n, x'^n) \in E_m \) iff there are PD's \( \pi, \pi' \in \mathcal{P}(S^m) \) such that

\[
\sum_{x^n} \pi(x^n)W^n(x^n, y^n) = \sum_{x^n} \pi(x^n)W^n(x^n, y'^n) \tag{1.10}
\]

(or

\[
\bigcap \{ W_m(x^n, y^n) : (x^n, y^n) \in E_m \} = \emptyset.
\]

Denote by \( E_m \) the family of independent sets of the graph. Then \( \mathcal{U} \subset E_m \) is an NRS-code and we have the following auxiliary result.

**Lemma 8:** For any \( \epsilon > 0, \lambda > 0 \), and sufficiently large \( n \), one can choose \( \{ D(u^n, u'^n) : u^n \neq u'^n, u^n \in \mathcal{U} \} \) suitable to obtain an SP-code with probability of error \( \lambda \), if the pairwise Hamming distances (with respect to alphabet \( \mathcal{U} \)) of codewords in \( \mathcal{U} \subset \mathcal{U}^n \) are not smaller than \( \epsilon \).

For a list code \( \{ (u, D_u)_{u \in \mathcal{U}} \} \) satisfying (1.1) we consider the worst case average list size \( \overline{L}_U(\mathcal{D}_u|u \in \mathcal{U}) = \overline{L}_U \) (defined in (1.6)) and define

\[
T_{U, \lambda_1} = \min_{u \in \mathcal{U}} \overline{L}_U(\mathcal{D}_u|u \in \mathcal{U}) \tag{1.11}
\]

In other terms clearly,

\[
T_{U, \lambda_1} = \min_{u \in \mathcal{U}} \max_{u' \neq u} \sum_{s^n \in \mathcal{D}_u \bigcup D_{u'}} W^n(A(u, u')|u, s^n) + 1. \tag{1.12}
\]

\[
\overline{L}_U = \max_{u \in \mathcal{U}} \min_{v \in \mathcal{D}_U} \overline{L}(u, s^n) \tag{1.13}
\]

C. On Randomization in the Decoding

We mention here the effects of randomization in the decoding on the transmission capacity \( C(W) \) and the identification capacities \( C_{\text{Ident}}(W) \) and \( C_{\text{Id}}(W) \) for the AVC \( W \), that is, if the maximal probability of error criterion is used.

**Theorem 3:** For every AVC \( W \) under the maximal error probability criterion randomization in the decoding does not lead to higher capacities than i) \( C(W) \), ii) \( C_{\text{Ident}}(W) \), and iii) \( C_{\text{Id}}(W) \), respectively.
Remark 1: It follows immediately from the elimination technique and the positivity characterization of [7] that also for the average error probability criterion the transmission capacity does not increase under randomized decoding.

D. A Lower Bound on $C_{\text{NH}}(W)$

Now we present a partial result for $C_{\text{NH}}(W)$, the quantity of our main interest in this correspondence.

Theorem 4: For $P \in \mathcal{P}(\mathcal{X})$ set

$$Q(P,W) = \{(X, X', Y) : P_{X|Y} P_{X'|Y} \in \overline{W},$$

$$P_X = P_{X'} = P \text{ and } X, X', Y \text{ form a Markov chain in this order}\}$$

and set

$$\hat{I}(P,W) = \min_{(X, X', Y) \in Q(P,W)} I(X' \land XY)$$

(where $\overline{W}$ is the row-convex hull of $W$) then

$$C_{\text{NH}}(W) \geq \max_P \hat{I}(P,W).$$

(1.14)

Remark 2: $\hat{I}(P,W) = H(P)$ for $P, W$ such that $(X, X', Y) \in Q(P,W)$ implies $H(X|X') = 0$.

Corollary: The quantities in the inequality

$$C_{\text{NH}}(W) \geq C(W)$$

(1.15)

can be different.

This follows from

Example 1: For

$$W = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & \delta \\ \delta & 1 - \delta \end{pmatrix} \right\}, \quad \delta \in \left(0, \frac{1}{2}\right)$$

Theorem 4 (or also the Lemma below) yields

$$C_{\text{NH}}(W) = 1$$

(1.16)

and

$$1 - h(\delta) < C(W) < 1 \quad \text{(by [1]).}$$

For the following class, including Example 1, (1.15) also follows from Theorem 4.

Example 2: Let

$$W = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 - q(s) \end{pmatrix} : s \in S \right\}$$

where $1 > q(s) > 0$ for all $s \in \mathcal{S}$ (finite). Then for $P \in \mathcal{P}(\mathcal{X})$ of the form $P = (p, 1 - p), p \in (0, 1)$, $Q(P,W) = \emptyset$, and by (1.13) we get $C_{\text{NH}}(W) \geq 1$. This is obviously tight.

Next we give a formula for the capacities of a special class of channels, including Examples 1 and 2.

Lemma: Let $X = \{1, 2, \ldots, \alpha\}, \mathcal{Y} = \{0, 1, \ldots, \beta\}, |S| < \infty$, and $\max_{x \in \mathcal{X}} \max_{s \in S} W(0|x,s) < 1$. Furthermore, consider for input alphabet $\mathcal{X} \cup \{0\}$ and output alphabet $\mathcal{Y}$ the AVC

$$W^* = \left\{W^*(y|x, s) = W(y|x, s) \right\}$$

for all $x \in X, y \in \mathcal{Y}, s \in S$, and $W^*(0|0, s) = 1$ for all $s \in S$. Then

$$C_{\text{NH}}(W^*) = \max_{p \in \mathcal{P}(\mathcal{X})} \left[ h(p) + p C_{\text{NH}}(W) \right].$$

E. Combinatorial Problem Related to $C_{\text{NH}}(\{(\begin{pmatrix} 0 & 0 \\ 1 & \delta \end{pmatrix}, \begin{pmatrix} 1 - \delta & \delta \end{pmatrix}, s))\}$

Finding $C_{\text{NH}}(W)$ for the special case

$$W = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 - \delta & \delta \\ \delta & 1 - \delta \end{pmatrix} \right\}, \quad \delta \in \left(0, \frac{1}{2}\right)$$

is already a formidable task.

By Theorem 4 and numerical computations of B. Balkenholl

$$C_{\text{NH}}(W) \geq C(W) = 1 - h(\delta)$$

where the identity is a very special case of the capacity theorem of [1]. The heart of the matter seems to be related to the following coding problem.

We denote by $B(w^a, d) \subset \{0, 1\}^n$ the Hamming ball with radius $d$. For numbers $1 < \beta < \delta < \frac{1}{2}$ and $\lambda \in (0, 1)$ find a subset $A \subset \{0, 1\}^n$ as large as possible such that for all $w^a \in A$

$$|B(w^a, n\delta) \cap \bigcup_{y \in A} B(y^a, n\lambda)| < |B(w^a, n\delta)|.$$

F. The Capacity of the Compound Channel (CC) for NLRI-Codes

Each member $V \{1, s\}$ in the compound channel with $|S| < \infty$ introduces a partition $\{X(1)s, \ldots, X(j)s\}$ of $X$ such that $x, x'$ are in the same subset exactly if $V(x, s) = V(x', s)$. Thus an RV $X$ taking values in $X$ induces an RV $\hat{X}(s)$ for every $s \in S$ such that $\hat{X}(s) = \hat{X}$ exactly if $X \in X(s) [\mathcal{L}]$.

Theorem 5: For a CC $V \{1, s\}, s \in S$ with $|S| < \infty$

$$C_{\text{NH}}(V) = \max_{s \in \mathcal{S}} H(\hat{X}(s)).$$

II. PROOF OF THEOREM 1

We color the $M$ codewords in $\mathcal{U}$ of an $(n, M, \lambda_1, \lambda_2, L)$ NAND-code $\{u, \mathcal{D}_u\} : u \in \mathcal{U}$ randomly and independently according to the uniform distribution with $K$ colors and show that the probability for the existence of a coloring satisfying the conditions in the theorem is positive. To estimate the probability, we first fix $s^b \in S^n$ and $u \in \mathcal{U}$ and partition $\mathcal{U}(\{i\})$ into two parts $\mathcal{U}(\{i\}) (i = 1, 2)$ such that $w' \in \mathcal{U}(\{1\})$ if

$$W^n(A(u^a, u)|u^a, s^b) \leq \frac{1}{n^2}.$$

Then

$$\sum_{w' \in \mathcal{U}(\{2\})} W^n(A(u^a, u)|u^a, s^b) \leq \sum_{w' \in \mathcal{U}(\{2\})} W^n(\mathcal{D}_u \cap \mathcal{D}_{w'}|u^a, s^b) \leq \sum_{w' \in \mathcal{U}(\{2\})} W^n(\mathcal{D}_u \cap \mathcal{D}_{w'}|u^a, s^b) \leq n^2(\mathcal{U}(u, s^b) - 1) \leq n^2 L$$

(2.2)

where for the equality we use the useful observation

$$\sum_{w' \in \mathcal{U}(\{1\})} W^n(\mathcal{D}_u \cap \mathcal{D}_{w'}|u^a, s^b) = \mathcal{I}(u, s^b) - 1$$

(2.3)

for all $u \in \mathcal{U}$ and $s^b \in S^n$. For the fixed $u$, $s^b$, and $i = 1, 2$ let

$$Z_w^{(i)} = \begin{cases} W^n(A(u^a, u)|u^a, s^b), & \text{if } w' \in \mathcal{U}(\{i\}) \text{ and } u^a \text{ is colored by the same color as } u \\ 0, & \text{otherwise} \end{cases}$$

(2.4)

If we can show that for all such $u$ and $s^b$, $i = 1, 2$

$$\Pr \left\{ \sum_{w \in \mathcal{U}(\{i\})} Z_w^{(i)} > \frac{\lambda - \frac{1}{2}}{2} \right\} \leq (|S^n| |\mathcal{U}|)^{-1} \frac{\theta}{2}$$

(2.5)
for a $\theta \in (0, 1)$, then we can find a (coloring) partition

$$\{u_k: k = 1, 2, \ldots, K\}$$

such that for all $u \in U_k$ ($k = 1, 2, \ldots, K$), $s^n \in S^n$

$$\sum_{u' \in U_k \setminus \{u\}} W^n(A(u', u) | u, s^n) < \lambda - \lambda_1$$

and so

$$W^n(D_u \setminus \left( \bigcup_{u' \in U_k \setminus \{u\}} A(u', u) \right) | u, s^n) 
\geq W^n(D_u | u, s^n) - \sum_{u' \in U_k \setminus \{u\}} W^n(A(u', u) | u, s^n)$$

$$> 1 - \lambda - (\lambda - \lambda_1) = 1 - \lambda.$$ 

Thus if we let

$$D'_u = D_u \setminus \left( \bigcup_{u' \in U_k \setminus \{u\}} A(u', u) \right)$$

for all $u \in U_k$ ($k = 1, 2, \ldots, K$) then $(u, D'_u): u \in U_k$ ($k = 1, 2, \ldots, K$) is the family of codes required by the theorem. We first show (2.5) for $i = 1$. By the definition of the $Z_i^{(1)}$'s

$$\Pr\left\{ \sum_{u' \in U_k} Z_i^{(1)} > \frac{\lambda - \lambda_1}{2} \right\} \leq \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} \right\} \prod_{u' \in U_k} \exp\left\{ n^{3/2} Z_i^{(1)} \right\}$$

$$= \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} \right\} \prod_{u' \in U_k \setminus \{u\}} \left( 1 + \frac{1}{K} \right) \exp\left\{ n^{3/2} W^n(A(u', u) | u, s^n) \right\}$$

$$= \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} \right\} \prod_{u' \in U_k \setminus \{u\}} \left[ 1 + \frac{1}{K} \exp\left\{ n^{3/2} W^n(A(u', u) | u, s^n) \right\} \right] - 1$$

$$\leq \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} \right\} \prod_{u' \in U_k \setminus \{u\}} \left[ 1 + \frac{K}{n^{3/2} W^n(A(u', u) | u, s^n)} \right]$$

$$\leq \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} \right\} \prod_{u' \in U_k \setminus \{u\}} \left[ 1 + \frac{c}{K n^{3/2} W^n(D_u | u, s^n)} \right]$$

$$= \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} + \frac{c}{K} n^{3/2} \right\} \prod_{u' \in U_k \setminus \{u\}} W^n(D_u | u, s^n)$$

$$= \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} + \frac{c}{K} \sum_{u' \in U_k \setminus \{u\}} W^n(D_u | u, s^n) \right\}$$

$$\leq \exp\left\{ -n^{3/2} \frac{\lambda - \lambda_1}{2} + \frac{c}{K} \sum_{u' \in U_k \setminus \{u\}} W^n(D_u | u, s^n) \right\}$$

Thus for any $\lambda - \lambda_1$, sufficiently large $n$, and fixed $u$ and $s^n$, (2.5) holds for $i = 1$ if we choose a $K$ satisfying

$$K \geq \frac{4cL}{\lambda - \lambda_1}. \tag{2.7}$$

To show (2.5) for $i = 2$, it is sufficient to show that the probability of the event, that the number of $u' \in U_i$ (with the same color in the (random) coloring as $u$ is larger than $[\frac{\lambda - \lambda_1}{2}] = \kappa_2$, say, is not larger than the right-hand side (RHS) of (2.5).

That is,

$$\sum_{j > \kappa_2} \left( \frac{K^{(2)}}{j} \right) \left( 1 - \frac{1}{K} \right)^{j-1} \leq \left[ \frac{1}{\theta} \right] \left( \frac{\kappa_2}{K} \right)^{\kappa_2 - 1} \frac{1}{\kappa_2} \tag{2.8}$$

Indeed, if

$$\frac{K^{(2)}}{K} < \kappa_2 \tag{2.9}$$

then by Stirling’s formula and with $\kappa_2 = \frac{1}{\theta} (\lambda - \lambda_1/2)$

$$\text{LHS of (2.8)} \leq \frac{K^{(2)}}{\sqrt{2\pi}} \exp\left\{ \frac{1}{2} \left( \frac{K^{(2)}}{\kappa_2} \right) - \frac{1}{\kappa_2} \left( \frac{K^{(2)}}{\kappa_2} \right)^{\kappa_2 - 1} \right\}$$

$$\leq \frac{K^{(2)}}{\sqrt{2\pi}} \exp\left\{ \frac{1}{2} \left( \frac{K^{(2)}}{\kappa_2} \right) \right\} \leq \frac{K^{(2)}}{\sqrt{2\pi}} \exp\left\{ \frac{1}{2} \left( \frac{K^{(2)}}{\kappa_2} \right)^{\kappa_2 - 1} \right\} \tag{2.10}$$

Thus (2.8) holds if we choose

$$K > \frac{4cL}{\lambda - \lambda_1} n^{3/2} \exp\left\{ \frac{1}{\theta} \right\}$$

and sufficiently small $\lambda^*$ (and, therefore, $\lambda_2$) depending on $\lambda$, $\lambda_1$, and $\varepsilon$.

This completes the proof.

III. PROOF OF THEOREM 2

The converse part is absolutely trivial because an $(n, M)$-code $\{(u, D_u): u \in U\}$ with maximal probability of error $\lambda$ satisfies

$$\lambda_1 (\lambda_1 - 1) = 1 \quad \text{and} \quad U \in T_m.$$
To apply Theorem 1, we have to estimate $L_I$. Let us write $D_{u_1} \times \cdots \times D_{u_t} = D_{u^t}$ and denote the Hamming ball with center $u^t$ and radius $r$ in $U^t$ by $B(u^t, r)$. Then for $L_1$ (in (1.8))

$$u^t = (u_1, \ldots, u_t) \in U^t$$

$$s^m = (s_1^m, \ldots, s_t^m) \in S^m$$

$$v^t = (v_1, \ldots, v_t) \in D_{u^t}$$

$$\sum_{v^t \in B(u^t, (1 + \varepsilon)^r)} L(v^t)W(m)(v^t | u^t, s^m) \leq \sum_{J \subseteq [r]} |U^J| |I|^{(1 + \varepsilon)^r} \prod_{J \subseteq [r]} \sum_{j \in J} |L(j)| W(m)(v_j | u_j, s^m_j)
$$

Thus one can find a $\beta(\lambda_1, \varepsilon)$ such that $\beta(\lambda_1, \varepsilon) \rightarrow 0$ as $\lambda_1, \varepsilon \rightarrow 0$, and

$$\frac{1}{\varepsilon} \log L_I \leq \frac{1}{\varepsilon} \log L_D + \beta(\lambda_1, \varepsilon). \quad (3.2)$$

Finally, we choose $n = m/f$ and apply Theorem 1 to

$$\{ (u^t, D_{u^t}); v^t \}$$

to obtain a (transmission) subcode with probability of error $\varepsilon + \lambda_1$ and rate arbitrarily close to $(1/n) \log |I|/|U^t|$ when $t$ is arbitrarily large (depending on $m$ and $\lambda_1$ and $\varepsilon$ in (3.2)) are arbitrarily small. Since $m$ is fixed when $m = m/r$, $r < m$, we asymptotically lose nothing, if we add $r$ dummy letters. This completes our proof.

IV. PROOF OF THEOREM 3

i) This is an exercise in [9, p. 226, Problem 11(c)] and a very easy consequence of Theorem 1 as well.

The proofs of ii) and iii) are essentially the same and so we only prove ii).

We are given a system $(U, Q)$ with $U \subset X^n$ and $Q; Y^n \rightarrow gU^n$ such that for all $u, u' \in U$, $s^n \in S^n$

$$\sum_{A \subseteq U, w \in A} Q(A|y^n) W^n(y^n | u, s^n) > 1 - \lambda_1 \quad (4.1)$$

$$\sum_{A \subseteq U, w \in A} Q(A|y^n) W^n(y^n | u, s^n) < \lambda_2. \quad (4.2)$$

(Here we note that $A$'s in (4.1) and (4.2) are "decoding sets" for $u$ and $u'$ respectively.) We extract an NRI-code $(u, D_u)_{u \in U}$ with error probability $\lambda'_1, \lambda'_2$ by letting

$$D_u = \left\{ y^n ; \sum_{A \subseteq U, w \in A} Q(A|y^n) \geq \alpha \right\}. \quad (4.3)$$

Then by (4.1) for all $s^n$

$$1 - \lambda_1 < \sum_{y^n \in D_u} \sum_{A \subseteq U, w \in A} Q(A|y^n) W^n(y^n | u, s^n) + \sum_{y^n \in D_u} \sum_{A \subseteq U, w \in A} Q(A|y^n) W^n(y^n | u, s^n)$$

$$\leq \sum_{y^n \in D_u} W^n(D_u | u, s^n) + \alpha W^n(D_u | u, s^n)$$

$$\leq W^n(D_u | u, s^n) + \alpha \quad (4.3)$$

On the other hand, (4.2) implies that for all $s^n$

$$\lambda_2 > \sum_{y^n \in D_u} \sum_{A \subseteq U, w \in A} Q(A|y^n) W^n(y^n | u, s^n)$$

$$\geq \alpha W^n(D_u | u, s^n) \quad (4.4)$$

Finally, for instance with the choice $\alpha = \sqrt{\lambda_2}$ we conclude from (4.3) and (4.4) that we can achieve

$$\lambda'_1 = \lambda + \sqrt{\lambda_2} \quad \text{and} \quad \lambda'_2 = \sqrt{\lambda_2}.$$
or, by (5.1), \( I(X \wedge Y' | X') > \delta \), if \( P_{X|Y}, P_{Y'|X'} \in \mathbb{W} \). Thus because the number of conditional types is a polynomial in \( n \), (5.8) follows from the fact that for \((X, X', Y) \in \mathbb{E}(P), P_{X|Y}, P_{Y'|X'} \in \mathbb{W} \)
\[
\frac{1}{n} \log \mathbb{W}^n(T_{X|X'}^{n}(u, u') \leq H(Y | X' X) - H(Y | X')
\]
and \( W^n(B(u'))(u', s^n) > 1 - 2^{-\eta n} \) for all \( s^n \) and suitable \( \eta > 0 \).

2) We have to show that for all \( u \) and \( s^n \)
\[
W^n(D_u | u, s^n) > 1 - \lambda_1, \tag{5.10}
\]
Since for all \( s^n \in S^n \), by (5.7) \( W^n(B(u'))(u', s^n) > 1 - 2^{-\eta n} \) for suitable \( \eta > 0 \), by (5.6) it is sufficient for (5.10) to show \( W^n(E(u, s^n)) \) is exponentially small. Indeed, by the definition of \( E(u, s^n) \) and (5.4)
\[
\frac{1}{n} \log W^n(E(u)) \leq \log \mathbb{W}^n(T_{X|X'}^{n}(u, u') \leq H(Y | X' X) - H(Y | X) = H(Y | X' X') + \theta - \rho(\alpha - \theta) < 0 \tag{5.11}
\]
if we choose \( \theta < \rho \).

VI. PROOF OF THE LEMMA

Denote by \( A_{n, \lambda_1, \lambda_2}(\mathcal{W})(A_{n, \lambda_1, \lambda_2}(W^n)) \) the maximal \( M \) such that an \((n, \lambda_1, \lambda_2) \) NRL-code for \( \mathcal{W} \) exists, \( \forall \mathcal{W} \).

1) \( C_{\mathbb{N}L}(W^n) \geq \max_p [h(p) + p C_{\mathbb{N}L}(W)] \)

Let \( \{D_{x,u} : u \in \mathcal{U} \} \) be an \((n, \lambda_1, \lambda_2) \) NRL-code for \( W^n \). We partition \( \mathcal{U} \) into subsets \( D_{x,u} \) in a suitable manner according to the number of zeros in the codewords. Then there must be a \( k \) such that
\[
|D_{x,u}| \geq \frac{1}{n} \log |D_{x,u}|. \tag{6.1}
\]
Moreover, the relation \( u \sim u' \) defined by the rule
\( u \sim u' \) if \( x_1 = 0 \) exactly if \( x'_1 = 0 \) for \( u = (x_1, \ldots, x_n), u' = (x'_1, \ldots, x'_n) \) holds in \( D_{x,u} \) is an equivalent relation, which further partitions \( D_{x,u} \) into at most \( \binom{n}{k} \) equivalence classes \( \{V_{u,j} \}_{j=1}^{n-k} \). J \( \leq \binom{n}{k} \).

All codewords in a fixed \( V_{u,j} \) have \( k \) zero components at the same coordinates. By our assumption at these coordinates the outputs are zeros whenever the inputs fall into \( V_{u,j} \). So we can obtain an \((n-k, |V_{u,j}|, \lambda_1, \lambda_2) \) NRL-code by deleting the \( k \) components from codewords in \( V_{u,j} \) and corresponding components from decoding sets. Therefore,
\[
A_{n, \lambda_1, \lambda_2}(W^n) \leq \binom{n}{k} A_{n-k, \lambda_1, \lambda_2}(W), \tag{6.2}
\]

2) \( C_{\mathbb{N}L}(W^n) \geq \max_p [h(p) + p C_{\mathbb{N}L}(W)] \)

We have to find an \((n, \lambda_1, \lambda_2) \) code for \( W^n \) with
\[
M \geq 2^{-\eta n} \binom{n}{k} A_{n-k, \lambda_1, \lambda_2}(W) \tag{6.2}
\]
for an arbitrarily small \( \eta \).

We first find, by an greedy algorithm, a set \( B \) of binary sequences with Hamming weight \( n - k \), pairwise Hamming distance not less than \( 2^{n-k} \), and size
\[
|B| \geq 2^{-\eta n} \binom{n}{k} \tag{6.3}
\]
where \( \eta \) is a positive constant depending on \( \varepsilon \) and \( \eta \to 0 \) as \( \varepsilon \to 0 \). Let \( \{u, D_u : u \in \mathcal{U} \} \) be an NRL-code of length \( n - k \) for \( W \) achieving \( A_{n-k, \lambda_1, \lambda_2}(W) \). We define for \( b^n \in B \) a subset \( \mathcal{U}^n(b^n) \) in \( \mathbb{X}^n \)
\[
\mathcal{U}^n(b^n) = \{x^n : x_i = 0 \text{ if } t_i \neq t_j, (x_{t_1}, \ldots, x_{t_{n-k}}) \in \mathcal{U} \}
\]
if \( b_j = 1 \), for \( 1 \leq t_1 < t_2 < \cdots < t_{n-k} \leq n \).

and let \( \mathcal{U}^n = \cup_{b^n \in B} \mathcal{U}^n(b^n) \).

For \( u^n = (x_1, \ldots, x_n) \in \mathcal{U}^n(b^n) \), the decoding set is defined by
\[
D_u^n = \{y^n : y_{t_i} = 0 \text{ if } t_i \neq t_j \text{ and } y_{t_1}, \ldots, y_{t_{n-k}} \in \mathcal{D}_u \}
\]
for \( t_1, \ldots, t_{n-k} \) in (6.5) and \( u = (x_1, \ldots, x_n) \in \mathcal{D}_u \).

Then for all \( s^n \in \mathcal{S}^n \) \( \mathcal{U}^n(b^n), s^n) > 1 - \lambda_1, \) and for all \( u^n, u'^n \in \mathcal{U}^n(b^n), s^n) \leq \lambda_2, \) since \( \{u, D_u : u \in \mathcal{U} \} \) has error probability \( \lambda_1 \). Moreover, for all \( s^n \in \mathcal{S}^n \), \( u^n \in \mathcal{U}^n(b^n), u'^n \in \mathcal{U}^n(b^n), b^n, b'^n \in B, \) and \( b^n \not= b'^n \)
\[
W^n(D_u^n, u^n, s^n) \leq W^n(b^n, b'^n) \leq \max_{s \in \mathcal{S}} W(0 | s, s)
\]
if \( u \) is sufficiently large.

Thus \( \{u^n, D_u^n : u^n \in \mathcal{U}^n \} \) is a desired code.

VII. PROOF OF THEOREM 5

Without loss of generality assume that for \( u \neq u', V(-, u) = V(-, u') \).

It was shown in [1] that for any channel \( V : X \rightarrow Y \) without two identical rows, any \( u_1, u_2 > 0, \) sufficiently large, and any \( U \subset \mathbb{X}^n \) such that for all \( u, u' \in U, d_H(u, u') \leq \varepsilon, \) there exists a family of subsets of \( \mathbb{X}^n \), say \( \mathcal{D}_u, u \in U \), such that \( V^n \mathcal{D}_u, u \in U \) \( \geq 1 - u_1 \) and \( V^n \mathcal{D}_u, u \in U \) \( \leq u_2 \) for all \( u, u' \neq u \), where \( d_H \) is the Hamming distance. Let us fix a family
\[
\{\{X(1|s), \ldots, X(j|s)\} : s \in S\}
\]

of partitions in Section 1.6. For \( x^n, x'^n \in \mathbb{X}^n, s \in S \), we define
\[
d_s(x^n, x'^n) = \{t : x_t \in X(j|s), x'_t \in X(j'|s) \text{ with } j \neq j'\}
\]

Thus by the above auxiliary result, we have that for any \( \lambda_1, \lambda_2 > 0, \) sufficiently large, and any \( U \subset \mathbb{X}^n \) such that for all \( s \in S, u, u' \in U \)
\[
d_s(u, u') \geq n \varepsilon
\]
there is a family of subsets in \( \mathbb{X}^n \), say \( \mathcal{D}_u, u \in U, u \neq u', u \in S \), such that for all \( u, u' \in U, u \neq u', s \in S \)
\[
V^n \mathcal{D}_u, u \in U \geq 1 - \lambda_1, \quad V^n \mathcal{D}_u, u \in U \leq \lambda_2, \tag{6.3}
\]

To find a good NRL-code for \( V \), we first find a \( U \) satisfying (7.2). Let \( X \) be the RV achieving the extremal value in the theorem. Then for any fixed \( u \in T_X^n \), if (7.2) is violated for \( s, u, u' \in T_X^n \), then there exists a pair \((X, X')\) such that with \( u' \in T_X^n(X, u) \), \( P_{X'} = P_X \), and \( E_u d(X, X') < \varepsilon \). For such \((X, X')\)
\[
\frac{1}{n} \log |T_{X|X'}^{n}(u)| = H(X' | X) + o(1)
\]

by the data processing inequality and by Fano’s inequality

\[ I(X \land X') \geq I(X(s) \land X'(s)) \geq H(\hat{X}(s)) - \alpha(e) \]  \tag{7.5} 

as \( E_d(X, X') < \varepsilon \) implies that \( Pr(\hat{X}(s) \neq \hat{X}'(s)) < \varepsilon \), where \( \alpha(e) \) is a constant depending on \( e \) and \( \alpha \to 0 \) as \( e \to 0 \).

Denote by \( \mathcal{Q} = \{(X, X') : F_X = F_X' \text{ and } d_4(X, X') < \varepsilon \text{ for some } s \in S\} \).

Then the total number of \( u \)'s in \( T_k^n \), such that for an \( s \in S \) \( (7.2) \) does not hold, are not larger than \( 2^H(X, X') \cdot \mathcal{Q} \cdot H(X) \cdot (e+1) \).

Consequently, by the greedy algorithm one can find a \( U \subset T_k^n \) satisfying (7.2) such that

\[ \frac{1}{n} \log |\mathcal{U}| \geq H(X) - \max_{(X, X') \in \mathcal{Q}} H(X') + o(1) \]

\[ = \min_{(X, X') \in \mathcal{Q}} I(X \land X') + o(1) \quad \text{(since } H(X') = H(X)) \]

\[ \geq \min H(\hat{X}(s)) - \alpha(s) + o(1) \quad \text{(by (7.5)).} \]

Then the following procedure works,

1) For all \( a \in \mathcal{X} \) define \( a' = (a_1, \ldots, a_l) \). Choose a sufficiently small \( \delta \) and a sufficiently large \( \ell \) such that for all \( a \in \mathcal{X}, V \in \mathcal{V} \)

\[ V'(T_{\beta, \lambda}(a')|a|) > 1 - \frac{1}{2|\mathcal{V}|} \lambda \]

and for all \( V, V' \in V \) there is an \( a \in \mathcal{X} \) such that

\[ T_{\beta, \lambda}(a') \cap T_{\beta, \lambda}(a') = \emptyset \]

where

\[ \lambda = \min(\lambda_1, \lambda_2). \]

Then the encoder uses \( |\mathcal{K}| \) blocks of length \( \ell \) to send \( a' \) for all \( a \in \mathcal{X} \). The decoder tries to find a \( V \in \mathcal{V} \) (and the corresponding state \( s \in S \)) such that for all \( a \in \mathcal{X} \), the ath-block output of length \( \ell \) falls into \( T_{\beta, \lambda}(a') \). If he can find it, it must be unique by our construction, otherwise, the decoder just declares an error. When any \( V \in \mathcal{V} \) governs the transmission, the decoder successfully estimates \( V \) with probability at least \( 1 - \frac{1}{2} \lambda \).

2) Knowing the state \( s \) governing the transmission, the decoder uses the decoding sets \( \{D_a(s) : s \in S\} \) in (7.3) to identify the message for which the two kind of error probabilities are \( \lambda_1/2 \) and \( \lambda_2/2 \), respectively. Thus the two kind of error probabilities totally do not exceed \( \lambda_1 \) and \( \lambda_2 \), respectively.

This and (7.6) complete the proof of the direct part (by choosing \( \ell/n \) arbitrarily small).

To prove the converse we partition the set \( \mathcal{U} \) of codewords of a given NRI-code of length \( n \) according to the types. Then we can find an RV \( X \) and a \( U \subset \mathcal{U} \) such that

\[ \mathcal{U} \subset T_k^n \quad \text{and} \quad |\mathcal{U}| \geq (n+1)^{-k} |\mathcal{K}|. \]

Let \( \varphi_s \) be the mapping \( X^n \to \{1, 2, \ldots, j_s\} \) for a fixed \( s \in S \) such that \( \varphi_s(x^n) = (i_1, \ldots, i_s) \), if \( x_s \in \mathcal{X}(i_s) \). Then for all \( s \in S \) there are no \( u, u' \in U \) with \( \varphi_s(u) = \varphi_s(u') \) and \( u \neq u' \). Furthermore, the mapping \( \varphi_s \) sends \( T_k^n \) to \( T_k^n \).

Consequently, for all \( s \)

\[ \frac{1}{n} \log |\mathcal{U}| \leq \frac{1}{n} \log |T_k^n(\varphi_s)| = H(\hat{X}(s)) + o(1). \]

Thus the converse holds.

REFERENCES


A Simple Randomized Algorithm for Sequential Prediction of Ergodic Time Series

László Györfi, Fellow, IEEE, Gábor Lugosi, Member, IEEE, and Gáspár Morvai

Abstract—We present a simple randomized procedure for the prediction of a binary sequence. The algorithm uses ideas from recent developments of the theory of the prediction of individual sequences. We show that if the sequence is a realization of a stationary and ergodic random process then the average number of mistakes converges to 1 minus the optimum, given by the Bayes predictor. The desirable finite-sample properties of the predictor are illustrated by its performance for Markov processes. In such cases the predictor exhibits near-optimal behavior even without knowing the order of the Markov process. Predictors with side information is also considered.

Index Terms—Ergodic processes, Markov processes, on-line learning, sequential prediction, universal prediction.

I. INTRODUCTION

We address the problem of sequential prediction of a binary sequence. A sequence of bits \( y_1, y_2, \ldots \in \{0, 1\} \) is hidden from the predictor. At each time instant \( t = 1, 2, \ldots \), the predictor is

Manuscript received May 11, 1998; revised June 3, 1999. The work of G. Lugosi was supported by DFG under Grant PB86-0300.
L. Györfi is with the Department of Computer Science and Information Theory, Technical University of Budapest, Budapest, Hungary (e-mail: gyorfi@inf.bme.hu).
G. Lugosi is with the Department of Economics, Pompeu Fabra University, 08005 Barcelona, Spain (e-mail: lugosi@upf.es).
G. Morvai is with the Research Group for Informatics and Elettronics, Hungarian Academy of Sciences, Budapest, Hungary (e-mail: morvai@inf.bme.hu).
Correspondence to P. Mó Richie, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks. Publisher Item Identifier S 0018-9448(99)08208-5.

0018-9448/99$10.00 © 1999 IEEE