[3] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[4] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 267–271, 1965.
[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[6] A. El-Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 851–857, Nov. 1982.
[7] A. Lapidoth, "On the role of missmatch in rate-distortion theory," *IEEE Trans. Inform, Theory*, vol. 43, pp. 38–47, Jan. 1997.
[8] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inform. Theory*, vol. 40, pp. 2026–2031, Nov. 1994.
[9] T. Linder, R. Zamir, and K. Zeger, "The multiple description rate region for high resolution source coding," in *Proc. Data Compression Conf.* (Snowbird, UT, Mar. 1998), pp. 149–158.
[10] L. H. Ozarow, "On the source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, pp. 1909–1922, 1980.
[11] D. J. Sakrison, "The rate of a class of random processes," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 10–16, Jan. 1970.
[12] V. A. Vaishampayan, "Design of entropy-constrained multiple-description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 40, pp. 245–251, Jan. 1994.
[13] R. Zamir and M. Feder, "Information rates of pre/post filtered dithered quantizers," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1340–1353, Sept. 1996.
[14] Z. Zhang and T. Berger, "Multiple description source coding with no excess marginal rate," *IEEE Trans. Inform. Theory*, vol. 41, pp. 349–357, Mar. 1995.
[15] J. Ziv, "On universal quantization," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 344–347, May 1985.
[16] Y. Frank-Dayan, "Dithered lattice quantization for multiple descriptions," M.S. thesis, Tel-Aviv University, 1999.
[17] R. Zamir, "Shannon-type bounds for multiple descriptions of a stationary source," *J. Statist. Planning and Inference*, to be published in a special issue devoted to the Maine Conference, held in July 1997.

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 (CC) 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 $C_{\mathrm{NRI}}(W)$, 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 $\mathcal{W}$. In fact, for 0-1-matrices only in $\mathcal{W}$ we are—exactly as for transmission—led to the equivalent zero-error-capacity of Shannon (see [3]). But for general $\mathcal{W}$ the identification capacity $C_{\mathrm{NRI}}(\mathcal{W})$ is quite different from the transmission capacity $C(\mathcal{W})$. An observation is that the separation codes of [1] are also relevant here. We present a lower bound on $C_{\mathrm{NRI}}(\mathcal{W})$. It implies for instance for

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

that $C_{\mathrm{NRI}}(\mathcal{W}) = 1$, which is obviously tight. It exceeds $C(\mathcal{W})$, 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 $\overline{L}$ (which we call an NRA-code) can be partitioned into $\overline{L}2^{n\varepsilon}$ transmission codes. This gives a nonsingle-letter 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 $C_I(\mathcal{W})$ and $C_{\mathrm{NRI}}(\mathcal{W})$.

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

$$\mathcal{W} = \{W(\cdot|\cdot, s)\colon\ s \in \mathcal{S} \text{ finite}\},$$

and the compound channel (CC) $\mathcal{C}(\mathcal{W})$ associated with $\mathcal{W}$.

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

$$W^n(\mathcal{D}_u|u, s^n) > 1 - \lambda_1 \tag{1.1}$$

and

$$W^n(\mathcal{D}_u|u', s^n) < \lambda_2. \tag{1.2}$$

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

The authors are with Facultät für Mathematik, Universität Bielefeld, 333501 Bielefeld, Germany.

Communicated by I. Csiszár, Associate Editor for Shannon Theory.
Publisher Item Identifier S 0018-9448(99)07672-5.

Here for $s^n = (s_1, \cdots, s_n)$, $y^n = (y_1, \cdots, y_n)$, and $u = (u_1, \cdots, u_n)$

$$W^n(y^n|u, s^n) = \prod_{t=1}^{n} W(y_t|u_t, s_t).$$

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

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 $\log|\mathcal{X}|$ even before the concept of identification was introduced in [2].

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

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

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

$$\mathcal{D}(u, u') \cap \mathcal{D}(u', u) = \emptyset (u \neq u'), \tag{1.3}$$

and

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

Notice that with the choice

$$\mathcal{D}(u, u') = \mathcal{D}_u \backslash \mathcal{D}_{u'} \tag{1.5}$$

we can associate with every $(n, M, \lambda_1, \lambda_2)$ ID-code (respectively, NRI-code) an (n, M, λ) 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 (soft)-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-codes, if in the definition $\mathcal{U} \subset \mathcal{P}(\mathcal{X}^n)$. For the CC only sequences $s^n = (s, \cdots, s) (s \in \mathcal{S})$ are considered and the code constraints are modified accordingly.

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

$$\overline{L}_{\mathcal{U}} = \max_{u \in \mathcal{U}, s^n \in \mathcal{S}^n} \overline{L}(u, s^n) \tag{1.6}$$

where

$$\overline{L}(u, s^n) = \sum_{y^n \in \mathcal{D}_u} L(y^n) W^n(y^n|u, s^n) \tag{1.7}$$

and

$$L(y^n) = |\{u' \in \mathcal{U}: y^n \in \mathcal{D}_{u'}\}|. \tag{1.8}$$

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

$$\overline{L}_{\mathcal{U}} \leq \overline{L} \tag{1.9}$$

and for all $u, u' \in \mathcal{U}$, $u \neq u'$, there is a partition of $\mathcal{D}_u \cap \mathcal{D}_{u'}$, say $\{A(u, u'), A(u', u)\}$, such that

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

Obviously, (1.2') holds for any partition of $\mathcal{D}_u \cap \mathcal{D}_{u'}$, $\lambda = \lambda_2$, if (1.2) is true and for $\mathcal{D}(u, u') = (\mathcal{D}_u \backslash \mathcal{D}_{u'}) \cup A(u, u')$, $\lambda = \lambda_1 + \lambda_2$, (1.4) holds whenever (1.2') holds. On the other hand, for $A(u', u) = \mathcal{D}_u \cap \mathcal{D}_{u'} \cap D(u', u)$, (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 $(n, M, \lambda_1, \lambda_2, \overline{L})$ NRA-code for the AVC $\mathcal{W}$ defined above. For every $\varepsilon > 0$, $0 < \lambda_1 < \lambda$ there exists a λ^* such that for all $\lambda_2 \leq \lambda^*$ and sufficiently large n the NRA-code can be partitioned into K transmission subcodes for $\mathcal{W}$ with maximal probability of error λ_2, if for $\overline{\ell} = (1/n) \log \overline{L}$

$$\frac{1}{n} \log K > \overline{\ell} + \varepsilon.$$

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

B. A Formula for $C(\mathcal{W})$

From Theorem 1 we get a nonsingle letter characterization for $C(\mathcal{W})$ 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 already 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 $\mathcal{W}$ a graph $G^m(\mathcal{W}) = \{\mathcal{X}^m, \mathcal{E}_m\}$ such that $(x^m, x'^m) \in \mathcal{E}_m$ iff there are PD's $\pi, \pi' \in \mathcal{P}(\mathcal{S}^m)$ such that

$$\sum_{s^m} \pi'(s^m) W^m(\cdot|x^m, s^m) \equiv \sum_{s^m} \pi(s^m) W^m(\cdot|x'^m s^m) \tag{1.10}$$

(or

$$\operatorname{conv}\{W^m(\cdot|x^m, s^m): s^m \in \mathcal{S}^m\} \cap \operatorname{conv}\{W^m(\cdot|x'^m, s^m): s^m \in \mathcal{S}^m\} \neq \emptyset).$$

Denote by $\mathcal{I}_m$ the family of independent sets of the graph. Then $\mathcal{U} \in \mathcal{I}_m$ is an NRS-code and we have the following auxiliary result.

Lemma 8 [1]: For any $\varepsilon > 0$, $\lambda > 0$, and sufficiently large n, one can choose $\{\mathcal{D}(u^n, u'^n): u^n \neq u'^n, u^n, u'^n \in \tilde{\mathcal{U}}\}$ suitable to obtain an SP-code with probability of error λ, if the pairwise Hamming distances (with respect to alphabet $\mathcal{U}$) of codewords in $\tilde{\mathcal{U}} \subset \mathcal{U}^n$ are not smaller than $n\varepsilon$.

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

$$\overline{L}_{\mathcal{U}, \lambda_1} = \min\{\overline{L}_{\mathcal{U}}((\mathcal{D}_u)_{u \in \mathcal{U}}): (\mathcal{D}_u)_{u \in \mathcal{U}} \text{ satisfies (1.1) for } \lambda_1\}. \tag{1.11}$$

In other terms clearly,

$$\overline{L}_{\mathcal{U}, \lambda_1} = \min_{(\mathcal{D}_{u''})_{u'' \in \mathcal{U}} \text{ with (1.1)}} \max_{u \in \mathcal{U}, s^n \in \mathcal{S}^n} \sum_{u' \in \mathcal{U} \backslash \{u\}} \cdot W^m(\mathcal{D}_u \cap \mathcal{D}_{u'}|u, s^m) + 1. \tag{1.12}$$

Theorem 2:

$$C(\mathcal{W}) = \sup_m \inf_{\lambda_1 > 0} \max_{\mathcal{U} \in \mathcal{I}_m} \frac{1}{m} \log \frac{|\mathcal{U}|}{|\overline{L}_{\mathcal{U}, \lambda_1}|}. \tag{1.13}$$

C. On Randomization in the Decoding

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

Theorem 3: For every AVC $\mathcal{W}$ under the maximal error probability criterion randomization in the decoding does not lead to higher capacities than i) $C(\mathcal{W})$, ii) $C_{\mathrm{NRI}}(\mathcal{W})$, and iii) $C_I(\mathcal{W})$, 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 $C_{\mathrm{NRI}}(\mathcal{W})$

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

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

$$\mathcal{Q}(P,\mathcal{W}) = \{(X,X',Y)\colon P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}}, \; P_X = P_{X'} = P \text{ and } X, X', Y \text{ form a Markov chain in this order}\}$$

and set

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

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

$$C_{\mathrm{NRI}}(\mathcal{W}) \geq \max_{P} \hat{I}(\mathcal{P},\mathcal{W}). \tag{1.14}$$

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

Corollary: The quantities in the inequality

$$C_{\mathrm{NRI}}(\mathcal{W}) \geq C(\mathcal{W}) \tag{1.15}$$

can be different.

This follows from

Example 1: For

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

Theorem 4 (or also the Lemma below) yields

$$C_{\mathrm{NRI}}(\mathcal{W}) = 1 \tag{1.16}$$

and

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

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

Example 2: Let

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

where $1 > q(s) > 0$ for all $s \in \mathcal{S}$ (finite). Then for $P \in \mathcal{P}(\mathcal{X})$ of the form $P = (p, 1-p)$, $p \in (0,1)$, $\mathcal{Q}(P,\mathcal{W}) = \emptyset$, and by (1.13) we get $C_{\mathrm{NRI}}(\mathcal{W}) \geq 1$. 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 $\mathcal{X} = \{1,2,\cdots,\alpha\}$, $\mathcal{Y} = \{0,1,\cdots,\beta\}$, $|\mathcal{S}| < \infty$, and $\max_{x\in\mathcal{X}} \max_{s\in\mathcal{S}} W(0|x,s) < 1$. Furthermore, consider for input alphabet $\mathcal{X} \cup \{0\}$ and output alphabet $\mathcal{Y}$ the AVC

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

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

Then

$$C_{\mathrm{NRI}}(\mathcal{W}^*) = \max_{0\leq p\leq 1} [h(p) + pC_{\mathrm{NRI}}(\mathcal{W})].$$

E. Combinatorial Problem Related to $C_{\mathrm{NRI}}\left(\left\{\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right), \left(\begin{smallmatrix}1-\delta & \delta\\ \delta & 1-\delta\end{smallmatrix}\right)\right\}\right)$

Finding $C_{\mathrm{NRI}}(\mathcal{W})$ for the special case

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

is already a formidable task.

By Theorem 4 and numerical computations of B. Balkenhol

$$C_{\mathrm{NRI}}(\mathcal{W}) \geq C(\mathcal{W}) = 1 - h(\delta)$$

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 $\mathcal{B}(u^n, d) \subset \{0,1\}^n$ 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 $x^n \in A$

$$\left|\mathcal{B}(x^n, n\delta) \cap \left[\bigcup_{y^n \in A\setminus\{x^n\}} B(y^n, n\beta)\right]\right| < \lambda |B(x^n, n\delta)|.$$

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

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

Theorem 5: For a CC $\mathcal{V} = \{V(\cdot|\cdot,s)\colon s \in \mathcal{S}\}$ with $|\mathcal{S}| < \infty$

$$C_{\mathrm{NRI}}(\mathcal{V}) = \max_{\hat{X}} \min_{s\in\mathcal{S}} H(\hat{X}(s)).$$

II. Proof of Theorem 1

We color the M codewords in $\mathcal{U}$ of an $(n, M, \lambda_1, \lambda_2, L)$ NRA-code $\{(u, \mathcal{D}_u)\colon u \in \mathcal{U}\}$ 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 $s^n \in \mathcal{S}^n$ and $u \in \mathcal{U}$ and partition $\mathcal{U}\setminus\{u\}$ into two parts $\mathcal{U}^{(i)}$ $(i = 1,2)$ such that $u' \in \mathcal{U}^{(1)}$ iff

$$W^n(A(u',u)|u,s^n) \leq \frac{1}{n^2}. \tag{2.1}$$

Then

$$\begin{aligned}|\mathcal{U}^{(2)}| &< n^2 \sum_{u'\in\mathcal{U}^{(2)}} W^n(A(u',u)|u,s^n)\\ &\leq n^2 \sum_{u'\in\mathcal{U}^{(2)}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u,s^n)\\ &\leq n^2 \sum_{u'\in\mathcal{U}\setminus\{u\}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u,s^n)\\ &= n^2(\overline{L}(u,s^n) - 1) < n^2\overline{L}\end{aligned} \tag{2.2}$$

where for the equality we use the useful observation

$$\sum_{u'\in\mathcal{U}\setminus\{u\}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u,s^n) = \overline{L}(u,s^n) - 1 \tag{2.3}$$

for all $u \in \mathcal{U}$ and $s^n \in \mathcal{S}^n$.

For the fixed u, s^n, and $i = 1,2$ let

$$Z_{u'}^{(i)} = \begin{cases} W^n(A(u',u)|u,s^n), & \text{if } u' \in \mathcal{U}^{(i)} \text{ and } u' \text{ is colored by the same color as } u\\ 0, & \text{otherwise.}\end{cases} \tag{2.4}$$

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

$$\Pr\left\{\sum_{u'\in\mathcal{U}^{(i)}} Z_{u'}^{(i)} > \frac{\lambda - \lambda_1}{2}\right\} \leq (|\mathcal{S}^n||\mathcal{U}|)^{-1}\frac{\theta}{2} \tag{2.5}$$

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

$$\{\mathcal{U}_k\colon\ k=1,2,\cdots,K\}$$

such that for all $u \in \mathcal{U}_k$ $(k=1,2,\cdots,K)$, $s^n \in \mathcal{S}^n$

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

and so

$$W^n\left(\mathcal{D}_u \setminus \left(\bigcup_{u' \in \mathcal{U}_k \setminus \{u\}} A(u',u)\right)|u,s^n\right) \geq W^n(\mathcal{D}_u|i,s^n) - \sum_{u' \in \mathcal{U}_k \setminus \{u\}} W^n(A(u',u)|u,s^n) > 1-\lambda-(\lambda-\lambda_1) = 1-\lambda.$$

Thus if we let

$$\mathcal{D}'_u = \mathcal{D}_u \setminus \left\{\bigcup_{u' \in \mathcal{U}_k \setminus \{u\}} A(u,u')\right\}$$

for all $u \in \mathcal{U}_k$ $(k=1,2,\cdots,K)$ then $\{(u,\mathcal{D}'_u)\colon\ u \in \mathcal{U}_k\}$ $(k=1,2,\cdots,K)$ is the family of codes required by the theorem. We first show (2.5) for $i=1$. By the definition of the $Z^{(1)}_{u'}$'s

$$\begin{aligned}
&\Pr\left(\sum_{u' \in \mathcal{U}^{(1)}} Z^{(1)}_{u'} > \frac{\lambda-\lambda_1}{2}\right) \\
&\quad \leq \exp_e\left\{-n^{3/2}\frac{\lambda-\lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \exp_e\{n^{3/2}Z^{(1)}_{u'}\} \\
&\quad = \exp_e\left\{-n^{3/2}\frac{\lambda-\lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \left[\left(1-\frac{1}{K}\right)+\frac{1}{K} \cdot \exp_e\{n^{3/2}W^n(A(u,u')|u,s^n)\}\right] \\
&\quad = \exp_e\left\{-n^{3/2}\frac{\lambda-\lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \cdot\left[1+\frac{1}{K}(\exp_e\{n^{3/2}W^n(A(u',u)|u,s^n)\}-1)\right] \\
&\quad \leq \exp_e\left\{-n^{3/2}\frac{\lambda-\lambda_1}{2}\right\} \prod_{u' \in \mathcal{U}^{(1)}} \cdot\left[1+\frac{e}{K}n^{3/2}W^n(A(u,u')|u,s^n)\right] \\
&\quad \leq \exp_e\left\{-n^{3/2}\frac{\lambda-\lambda_1}{2}\right\} \prod_{u \in \mathcal{U} \setminus \{u\}} \cdot\left[1+\frac{e}{K}n^{3/2}W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u,s^n)\right] \\
&\quad \leq \exp_e\left\{-n^{3/2}\frac{\lambda-\lambda_1}{2}+\frac{e}{K}n^{3/2} \cdot \sum_{u \in \mathcal{U} \setminus \{u\}} W^n(\mathcal{D}_u \cap \mathcal{D}_{u'}|u,s^n)\right\} \\
&\quad = \exp_e\left\{-n^{3/2}\left[\frac{\lambda-\lambda_1}{2}-\frac{e}{K}(\overline{L}(u,s^n)-1)\right]\right\} \quad \text{(by (2.3))} \\
&\quad \leq \exp_e\left\{-n^{3/2}\left[\frac{\lambda-\lambda_1}{2}-\frac{e}{K}\overline{L}\right]\right\}. \qquad (2.6)
\end{aligned}$$

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

$$K \geq \frac{4e\overline{L}}{\lambda-\lambda_1}. \qquad (2.7)$$

To show (2.5) for $i=2$, it is sufficient to show that the probability of the event, that the number of $u' \in \mathcal{U}^{(2)}$ with the same color in the (random) coloring as u is larger than $\lfloor \lambda_2^{-1}(\lambda-\lambda_1/2)\rfloor = \kappa_2$, say, is not larger than the right-hand side (RHS) of (2.5).

That is,

$$\sum_{j>\kappa_2} \binom{|\mathcal{U}^{(2)}|}{j}\left(\frac{1}{K}\right)^j\left(1-\frac{1}{K}\right)^{(|\mathcal{U}^{(2)}|-j} \leq [|\mathcal{S}^n||\mathcal{U}||]^{-1}\frac{\theta}{2}. \qquad (2.8)$$

Indeed, if

$$\frac{|\mathcal{U}^{(2)}|}{K} < k_2 \qquad (2.9)$$

then by Stirling's formula and with $\kappa_2 = \lfloor \lambda_2^{-1}(\lambda-\lambda_1/2)\rfloor$

LHS of (2.8)

$$\begin{aligned}
&\leq |\mathcal{U}^{(2)}|\frac{e}{\sqrt{2\pi}}\sqrt{\frac{|\mathcal{U}^{(2)}|}{\kappa_2|\mathcal{U}^{(2)}|-\kappa_2}}\left(\frac{|\mathcal{U}^{(2)}|}{K\kappa_2}\right)^{\kappa_2} \cdot\left(\frac{|\mathcal{U}^{(2)}|(K-1)}{K(|\mathcal{U}^{(2)}|-\kappa_2)}\right)^{|\mathcal{U}^{(2)}|-\kappa_2} \\
&\leq |\mathcal{U}^{(2)}|\frac{e}{\sqrt{2\pi}}\left(\frac{|\mathcal{U}^{(2)}|}{K\kappa_2}\right)^{\kappa_2}\left(1+\frac{\kappa_2}{|\mathcal{U}^{(2)}|-\kappa_2}\right)^{|\mathcal{U}^{(2)}|-\kappa_2} \\
&\leq |\mathcal{U}^{(2)}|\frac{e}{\sqrt{2\pi}}\left(\frac{|\mathcal{U}^{(2)}|e}{K\kappa_2}\right)^{\kappa_2} \leq |\mathcal{U}^{(2)}|e\left(\frac{e|\mathcal{U}^{(2)}|^2}{K}\right)^{\kappa_2}.
\end{aligned}$$

Thus (2.8) holds if we choose

$$\begin{aligned}
K &> e|\mathcal{U}^{(2)}|\exp\left\{\frac{n}{\kappa_2}\log|\mathcal{S}||\mathcal{X}|^2+\frac{1}{\kappa_2}\log\frac{2e}{\theta}\right\} \\
&= |\mathcal{U}^{(2)}|\exp\{n\varphi_\theta(\kappa_2)\} \qquad (2.10)
\end{aligned}$$

where $\varphi_\theta(z)$ is a function of z whose values are arbitrarily small when z is arbitrarily large. By (2.2) it is sufficient for (2.9) and (2.10) to hold that

$$K > n^2\overline{L}\exp\{n\varphi_\theta(\kappa_2)\} \quad \text{(for sufficiently large } n\text{)}.$$

To satisfy the above inequality (2.7), we only need to choose

$$K > \frac{4\overline{L}}{\lambda-\lambda_1}n^2\exp\{n\varphi_\theta(\kappa_2)\}$$

and sufficiently small λ^* (and, therefore, λ_2) depending on λ, λ_1, and ε.

This completes the proof.

III. Proof of Theorem 2

The converse part is absolutely trivial because an (n,M)-code $\{(u,\mathcal{D}_u)\colon\ u \in \mathcal{U}\}$ with maximal probability of error λ_1 satisfies

$$\overline{L}_{\mathcal{U},\lambda_1} = 1 \quad \text{and} \quad \mathcal{U} \in \mathcal{I}_m.$$

The issue of the theorem is to show that one cannot do better by increasing the size of lists, namely, the direct part. This is an easy consequence of Theorem 1 and [1, Lemma 8] (see Section I-B).

For a fixed m, $\lambda_1 > 0$ assume that $\mathcal{U} \in \mathcal{I}_m$, $\{(u,\mathcal{D}_u)\colon\ u \in \mathcal{U}\}$ achieves the maximum in (1.13). Then we treat $\mathcal{U}$ as an input alphabet and $\mathcal{Y}^m$ as an output alphabet. Then one can find, by the greedy algorithm, a subset of codewords $\tilde{\mathcal{U}} \subset \mathcal{U}^\ell$ such that for all u^ℓ, $u'^\ell \in \tilde{\mathcal{U}}$, $d_H(u^\ell,u'^\ell) \geq \ell\varepsilon$ for any fixed ε, and

$$\log|\mathcal{U}| - \frac{1}{\ell}\log|\tilde{\mathcal{U}}| = o(1) \quad (\text{as } \varepsilon \to 0, \ell \to \infty). \qquad (3.1)$$

Let $\tilde{\mathcal{D}}_{u^\ell}$ for $u^\ell \in \tilde{\mathcal{U}}$ to be the union of Hamming balls with radius $\ell(\lambda_1+\varepsilon^2)$ and centers at the points in the Cartesian product

$\mathcal{D}_{u_1} \times \cdots \times \mathcal{D}_{u_\ell}$, for $u^\ell = (u_1, \cdots, u_\ell)$. Then (1.1) holds for sufficiently large ℓ, if λ_1 is replaced by $2\lambda_1$. Moreover, for any λ and sufficiently large ℓ, by [1, Lemma 8], (1.4) holds, for suitable $D(u^\ell, u'^\ell)$ and, therefore, (1.2') holds for suitable $A(u^\ell, u'^\ell)$ for all $u^\ell, u'^\ell \in \tilde{\mathcal{U}}$, $u^\ell \neq u'^\ell$.

To apply Theorem 1, we have to estimate $\overline{L}_{\tilde{\mathcal{U}}}$. Let us write $\mathcal{D}_{u_1} \times \cdots \times \mathcal{D}_{u_\ell} = \mathcal{D}_{u^\ell}$ and denote the Hamming ball with center u^ℓ and radius r in $\mathcal{U}^\ell$ by $B(u^\ell, r)$. Then for $L(\cdot)$ in (1.8)

$$u^\ell = (u_1, \cdots, u_\ell) \in \tilde{\mathcal{U}}$$
$$s^{m\ell} = (s_1^m, \cdots, s_\ell^m) \in \mathcal{S}^{m\ell}$$
$$v^\ell = (v_1, \cdots, v_\ell) \in \mathcal{D}_{u^\ell}$$

$$\sum_{v^\ell \in B(u^\ell, \ell(\lambda_1+\varepsilon^2))} L(v^\ell) W^{m\ell}(v^\ell | u^\ell, s^{m\ell}) \leq \sum_{J \subset [\ell]:\, |J| = \ell(\lambda_1+\varepsilon^2)} |\mathcal{U}|^{\ell(\lambda_1+\varepsilon^2)} \prod_{j \notin J} L(v_j) W^m(v_j | u_j, s_j^m).$$

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

$$\frac{1}{\ell} \log \overline{L}_{\tilde{\mathcal{U}}} \leq \frac{1}{\ell} \log \overline{L}_{\mathcal{U}} + \beta(\lambda_1, \varepsilon). \tag{3.2}$$

Finally, we choose $n = m\ell$ and apply Theorem 1 to

$$\{(u^\ell, \tilde{\mathcal{D}}_{u^\ell}) \colon u^\ell \in \tilde{\mathcal{U}}\}$$

to obtain a (transmission) subcode with probability of error $\varepsilon + \lambda_1$ and rate arbitrarily close to $(1/n) \log(|\mathcal{U}|/\overline{L}_{\tilde{\mathcal{U}}})$, when ℓ is arbitrarily large (depending on m) and λ_1 and ε in (3.2) are arbitrarily small. Since m is fixed when $n = m\ell + r$, $r < m$, we asymptotically loose 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 $(\mathcal{U}, Q)$ with $\mathcal{U} \subset \mathcal{X}^n$ and $Q \colon \mathcal{Y}^n \to 2^{\mathcal{U}}$ such that for all u, $u' \in \mathcal{U}$, $s^n \in \mathcal{S}^n$

$$\sum_{A \colon u \in A} \sum_{y^n \in \mathcal{Y}^n} Q(A|y^n) W^n(y^n | u, s^n) > 1 - \lambda_1 \tag{4.1}$$

$$\sum_{A \colon u' \in A} \sum_{y^n \in \mathcal{Y}^n} Q(A|y^n) W^n(y^n | u, s^n) < \lambda_2. \tag{4.2}$$

(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 $(u, \mathcal{D}_u)_{u \in \mathcal{U}}$ with error probability λ_1', λ_2' by letting

$$\mathcal{D}_u = \left\{ y^n \colon \sum_{A \colon u \in A} Q(A|y^n) \geq \alpha \right\}.$$

Then by (4.1) for all s^n

$$\begin{aligned} 1 - \lambda_1 &< \sum_{y^n \in \mathcal{D}_u} \sum_{A \colon u \in A} Q(A|y^n) W^n(y^n | u, s^n) + \sum_{y^n \in \mathcal{D}_u^c} \sum_{A \colon u \in A} Q(A|y^n) W^n(y^n | u, s^n) \\ &< \sum_{y^n \in \mathcal{D}_u} 1 \cdot W^n(y^n | u, s^n) + \sum_{y^n \in \mathcal{D}_u^c} \alpha W^n(y^n | u, s^n) \\ &\leq W^n(\mathcal{D}_u | u, s^n) + \alpha \quad \text{or} \quad W^n(\mathcal{D}_u | u, s^n) > 1 - \lambda_1 - \alpha. \end{aligned} \tag{4.3}$$

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

$$\begin{aligned} \lambda_2 &> \sum_{y^n \in \mathcal{D}_{u'}} \sum_{A \colon u' \in A} Q(A|y^n) W^n(y^n | u, s^n) \\ &\geq \alpha W^n(\mathcal{D}_{u'} | u, s^n) \quad \text{or} \quad W^n(\mathcal{D}_{u'} | u, s^n) < \frac{\lambda_2}{\alpha}. \end{aligned} \tag{4.4}$$

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

$$\lambda_1' = \lambda_2 + \sqrt{\lambda_2} \quad \text{and} \quad \lambda_2' = \sqrt{\lambda_2}.$$

Since for $(\lambda_1, \lambda_2) \to (0,0)$ also $(\lambda_1', \lambda_2') \to (0,0)$ we have established equality of the (weak) capacities.

V. Proof of Theorem 4

Fix $R < \hat{I}(P, \mathcal{W}) - (\varepsilon/2)$ and let

$$\mathcal{Q}_\delta(P, \mathcal{W}) \triangleq \{(X, X', Y) \colon P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}}, I(X \wedge Y | X') \leq \delta\}. \tag{5.1}$$

Then

$$\mathcal{Q}(P, \mathcal{W}) = \bigcap_{\delta > 0} \mathcal{Q}_\delta(P, \mathcal{W}) \tag{5.2}$$

and, by the continuity of the mutual information, there are $\alpha, \delta > 0$ such that for all $(X, X', Y) \in \mathcal{Q}_\delta(P, \mathcal{W})$

$$R < I(X' \wedge XY) - \alpha. \tag{5.3}$$

Next we apply the large deviation method in the standard way or directly use [10, Lemma 3] to obtain a set of codewords $\mathcal{U}' \subset \mathcal{X}^n$ such that $(1/n) \log |\mathcal{U}'| \sim R$ and for all $U \in \mathcal{U}'$, $P_{XX'} \in \mathcal{P}_n(\mathcal{X} \times \mathcal{X})$ (where $\mathcal{P}_n(\mathcal{Z})$ is the set of n-types over $\mathcal{Z}$)

$$\frac{1}{n} \log |\{u' \in \mathcal{U}' \colon (u, u') \in T_{XX'}^n\}| < (R - I(X \wedge X'))^+ + \theta \tag{5.4}$$

where θ is a positive number and can be choosen arbitrarily small and $a^+ = \max\{0, a\}$, if n is arbitrarily large. Thus by deleting the "bad codewords" from the neighborhoods of the codewords, we can obtain a subset $\mathcal{U} \subset \mathcal{U}'$ (for sufficiently large n) such that

$$\frac{1}{n} \log |\mathcal{U}| \geq \frac{1}{n} \log |\mathcal{U}'| - \frac{\varepsilon}{2} \tag{5.5}$$

and there is no pair (u, u') of codewords in $\mathcal{U}$ with $(u, u') \in T_{XX'}^n$ for RV's X and X' with $R \leq I(X \wedge X')$. We choose $\mathcal{U}$ as our set of codewords and

$$\mathcal{D}_u = B(u) \backslash E(u) \tag{5.6}$$

as decoding set for $u \in \mathcal{U}$, where

$$B(u) = \bigcup_{\overline{\overline{W}} \in \overline{\overline{\mathcal{W}}}} T_{\overline{\overline{W}}, \delta}^n(u), \qquad \text{for } \delta \text{ in (5.3)} \tag{5.7}$$

and $E(u)$ is the set of y^n's (in $\mathcal{Y}^n$) such that there exist a $u' \neq u$ and a triple $(X, X', Y) \in \mathcal{Q}_\delta$ with $(u, u', y^n) \in T_{XX'Y}^n$.

Analysis:

1) To show that for all $s^n \in \mathcal{S}^n$

$$W^n(\mathcal{D}_u | u', s^n) < \lambda_2, \qquad \text{if } u \neq u' \tag{5.8}$$

we partition $\mathcal{D}_u$ into polynomially many subsets according to the conditional types of y^n's, $P_{Y|XX'}(\cdot | u, u')$, for the u, u' in (5.8). By (5.6)

$$T_{Y|XX'}^n(u, u') \cap \mathcal{D}_u \neq \emptyset$$

implies

$$(X, X', Y) \notin \mathcal{Q}_\delta(P, \mathcal{W})$$

or, by (5.1), $I(X \wedge Y'|X') > \delta$, if $P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}}$. Thus because the number of conditional types is a polynomial in n, (5.8) follows from the fact that for $(X, X', Y) \notin \mathcal{Q}_\delta(P)$, $P_{Y|X}, P_{Y|X'} \in \overline{\overline{\mathcal{W}}}$

$$\frac{1}{n}\log \overline{\overline{W}}^n(\mathcal{T}^n_{Y|XX'}(u,u')) \lesssim H(Y|XX') - H(Y|X') = -I(X \wedge Y|X') < -\delta \quad (5.9)$$

and $W^n(B(u')|u', s^n) > 1 - 2^{-n\eta}$ for all s^n and suitable $\eta > 0$.

2) We have to show that for all u and s^n

$$W^n(\mathcal{D}_u|u, s^n) > 1 - \lambda_1. \quad (5.10)$$

Since for all $s^n \in \mathcal{S}^n$, by (5.7) $W^n(B(u)|u, s^n) > 1 - 2^{-u\eta}$ for suitable $\eta > 0$, by (5.6) it is sufficient for (5.10) to show $W^n(E(u)|u, s^n)$ is exponentially small. Indeed, by the definition of $E(u)$ and (5.4)

$$\begin{aligned}\frac{1}{n}\log W^n(E(u)|u,s^n) &\lesssim \frac{1}{n}\max_{(X,X',Y)\in\mathcal{Q}_\delta(P,\mathcal{W})} \log|\{u': u' \in T^n_{X'|X}(u)\}| \cdot |T^n_{Y|XX'}(u,u')|2^{-nH(Y|X)} \\ &\sim R - I(X \wedge X') + \theta + H(Y|XX') - H(Y|X) \\ &= R - I(X' \wedge XY) + \theta < -(\alpha - \theta) < 0 \end{aligned} \quad (5.11)$$

if we chose $\theta < \alpha$.

VI. Proof of the Lemma

Denote by $A_{n,\lambda_1,\lambda_2}(\mathcal{W})(\mathcal{A}_{n,\lambda_1,\lambda_2}(\mathcal{W}^*))$ the maximal M such that an $(n, M, \lambda_1, \lambda_2)$ NRI-code for $\mathcal{W}$ (for $\mathcal{W}^*$) exists.

1) $C_{\mathrm{NRI}}(\mathcal{W}^*) \le \max_p[h(p) + pC_{\mathrm{NRI}}(\mathcal{W})]$.

Let $\{(u, \mathcal{D}_u): u \in \mathcal{U}\}$ be an $(n, M, \lambda_1, \lambda_2)$ NRI-code for $\mathcal{W}^*$. We partition $\mathcal{U}$ into subsets $\{\mathcal{U}_k\}_{k=0}^n$ according to the number of zeros in the codewords. Then there must be a k such that

$$|\mathcal{U}_k| \ge \frac{1}{n}\log|\mathcal{U}_k|. \quad (6.1)$$

Moreover, the relation $u \sim u'$ in $\mathcal{U}_k$ defined by the rule "$u \sim u'$ if $x_t = 0$ exactly if $x'_t = 0$ for $u = (x_1, \cdots, x_n)$, $u' = (x'_t, \cdots, x'_n) \in \mathcal{U}_k$" is an equivalent relation, which further partitions $\mathcal{U}_k$ into at most $\binom{n}{k}$ equivalence classes $\{\mathcal{V}_{k,j}\}_{j=1}^J$, $J \le \binom{n}{k}$.

All codewords in a fixed $\mathcal{V}_{k,j}$ have k zero components at the same coordinates. By our assumption at all these coordinates the outputs are zeros whenever the inputs fall into $\mathcal{V}_{k,j}$. So we can obtain an $(n-k, |\mathcal{V}_{k,j}|, \lambda_1, \lambda_2)$ NRI-code by deleting the k components from codewords in $\mathcal{V}_{k,j}$ (and corresponding components from decoding sets). Therefore,

$$A_{n,\lambda_1,\lambda_2}(\mathcal{W}^*) \le n\binom{n}{n-k}A_{n-k,\lambda_1,\lambda_2}(\mathcal{W}).$$

2) $C_{\mathrm{NRI}}(\mathcal{W}^*) \ge \max_p[h(p) + pC_{\mathrm{NRI}}(\mathcal{W})]$

We have to find an $(n, M, \lambda_1, \lambda_2)$-code for $\mathcal{W}^*$ with

$$M \ge 2^{-n\eta}\binom{n}{n-k}A_{n-k,\lambda_1,\lambda_2}(\mathcal{W}) \quad (6.2)$$

for an arbitrarily small η.

We first find, by a greedy algorithm, a set B of binary sequences with Hamming weight $n-k$, pairwise Hamming distance not less than $2n\varepsilon$, and size

$$|B| \ge 2^{-n\eta}\binom{n}{k} \quad (6.3)$$

where η is a positive constant depending on ε and $\eta \to 0$ as $\varepsilon \to 0$. Let $\{(u, \mathcal{D}_u): u \in \mathcal{U}\}$ be an NRI-code of length $n-k$ for $\mathcal{W}$ achieving $A_{n-k,\lambda_1,\lambda_2}(\mathcal{W})$. We define for $b^n \in B$ a subset $\mathcal{U}^*(b^n)$ in $\mathcal{X}^{*n}$

$$\mathcal{U}^*(b^n) = \{x^n: x_t = 0 \text{ if } t \ne t_j, (x_{t_1}, \cdots, x_{t_{n-k}}) \in \mathcal{U}\} \quad (6.4)$$

if

$$b_{t_j} = 1, \qquad \text{for } 1 \le t_1 < t_2 < \cdots \le t_{n-k} \le n \quad (6.5)$$

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

For $u^* = (x_1, \cdots, x_n) \in \mathcal{U}^*(b^n)$, the decoding set is defined by

$$\mathcal{D}^*_{u^*} = \{y^n: y_t = 0 \text{ if } t \ne t_j \text{ and } (y_{t_1}, \cdots, y_{t_{n-k}}) \in \mathcal{D}_u\}$$

for $t_1, \cdots, t_{n-k}$ in (6.5) and $u = (x_{t_1}, \cdots, x_{t_{n-1}})$.

Then for all s^n $W^{*n}(\mathcal{D}^*_{u^*}|u^*, s^n) > 1 - \lambda_1$, and for all $u^*, u^{*'} \in \mathcal{U}^*(b^n), s^n \in \mathcal{S}^n, W^n(\mathcal{D}^*_{u^*}|u^{*'}, s^n) < \lambda_2$, since $\{(u, \mathcal{D}_u): u \in \mathcal{U}\}$ has error probability (λ_1, λ_2). Moreover, for all $s^n \in \mathcal{S}^n$, $u^* \in \mathcal{U}^*(b^n)$, $u^{*'} \in \mathcal{U}^*(b'^n)$, $b^n, b'^n \in B$, and $b^n \ne b'^n$

$$W^n(\mathcal{D}^*_{u^*}|u^{*'}, s^n) \le \overline{w}^{(1/2)d_H(b^n, b'^n)} \le \overline{w}^{n\varepsilon} < \lambda_2$$

for

$$\overline{w} \triangleq \max_{s\in\mathcal{S}}\max_{x\in\mathcal{X}} W(0|x, s)$$

if n is sufficiently large.

Thus $\{(u^*, \mathcal{D}^*_{u^*}): u^* \in \mathcal{U}^*\}$ is a desired code.

VII. Proof of Theorem 5

Without loss of generality assume that for $s \ne s'$, $V(\cdot|\cdot, s) \ne V(\cdot|\cdot, s')$.

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

$$\{\{\mathcal{X}(1|s), \cdots, \mathcal{X}(j_s|s)\}: s \in \mathcal{S}\}$$

of partitions in Section I-F. For $x^n, x'^n \in \mathcal{X}^n$, $s \in \mathcal{S}$, we define

$$d_s(x^n, x'^n) = |\{t: x_t \in \mathcal{X}(j|s), x'_t \in \mathcal{X}(j'|s) \text{ with } j \ne j'\}|. \quad (7.1)$$

Thus by the above auxiliary result, we have that for any $\lambda_1, \lambda_2, \varepsilon > 0$, sufficiently large n, and any $\mathcal{U} \subset \mathcal{X}^n$ such that for all $s \in \mathcal{S}$, $u, u' \in \mathcal{U}$

$$d_s(u, u') \ge n\varepsilon \quad (7.2)$$

there is a family of subsets in $\mathcal{Y}^n$, say $\mathcal{D}_u(s)$, $u \in \mathcal{U}$, $s \in \mathcal{S}$, such that for all $u, u' \in \mathcal{U}$, $u \ne u'$, $s \in \mathcal{S}$

$$V^n(\mathcal{D}_u(s)|u, s) > 1 - \frac{\lambda_1}{2} \quad \text{and} \quad V^n(\mathcal{D}_u(s)|u', s) < \frac{\lambda_2}{2}. \quad (7.3)$$

To find a good NRI-code for $\mathcal{V}$, we first find a $\mathcal{U}$ satisfying (7.2). Let X be the RV achieving the extremal value in the theorem. Then for any fixed $u \in \mathcal{T}^n_X$, if (7.2) is violated for s, u, and $u' \in \mathcal{T}^n_X$, then there exists a pair (X, X') such that with $u' \in \mathcal{T}^n_{X'|X}(u)$, $P_{X'} = P_X$, and $E_s\, d(X, X') < \varepsilon$. For such (X, X')

$$\frac{1}{n}\log|\mathcal{T}^n_{X'|X}(u)| = H(X'|X) + o(1) \quad (7.4)$$

and by the data processing inequality and by Fano's inequality

$$I(X \wedge X') \geq I(\hat{X}(s) \wedge \hat{X}'(s)) \geq H(\hat{X}(s)) - \alpha(\varepsilon) \qquad (7.5)$$

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

Denote by $\mathcal{Q} = \{(X, X'): P_{X'} = P_X$ and $d_s(X, X') < \varepsilon$ for some $s \in \mathcal{S}\}$.

Then the total number of u's in $\mathcal{T}_X^n$, such that for an $s \in \mathcal{S}$ (7.2) does not hold, are not larger than $2^{n[\max_{(X,X')\in\mathcal{Q}} H(X'|X)+o(1)]}$. Consequently, by the greedy algorithm one can find a $\mathcal{U} \subset \mathcal{T}_X^n$ satisfying (7.2) such that

$$\begin{aligned}\frac{1}{n}\log|\mathcal{U}| &\geq H(X) - \max_{(X,X')\in\mathcal{Q}} H(X'|X) + o(1)\\ &= \min_{(X,X')\in\mathcal{Q}} I(X \wedge X') + o(1) \quad (\text{since } H(X') = H(X))\\ &\geq \min_s H(\hat{X}(s)) - \alpha(s) + o(1) \quad (\text{by (7.5)}). \qquad (7.6)\end{aligned}$$

Then the following procedure works.

1) For all $a \in \mathcal{X}$ define $a^\ell = (a, \cdots, a)$. Choose a sufficiently small δ and a sufficiently large ℓ such that for all $a \in \mathcal{X}, V \in \mathcal{V}$

$$V^\ell(\mathcal{T}_{V,\delta}^\ell(a^\ell)|a^b) > 1 - \frac{1}{2|\mathcal{X}|}\lambda$$

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

$$\mathcal{T}_{V,\delta}^\ell(a^\ell) \cap \mathcal{T}_{V',\delta}^\ell(a^\ell) = \emptyset$$

where

$$\lambda \triangleq \min(\lambda_1, \lambda_2).$$

Then the encoder uses $|\mathcal{X}|$ blocks of length ℓ to send a^ℓ for all $a \in \mathcal{X}$. The decoder tries to find a $V \in \mathcal{V}$ (and the corresponding state $s \in \mathcal{S}$) such that for all $a \in \mathcal{X}$, the ath-block output of length ℓ falls into $\mathcal{T}_{V,\delta}^\ell(a^\ell)$. 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 succesfully 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 $\{\mathcal{D}_u(s): s \in \mathcal{S}\}$ 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 λ_1 and λ_2, respectively.

This and (7.6) complete the proof of the direct part (by choosing ℓ/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 $\mathcal{U}' \subset \mathcal{U}$ such that

$$\mathcal{U}' \subset \mathcal{T}_X^n \quad \text{and} \quad |\mathcal{U}'| \geq (n+1)^{-|\mathcal{X}|}|\mathcal{U}|.$$

Let φ_s be the mapping $\mathcal{X}^n \to \{1, 2, \cdots, j_s\}$ for a fixed $s \in \mathcal{S}$ such that $\varphi_s(x^n) = (i_1, \cdots, i_n)$, if $x_t \in \mathcal{X}(i_t|s)$. Then for all $s \in \mathcal{S}$ there are no $u, u' \in \mathcal{U}$ with $\varphi_s(u) = \varphi_s(u')$ and $u \neq u'$. Furthermore, the mapping φ_s sends $\mathcal{T}_X^n$ to $\mathcal{T}_{\hat{X}(s)}^n$. Consequently, for all s

$$\frac{1}{n}\log|\mathcal{U}'| \leq \frac{1}{n}\log|\mathcal{T}_{\hat{X}(s)}^n| = H(\hat{X}(s)) + o(1).$$

Thus the converse holds.

REFERENCES

[1] R. Ahlswede, "A method of coding and application to arbitrarily varying channels," *J. Combin. Inform. Syst. Sci.*, vol. 5, pp. 10–35, 1980.

[2] R. Ahlswede and G. Dueck, "Identification via channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 15–29. 1989.

[3] R. Ahlswede, "A note on the existence of the weak capacity for channels with arbitrarily varying channel probability functions and its relation to Shannon's zero error capacity," *Ann. Math. Statist.*, vol. 41, no. 3, pp. 1027–1033, 1970.

[4] ——, "A general theory of information transfer," *IEEE Trans. Inform. Theory*, submitted for publication.

[5] R. Ahlswede and B. Balkenhol, "Data compression for identification," in preparation.

[6] I. Csiszár and J. Körner, "On the capacity of the arbitrarily varying channel for maximum probability of error," *Z. Wahrscheinlichkeitstheorie u. verw. Geb.*, vol. 57, pp. 87–101, 1981.

[7] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie und verw. Geb.*, vol. 44, pp. 159–175, 1978.

[8] C. E. Shannon, "The zero error capacity of a noisy channel," *Trans. IRE Prof. Group Inform. Theory*, vol. PGIT-2, pp. 8–19, 1956.

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[10] I. Csiszár and P. Narayan, "The capacity of arbitrarily varying channels revisited: Positivity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, pp. 181–193, 1988.

A Simple Randomized Algorithm for Sequential Prediction of Ergodic Time Series

László Györfi, *Fellow, IEEE*, Gábor Lugosi, *Member, IEEE*, and Gusztáv 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, almost surely, to that of 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. Prediction 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 $y_1, y_2, \cdots \in \{0, 1\}$ is hidden from the predictor. At each time instant $i = 1, 2, \cdots$, the predictor is

Manuscript received May 11, 1998; revised June 3, 1999. The work of G. Lugosi was supported by DGES under Grant PB96-0300.

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 Electronics, Hungarian Academy of Sciences, Budapest, Hungary (e-mail: morvai@inf.bme.hu).

Communicated by P. Moulin, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(99)08208-5.