

# New Results in the Theory of Identification via Channels

Te Sun Han, *Fellow, IEEE*, and Sergio Verdú, *Senior Member, IEEE*

**Abstract**—The identification capacity is the maximal iterated logarithm of the number of messages divided by the blocklength that can be reliably transmitted when the receiver is only interested in deciding whether a specific message was transmitted or not. The identification coding theorem of Ahlswede and Dueck for single-user discrete memoryless channels states that the identification capacity is equal to the Shannon capacity. A new method to prove the converse to the identification coding theorem is shown to achieve the strong version of the result. Identification plus transmission (IT) coding, a variant of the original problem of identification via channels, is proposed in the context of a common problem in point-to-multipoint communication, where a central station wishes to transmit information reliably to one of  $N$  terminals, whose identity is not predetermined. We show that as long as  $\log \log N$  is smaller than the number of bits to be transmitted, IT codes allow information transmission at channel capacity.

**Index Terms**—Identification via channels, channel capacity, coding theorems, point-to-multipoint communication.

## I. INTRODUCTION

THE award-winning work of R. Ahlswede and G. Dueck [1], [2] has opened a new and fertile area in the Shannon theory. In Shannon's formulation of the problem of reliable transmission through noisy channels, the decoder selects one of the possible messages based on the observation of the output codeword in such a way that

$$P[a \text{ is selected} | a \text{ is transmitted}] \geq 1 - \lambda_1, \quad (1)$$

for all  $a = 1, \dots, M$ ,

with arbitrarily small  $\lambda_1$ .

In the formulation of the problem of identification through noisy channels [1] the decoder is allowed to select a list of messages (whose size is not constrained) such that not only (1) is satisfied but any given message is unlikely to belong to

the output list unless it is equal to the transmitted message:

$$P[b \text{ is selected} | a \text{ is transmitted}] \leq \lambda_2, \quad \text{for all } b \neq a. \quad (2)$$

Identification coding is suitable in situations where the recipient is only interested in verifying whether a certain message (unknown to the encoder) is the transmitted message or not. If an identification code is used, then the recipient simply checks whether its message is in the output list. The number of messages that can be reliably transmitted using an identification code turns out to be *doubly exponential* in the blocklength, in contrast to the single exponential behavior of conventional transmission codes. This is due to the weaker reliability imposed by the identification coding formulation. The identification coding theorem for single-user channels without feedback states that the identification capacity (the maximum achievable iterated logarithm of the number of messages divided by the blocklength as the blocklength goes to infinity) is equal to the Shannon capacity of the channel. This theorem has been shown in [1] for discrete memoryless channels (DMC). The direct part of the theorem follows from a combinatorial result (akin to the Gilbert bound on the rate of codes with prescribed minimum distance) giving the largest number of fixed-size subsets of any set whose pairwise overlap does not exceed a certain percentage of their size. The major achievement in [1] is the converse part of the theorem, which turns out to be a much more involved result than the direct part. Ahlswede and Dueck [1] succeeded in proving a weaker version of the result which they refer to as a *soft converse* where the error probabilities are forced to vanish exponentially with the blocklength. The method of proof in [1] requires fairly delicate arguments and specialized combinatorial results on coloring hypergraphs.

The main result in this paper is the proof of the *strong converse* for DMC's, i.e., for any fixed pair of probabilities of missed identification  $\lambda_1$  and false identification  $\lambda_2$ , the  $(\lambda_1, \lambda_2)$ -identification capacity is upper bounded by the Shannon capacity. As argued by Wolfowitz [3] strong converse results are important in order to strengthen the significance of coding theorems; but much of our contribution lies in the method of proof. Our arguments are entirely probabilistic, and even though the formalism of the method of types popularized by Csiszár and Körner [4] is used, no combinatorial results are invoked beyond the elementary upper bounds

Manuscript received November 19, 1990; revised June 12, 1991. This work was supported by the U.S. Office of Naval Research under Grant No. N00014-90-J-1734. This work was presented at the IEEE International Symposium on Information Theory, Budapest, Hungary, June 24–28, 1991.

T. S. Han is with the Department of Information Systems, Senshu University, Higashimito 2-1-1 Tama-ku, Kawasaki 214, Japan.

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544.

IEEE Log Number 9102765.

on the number of different conditional and unconditional types. The central technical result shows that unconditional distributions on the space of output codewords are accurately approximated by replacing the input distribution with a uniform distribution on a collection of codewords whose cardinality is equal to the number of codewords that can be reliably transmitted through the channel. In order to prove that result we introduce a novel canonical decomposition of any DMC into *equitype* channels which map codewords of common type to codewords of common type.

We believe that it is important for the future development of the theory of identification via channels to investigate whether this theory can provide performance limits in meaningful communication problems of practical interest. The second contribution of this paper is a variant of identification coding that provides a natural setting to an engineering problem of potential practical importance in multiuser communication. This problem, *identification plus transmission*, is, perhaps, mathematically more natural than the original identification problem, as it lends itself to a general coding theorem that admits an elementary proof for any noisy channel (not necessarily memoryless). In identification plus transmission, a central station wishes to transmit  $B$  bits of information reliably through a noisy channel to one of  $N$  terminals whose identity is not predetermined. Every terminal listens to the channel, decides whether it is indeed the intended recipient of the message, and if so, it decodes the message sent by the transmitter. The straightforward strategy of encoding the address and the message separately with an identification and a transmission code, respectively, requires asymptotically  $(\log \log N + B)/C$  channel symbols, where  $C$  is the channel capacity. This performance is dramatically improved by the class of Identification + Transmission codes introduced here, which brings down the required number of channel symbols to  $\max \{\log \log N, B\}/C$ .

The paper is organized as follows. Section II contains the main definitions pertaining to identification codes and a brief discussion of the achievability results obtained with both maximal and average versions of the reliability measures (1) and (2). The discussion in Section II assumes completely arbitrary single-user channels without feedback. It is interesting to note that in identification coding theorems, the easier part, provable in full generality, is the direct part, whereas in Shannon coding theorems it is the (weak) converse part that is easy to prove in complete generality. Section III is devoted to the proof of the strong converse to the identification coding theorem. Section IV introduces the problem of identification plus transmission coding and gives a simple coding theorem which shows that the aforementioned required blocklength is optimal for any noisy channel with finite input alphabet. Sections III and IV are independent and can be read accordingly.

## II. DEFINITIONS AND ACHIEVABILITY RESULTS

This preliminary section summarizes the basic definitions of achievable rates and identification codes, and discusses the direct part of the identification coding theorem for *arbitrary* channels under both maximal and average probability con-

straints. Let  $A$  and  $B$  be the input and output alphabets of a noisy channel, and let  $W^{(n)}(\cdot | x^n)$  be the conditional distribution on  $B^n$ , given that the input is equal to the codeword  $x^n \in A^n$ .

The Ahlswede–Dueck identification codes are defined as the following.

*Definition:* An  $(n, N, \lambda_1, \lambda_2)$  ID code is a collection  $\{(Q_a, D_a), a = 1, \dots, N\}$  such that

- 1)  $Q_a$  is a probability distribution (PD) on  $A^n$ ,
- 2)  $D_a \subset B^n$ ,
- 3)  $Q_a W^{(n)}(D_a) \geq 1 - \lambda_1$ ,  $a = 1, \dots, N$ ,
- 4)  $Q_a W^{(n)}(D_b) \leq \lambda_2$ , for all  $a \neq b$ ,

where we have used the notation for unconditional output distributions:

$$QW(D) = \int W(D|x) dQ(x).$$

The collection  $\{Q_a, a = 1, \dots, N\}$  will be referred to as the *codebook*, and each of its elements will be referred to as a *codistribution* (as the counterparts to the *codewords* in a channel transmission code are now distributions on the set of codewords or blocks of symbols sent through the channel).

The *rate* of an  $(n, N, \lambda_1, \lambda_2)$  ID code is defined as<sup>1</sup>  $\frac{1}{n} \log \log N$ . The counterpart to the standard definitions [4, p. 101] of achievable rates and capacity of transmission codes is given by the following definition.

*Definition:*  $R$  is a  $(\lambda_1, \lambda_2)$ -achievable ID rate if for every  $\gamma > 0$  and for all sufficiently large  $n$ , there exist  $(n, N, \lambda_1, \lambda_2)$  ID codes whose rates satisfy

$$\frac{1}{n} \log \log N > R - \gamma.$$

The supremum of the  $(\lambda_1, \lambda_2)$ -achievable ID rates is called the  $(\lambda_1, \lambda_2)$ -ID capacity of the channel.

The direct (positive) part of the identification coding theorem holds for arbitrary channels.

*Theorem 1:* For any arbitrary channel  $\{W^{(n)}: A^n \rightarrow B^n\}_{n=1}^\infty$ , its  $\lambda$ -capacity<sup>2</sup> is a  $(\lambda_1, \lambda_2)$ -achievable ID rate, provided that  $0 < \lambda \leq \lambda_1$  and  $\lambda < \lambda_2$ .

*Proof:* A minor extension of [1, Theorem 1a)] (also a corollary to Theorem 3 in Section III).  $\square$

If either probability of error is allowed to be so large that  $\lambda_1 + \lambda_2 \geq 1$ , it is possible to sharpen Theorem 1 because then the  $(\lambda_1, \lambda_2)$ -ID capacity is infinite. To see this let the decoder include each message in the output list with probability  $\lambda_2$  (independent of everything else including the received codeword). Then, regardless of  $N$ , we get a  $(n, N, 1 - \lambda_2, \lambda_2)$  ID code. The proof of the optimality of the lower bound in Theorem 1 when  $\lambda_1 + \lambda_2 < 1$  is the purpose of Section III.

In the definition of ID code, the Constraints 3) and 4) on the probabilities of missed and false identification are placed

<sup>1</sup> The basis of all logarithms and exponentials is identical and arbitrary.  
<sup>2</sup> In the maximal error probability sense.

on each individual codeword and pair of codewords, respectively. This kind of error probability criterion is the counterpart of the maximal error probability criterion used in conjunction with conventional transmission codes. In most information theoretic problems (notable exceptions include the multiple-access channel and the arbitrarily varying channel), channel capacity is invariant to whether the error probability is defined on an average or maximal sense. It is of interest to check whether this property holds for identification coding by investigating the effect of averaging the probabilities of missed and false identification. If all  $N$  codistributions are *a priori* equiprobable, the averaged versions of Conditions 3) and 4) become

$$\begin{aligned} 3a) \quad & \frac{1}{N} \sum_{a=1}^N Q_a W^{(n)}(D_a) \geq 1 - \lambda_1, \\ 4a) \quad & \frac{1}{N-1} \sum_{a:a \neq b} Q_a W^{(n)}(D_b) \leq \lambda_2, \quad b = 1, \dots, N. \end{aligned}$$

When Conditions 3) and 4a) are in effect, we shall denote the alternative meaning of the false identification probability by a subscript  $AF$  in the notation of ID codes and achievable ID rates. Analogously, a subscript  $AM$  will indicate that Conditions 3a) and 4) are in effect. We will not deal separately with the case where both conditions 3a) and 4a) are in effect as the behavior of ID capacity in that case is clear from the following result.

**Proposition 1:** For any arbitrary channel  $\{W^{(n)}: A^n \rightarrow B^n\}_{n=1}^\infty$ , if the  $(\lambda_1, \lambda_2)$ -ID capacity is nonzero, then the  $(\lambda_1, \lambda_2)_{AF}$ -ID capacity is infinite, for all  $\lambda_2 < \lambda_1$ .

**Proof:** For any  $L \geq 1$  and any  $(n, N, \lambda_1, \lambda_2)$  ID code  $\{(Q_a, D_a), a = 1, \dots, N\}$ , define the following  $(n, NL, \bar{\lambda}_1, \bar{\lambda}_2)_{AF}$ -ID code,

$$(Q_{a+lN}, D_{a+lN}) = (Q_a, D_a), \quad a = 1, \dots, N, \\ l = 0, \dots, L-1,$$

where  $\bar{\lambda}_1 = \lambda_1$  because the constraint on the missed identification probability remains intact. In order to evaluate  $\bar{\lambda}_2$  consider the average false identification probability in 4a); for every  $b = 1, \dots, N, j = 0, \dots, L-1$ ,

$$\begin{aligned} & \frac{1}{NL-1} \sum_{(a,l):(a,l) \neq (b,j)} Q_{a+lN} W^{(n)}(D_{b+jN}) \\ &= \frac{1}{NL-1} \sum_{l=0}^{L-1} \sum_{a:a \neq b} Q_{a+lN} W^{(n)}(D_{b+jN}) \\ & \quad + \frac{1}{NL-1} \sum_{l:l \neq j} Q_{b+lN} W^{(n)}(D_{b+jN}) \\ & \leq \frac{L}{NL-1} \sum_{a:a \neq b} Q_a W^{(n)}(D_b) + \frac{L-1}{NL-1} \\ & \leq \frac{NL-L}{NL-1} \lambda_2 + \frac{L-1}{NL-1} \\ & \leq \lambda_2 + \frac{2}{N}. \end{aligned}$$

Since the  $(\lambda_1, \lambda_2)$ -ID capacity is nonzero, we can find a sequence of  $(n, N, \lambda_1, \lambda_2)$ -ID codes, where  $N$  goes to infinity. Therefore, we can take  $\bar{\lambda}_2 = \lambda_2 > \lambda_2$ . This con-

cludes the proof of the result because  $L$  can be arbitrarily large.  $\square$

The probability in 4a) is the probability that message  $b$  is falsely identified (when all messages are equiprobable). Proposition 1 reveals that a trivially large amount of overlapping between decoding sets is allowed no matter how small the constraint in 4a).

Even if the constraint is forced to go to zero exponentially fast with the blocklength, it is possible to use Theorem 2a) in [1] to achieve the conclusion of Proposition 1.

In contrast, replacing the maximal constraint of the probability of missed identification 3) by the corresponding average constraint 3a) has no effect on the achievable ID rates.

**Proposition 2:** For any arbitrary channel  $\{W^{(n)}: A^n \rightarrow B^n\}_{n=1}^\infty$ , a  $(\lambda_1, \lambda_2)_{AM}$ -achievable ID rate is also a  $(\lambda_1, \lambda_2)$ -achievable ID rate if  $\lambda_1 < \lambda_2$ .

**Proof:** Let  $R$  be a  $(\lambda_1, \lambda_2)_{AM}$ -achievable ID rate. Fix arbitrary  $\gamma > 0$  and  $\delta > 0$  and choose a sequence of  $(n, N, \lambda_1, \lambda_2)_{AM}$ -ID codes such that for sufficiently large  $n$

$$\frac{1}{n} \log \log N > R - \gamma.$$

Remove from this ID code the  $N(1 - \exp(-\delta n))$  messages with the largest  $Q_a W^{(n)}(D_a^c)$ . The remaining  $N \exp(-\delta n)$  codistributions and decoding sets form a subcode that satisfies

$$Q_b W^{(n)}(D_b^c) \leq \lambda_1 / (1 - \exp(-\delta n)),$$

because, otherwise,

$$\frac{1}{N} \sum_{a=1}^N Q_a W^{(n)}(D_a) > \lambda_1.$$

Therefore, for all sufficiently large  $n$ , the resulting code satisfies

$$Q_b W^{(n)}(D_b) \geq 1 - \lambda_1$$

and its rate is lower bounded by

$$\frac{1}{n} \log \log (N \exp(-\delta n)) \geq R - 2\gamma.$$

Consequently,  $R$  is a  $(\lambda_1, \lambda_2)$ -achievable ID rate.  $\square$

Motivated by the results in Propositions 1 and 2, the original maximal error probability criterion is adopted in Section III.

### III. THE STRONG CONVERSE TO THE IDENTIFICATION CODING THEOREM

With the nontrivial choice of probabilities of missed and false identification  $\lambda_1 + \lambda_2 < 1$ , the objective is to show the converse identification coding theorem, namely, that Theorem 1 cannot be improved. Such a result has been shown by Ahlswede and Dueck [1] for discrete-memoryless channels in a so-called *soft converse* version (a weaker form than the converse to the Shannon theorem), in which  $\lambda_1$  and  $\lambda_2$  are

required to vanish exponentially with the blocklength. The main contribution of this paper is the strong converse in Theorem 2 which, together with Theorem 1 and the direct part of the Shannon theorem for DMC's (i.e., that  $\max_P I(P, W)$  is  $\lambda$ -achievable for  $0 < \lambda < 1$ ), implies that the  $(\lambda_1, \lambda_2)$ -ID capacity of a DMC is equal to its Shannon capacity if  $0 < \lambda_1 + \lambda_2 < 1$ .

**Theorem 2:** Consider a discrete memoryless channel with transition probability matrix  $W: A \rightarrow B$ . If  $\lambda_1 + \lambda_2 < 1$ , then the  $(\lambda_1, \lambda_2)$ -ID capacity of the channel is upperbounded by  $C = \max_P I(P, W)$ .

*Proof:* The proof is divided into two major parts. In the first, we show that a comparatively narrow class of ID codes is essentially optimal, leading to an upper bound to the maximum size of any code with such a structure that satisfies  $\lambda_1 + \lambda_2 < 1$ . The second part is devoted to the proof of Lemma 1, a result that may be of independent interest, and that shows that unconditional distributions on the space of output codewords are accurately approximated by replacing the input distribution with a uniform distribution on a collection of codewords whose cardinality is equal to the number of codewords that can be reliably transmitted through the channel.

*Part I:* We will adopt the following terminology and notation, some of which is standard in the method of types [4].

**Definition:** For any positive integer  $\kappa$  and finite set  $\Omega$ , a PD  $Q$  defined on  $\Omega$  is said to be  $\kappa$ -type if

$$Q(\omega) \in \left\{ 0, \frac{1}{\kappa}, \frac{2}{\kappa}, \dots, 1 \right\}, \quad \text{for all } \omega \in \Omega.$$

The number of different  $\kappa$ -types on  $\Omega$  is upperbounded by both  $|\Omega|^\kappa$  and  $(\kappa + 1)^{|\Omega|}$  [4, p. 29]. (Both bounds will be used in the sequel, depending on the relative size of the specific  $\Omega$  and  $\kappa$ .) In the frequent case where  $\Omega = A$  and  $\kappa = n$ , we will simply refer to  $Q$  as a *type*, and we denote the set of types by  $\Gamma$  and its cardinality by  $K$ .

The set of elements in  $A^n$  whose empirical distribution is equal to type  $P$  is denoted by  $T_P$ . The restriction of any distribution  $Q$  on  $A^n$  to  $T_P$  is denoted by

$$Q^P(x^n) = \begin{cases} Q(x^n)/Q(T_P), & \text{if } x^n \in T_P, \\ 0, & \text{otherwise.} \end{cases}$$

**Definition:** An ID code  $\{(Q_a, D_a), a = 1, \dots, N\}$  such that for every  $P \in \Gamma$

$$Q_1(T_P) = \dots = Q_N(T_P)$$

is called *homogeneous*.

The following result demonstrates that asymptotically there is no penalty in either rate or performance by restricting attention to homogeneous ID codes.

**Proposition 3:** For every  $(n, N, \lambda_1, \lambda_2)$  ID code,  $\delta > 0$ ,  $\lambda_1 > \lambda_1$ ,  $\lambda_2 > \lambda_2$ , and all sufficiently large  $n$ , there exists a homogeneous  $(n, N \exp(-\delta n(n+1)^{|A|}), \lambda_1, \lambda_2)$  ID code.

*Proof:* Let us partition the collection of  $N$  codistributions in the original code  $\{(Q_a, D_a), a = 1, \dots, N\}$  according to the equivalence relationship  $Q_a \leftrightarrow Q_b$ , if and only if for every  $P \in \Gamma$ , there exists an integer  $i_P \in \{0, \dots, \lfloor \exp(n\delta/2) \rfloor\}$  such that both  $Q_a(T_P)$  and  $Q_b(T_P)$  belong to the interval

$$[i_P \exp(-n\delta/2), (i_P + 1) \exp(-n\delta/2)].$$

Let  $E_\delta$  be the largest equivalence class (it is immaterial how to break ties in that choice), and find the smallest  $\bar{a} \in 1, \dots, N$  such that  $Q_{\bar{a}} \in E_\delta$ . For each codistribution  $Q_b \in E_\delta$ , we derive a new codistribution via

$$\hat{Q}_b(x^n) = Q_{\bar{a}}(T_P) Q_b^P(x^n), \quad \text{if } x^n \in T_P,$$

where we notice that, for all  $Q_b \in E_\delta$  and  $P \in \Gamma$ ,

$$Q_b(T_P) = Q_{\bar{a}}(T_P) \pm \exp(-n\delta/2).$$

Consider the homogeneous ID code  $\{(\hat{Q}_b, D_b): Q_b \in E_\delta\}$ . In order to estimate its size  $|E_\delta|$  note that the number of equivalence classes is equal to  $\lfloor \exp(n\delta/2) \rfloor^K$ . Therefore,

$$\begin{aligned} |E_\delta| &\geq N / \lfloor \exp(n\delta/2) \rfloor^K \\ &\geq N \exp(-nK\delta) \\ &\geq N \exp(-n\delta(n+1)^{|A|}). \end{aligned}$$

In order to estimate the error probability of the new ID code, take an arbitrary  $D \subset B^n$ :

$$\begin{aligned} \hat{Q}_b W^n(D) &= \sum_{P \in \Gamma} \sum_{x^n \in T_P} W^n(D | x^n) \hat{Q}_b(x^n) \\ &= \sum_{P \in \Gamma} \sum_{x^n \in T_P} W^n(D | x^n) Q_{\bar{a}}(T_P) Q_b^P(x^n) \\ &= \sum_{P \in \Gamma} \sum_{x^n \in T_P} W^n(D | x^n) Q_b(T_P) Q_b^P(x^n) \\ &\quad \pm \sum_{P \in \Gamma} \exp(-n\delta/2) \\ &= Q_b W^n(D) \pm (n+1)^K \exp(-n\delta/2), \end{aligned}$$

where the error term is smaller than  $\min\{\lambda_1 - \lambda_1, \lambda_2 - \lambda_2\}$  for sufficiently large  $n$ , and the proposition is proved.  $\square$

**Definition:** An ID code such that, for every  $P \in \Gamma$  and  $a = 1, \dots, N$ ,  $Q_a^P$  is  $M$ -type is called  *$M$ -regular*.

A homogeneous code places some structure on the weight of each type, but puts no restrictions on the codistributions restricted to each type. The opposite holds for an  $M$ -regular code. When both restrictions are imposed simultaneously we can prove an upper bound on the size of the code.

**Proposition 4:** A homogeneous  $M$ -regular  $(n, N, \lambda_1, \lambda_2)$  ID code such that  $\lambda_1 + \lambda_2 < 1$  satisfies

$$\log N \leq n(n+1)^{|A|} M \log |A|.$$

*Proof:* If  $\lambda_1 + \lambda_2 < 1$ , then all  $N$  codistributions must be different; otherwise, say  $Q_a = Q_b$ ,

$$\lambda_2 \geq Q_a W^n(D_b) = Q_b W^n(D_b) \geq 1 - \lambda_1.$$

Since the number of different  $M$ -types on  $A^n$  is upperbounded by  $|A|^{nM}$ , and the number of different types is  $K$ ,

the number of different codistributions is upperbounded by

$$N \leq |A|^{nMK},$$

and the proposition follows from  $K \leq (n+1)^{|A|}$ .  $\square$

Note that in the foregoing discussion  $M$  is allowed to grow with  $n$ , and, therefore, Proposition 4 states that asymptotically the rate of a homogeneous  $M$ -regular ID code with  $\lambda_1 + \lambda_2 < 1$  cannot be larger than  $\frac{1}{n} \log M$ . Consequently,

Theorem 2 would be proved if we could find for every ID code a homogeneous  $M$ -regular ID code with roughly the same size and error probabilities and with  $M$  growing as  $\exp(nC)$ . Such a result is formalized as follows.

**Proposition 5:** For every homogeneous  $(n, N, \lambda_1, \lambda_2)$  ID code,  $\lambda'_1 > \lambda_1$ ,  $\lambda'_2 > \lambda_2$ ,  $\gamma > 0$ , and for all sufficiently large  $n$  there exists a homogeneous  $\exp(nC + \gamma)$ -regular  $(n, N, \lambda'_1, \lambda'_2)$  ID code.

*Proof:* The new ID code is obtained by modifying the original code  $\{(Q_a, D_a), a = 1, \dots, N\}$  as follows: The decoding sets remain unchanged and the new codistributions are

$$\tilde{Q}_a(x^n) = Q_a(T_P) \tilde{Q}_a^P(x^n), \quad \text{if } x^n \in T_P,$$

where  $\tilde{Q}_a^P$  is an  $\exp(nC + n\gamma)$ -type obtained from  $Q_a^P$  via the approximation in Lemma 1 (in Part II) chosen with  $\delta = \rho^{-1}(\gamma)$  and

$$\epsilon < \min \left\{ \frac{\lambda'_2}{\lambda_2} - 1, \frac{\lambda'_1 - \lambda_1}{1 - \lambda_1} \right\}.$$

The resulting ID code is homogeneous,  $\exp(nC + n\gamma)$ -regular and has the same size as the original code. Its error probabilities satisfy, for every  $a \neq b$ ,

$$\begin{aligned} \tilde{Q}_a W^n(D_b) &= \sum_{P \in \Gamma} Q_a(T_P) \tilde{Q}_a^P W^n(D_b) \\ &\leq \exp(-n\delta) \\ &\quad + \sum_{P \in \Gamma} Q_a(T_P) (1 + \epsilon) (1 - \exp(-n\delta))^{-1} \\ &\quad \cdot Q_a^P W^n(D_b) \\ &= (1 + \epsilon) (1 - \exp(-n\delta))^{-1} Q_a W^n(D_b) \\ &\quad + \exp(-n\delta) \\ &\leq (1 + \epsilon) (1 - \exp(-n\delta))^{-1} \lambda_2 + \exp(-n\delta) \\ &\leq \lambda'_2, \end{aligned}$$

where the last inequality holds for all sufficiently large  $n$  because of the choice of  $\epsilon$ . Analogously,

$$\begin{aligned} \tilde{Q}_a W^n(D_a) &= \sum_{P \in \Gamma} Q_a(T_P) \tilde{Q}_a^P W^n(D_a) \\ &\geq (1 - \epsilon) (1 - \exp(-n\delta)) Q_a W^n(D_a) \\ &\quad - \exp(-n\delta) \\ &\geq (1 - \epsilon) (1 - \exp(-n\delta)) (1 - \lambda_1) \\ &\quad - \exp(-n\delta) \\ &\geq 1 - \lambda'_1, \end{aligned}$$

for all sufficiently large  $n$ , concluding the proof of Proposition 5.  $\square$

It is now straightforward to see how Propositions 3, 4, and 5 lead to the proof of Theorem 2. Starting with arbitrary  $\lambda_1, \lambda_2$  that satisfy  $\lambda_1 + \lambda_2 < 1$  and an arbitrary sequence of  $(n, N, \lambda_1, \lambda_2)$  ID codes, we chose arbitrary  $\delta > 0$ ,  $\gamma > 0$ ,  $\lambda'_1 > \lambda_1 > \lambda_1$ , and  $\lambda'_2 > \lambda_2 > \lambda_2$  such that  $\lambda'_1 + \lambda'_2 < 1$  and  $\lambda'_1 + \lambda'_2 < 1$ . Proposition 3 guarantees the existence, for all sufficiently large  $n$ , of a homogeneous  $(n, N, \lambda'_1, \lambda'_2)$  ID code with

$$N' = N \exp(-\delta n(n+1)^{|A|}).$$

Moreover, Proposition 5 guarantees the existence of homogeneous  $\exp(nC + n\gamma)$ -regular  $(n, N', \lambda'_1, \lambda'_2)$  ID codes. Since  $\lambda'_1 + \lambda'_2 < 1$ , Proposition 4 requires that

$$\begin{aligned} \log N' &= \log N - \delta n(n+1)^{|A|} \\ &\leq n(n+1)^{|A|} M \log |A|. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\log \log N}{n} &\leq \frac{\log n(n+1)^{|A|}}{n} \\ &\quad + \frac{1}{n} \log \{ \exp(nC + n\gamma) \log |A| + \delta \} \\ &\leq \gamma + \frac{1}{n} \log \log |A| + C + 2\gamma \\ &\leq C + 4\gamma, \end{aligned}$$

where the second and third inequalities hold for all sufficiently large  $n$ . Since  $\gamma > 0$  was chosen arbitrarily, the conclusion is that the  $(\lambda_1, \lambda_2)$ -ID capacity is upperbounded by  $C$ .  $\square$

**Part II:** The second part of the proof of the strong converse is entirely devoted to the proof of Lemma 1, a result which was invoked in the proof of Proposition 5 and which is not specific to identification or to channel coding. Any PD can be approximated by an  $M$ -type PD provided that  $M$  is large enough. The question is how large  $M$  need be. If we are actually interested in approximating the unconditional output distribution by approximating the input distribution, and the input distribution puts mass on only one type, then Lemma 1 tells us that  $M$  need not grow faster than the number of different inputs that can be reliably discriminated at the output.

**Lemma 1:** For every  $P \in \Gamma$ ,  $\epsilon \in [0, \epsilon_0]$ ,  $\delta \in [0, \delta_0]$ ,  $PD$   $Q$  defined on  $T_P$  and all sufficiently large  $n$ , there exists an  $\exp(nC + n\gamma)$ -type distribution  $\tilde{Q}$  defined on  $T_P$  such that, for every  $D \subset B^n$ ,

$$\begin{aligned} \tilde{Q} W^n(D) &\leq (1 + \epsilon) (1 - \exp(-n\delta))^{-1} \\ &\quad \cdot Q W^n(D) + \exp(-n\delta) \quad (3) \end{aligned}$$

$$\begin{aligned} \tilde{Q} W^n(D) &\geq (1 - \epsilon) (1 - \exp(-n\delta)) \\ &\quad \cdot Q W^n(D) - \exp(-n\delta), \quad (4) \end{aligned}$$

where  $\gamma = \rho(\delta)$ , and  $\rho: [0, \delta_0] \rightarrow R^+$  is a continuous strictly increasing function such that  $\rho(0) = 0$ .

*Proof:*

*Step 1) Canonical Decomposition into Equitype Channels:* For any stochastic matrix  $V: A \rightarrow B$ , the set of output sequences with conditional type  $V$  given  $x^n \in A^n$  will be denoted by  $T_V(x^n)$  (cf. [4, p. 31]). Note that  $|T_V(x^n)|$  is identical for all  $x^n \in T_P$ ; so we can use the notation

$$L_V^P = |T_V(x^n)|, \quad \text{if } x^n \in T_P.$$

Let us define for all types  $P$  and conditional types  $V$

$$W_V^P(y^n | x^n) = \begin{cases} 1/L_V^P, & \text{if } x^n \in T_P \text{ and } y^n \in T_V(x^n), \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

$W_V^P$  is a stochastic matrix<sup>3</sup>  $T_P \rightarrow T_{PV}$  with identical nonzero entries for all  $V$  and  $P$  such that  $L_V^P > 0$ . Such a stochastic matrix will be referred to as an *equitype channel* as it only connects inputs of a common type to outputs of a common type. Each sequence  $x^n$  has a unique type; however the conditional type of  $y^n$  given  $x^n$  is not uniquely determined (if  $x^n$  does not contain every letter in  $A$ ). In order to find a unique representation of input/output pairs in terms of unconditional/conditional types it is convenient to define the set of conditional types congruent with  $P$  by

$$\Lambda^P = \{V: L_V^P > 0 \text{ and } V(\cdot | x) = W(\cdot | x) \text{ if } P(x) = 0\},$$

where in cases where the set  $T_V(x^n)$  coincides for several  $V$  we have chosen a specific representative stochastic matrix.

Since  $W^n(y^n | x^n)$  depends on its arguments only through the type of  $x^n$  and the conditional type of  $y^n$  given  $x^n$ , we may denote

$$c_V^P = W^n(T_V(x^n) | x^n), \quad \text{for any } x^n \in T_P.$$

Furthermore, for every  $x^n \in T_P$  and  $y^n \in T_V(x^n)$ , we may write

$$\begin{aligned} W^n(y^n | x^n) &= c_V^P / L_V^P \\ &= c_V^P W_V^P(y^n | x^n) \end{aligned} \quad (6)$$

and, since for every  $(x^n, y^n)$  there is only one element in  $\{(P, V): V \in \Lambda^P, P \in \Gamma\}$  such that  $x^n \in T_P$  and  $y^n \in T_V(x^n)$ , we are able to conclude from (5) and (6) that

$$W^n(y^n | x^n) = \sum_{P \in \Gamma} \sum_{V \in \Lambda^P} c_V^P W_V^P(y^n | x^n) \quad (7)$$

with

$$\sum_{V \in \Lambda^P} c_V^P = 1, \quad \text{for every } P \in \Gamma, \quad (8)$$

which we will refer to as the *canonical decomposition* of an arbitrary DMC into *equitype channels*. Note that  $W_V^P$  does not depend on the transition matrix  $W$ ; the information contained in  $W^n$  is summarized by the collection of distributions  $\{c_V^P, V \in \Lambda^P\}$  for all  $P \in \Gamma$ .

The main benefit of introducing this canonical decomposition is that, as far as approximating the output distributions,

<sup>3</sup> In a slight abuse of notation we denote by  $T_{PV}$  the subset of output sequences in  $B^n$  with unconditional type  $PV$ .

we may focus attention on each equitype channel separately, and then simply take the average of the resulting approximations under (7). Because the equitype channels are highly structured (and are independent of  $W$ ), they lend themselves to general results.

*Step 2) Estimating the Probability of Inverse Images:* The subset of input sequences connected to a specific  $y^n$  by the equitype channel  $W_V^P$  is denoted by

$$H_V^P(y^n) = \{x^n \in T_P, W_V^P(y^n | x^n) > 0\} \quad (9)$$

and is informally referred to as the *inverse image* of  $y^n$ .

It is important to estimate  $Q(H_V^P(y^n))$  in connexion with the approximation of the unconditional output distribution because

$$QW_V^P(y^n) = Q(H_V^P(y^n)) / L_V^P. \quad (10)$$

Such an estimation is carried out in the following result.

*Lemma 2:* Define for every  $P \in \Gamma$ ,  $V \in \Lambda^P$ , and  $\delta > 0$ ,

$$G_V^P = \{y^n \in B^n: Q(H_V^P(y^n)) \geq \exp(-nI(P, V) - n\delta)\}.$$

Then, for every blocklength  $n$ ,

$$QW_V^P(G_V^P) \geq 1 - \exp(-n\delta)(n+1)^{|A|+|B|}. \quad (11)$$

*Proof:* Define

$$\begin{aligned} F_V^P &= \{y^n \in T_{PV}: Q(H_V^P(y^n)) / L_V^P \\ &> (n+1)^{|A|+|B|} \exp(-n\delta) / |T_{PV}|\}. \end{aligned}$$

Using (10), the definition in (12) and the fact that  $P[\{\omega \in \Omega: P[\{\omega\}] \leq \mu | \Omega\}^{-1}] \leq \mu$ , particularized to  $\Omega = T_{PV}$ ,  $\mu = (n+1)^{|A|+|B|} \exp(-n\delta)$ , we obtain

$$QW_V^P(F_V^P) \geq 1 - (n+1)^{|A|+|B|} \exp(-n\delta).$$

Finally, note that  $F_V^P \subset G_V^P$  because if  $x^n \in T_P$  then [4, p. 40]

$$\begin{aligned} (n+1)^{-|A|+|B|} \exp(-nI(P, V)) &\leq |T_V(x^n)| / |T_{PV}| \\ &= L_V^P / |T_{PV}|. \end{aligned} \quad (13)$$

□

Using the fact that (13) is tight (in the exponential rate of growth with  $n$ ) it can be shown that for most output sequences  $y^n$ ,  $Q(H_V^P(y^n)) \sim \exp(-nI(P, V))$ . Therefore, with high probability the empirical distribution of a sample of  $M = \exp(nI(P, V) + n\delta)$  values of  $x^n$  generated under  $Q$  will be an  $M$ -type distribution approximating  $Q$  as desired. However, this falls short of showing Lemma 1 because such an empirical distribution depends on  $y^n$ ,  $P$ , and  $V$ , and there is an exponential number of such choices. The way around this hurdle will be to show that the set of empirical distributions that do not offer the desired degree of approximation is doubly exponentially small for every  $y^n$ ,  $P$ , and  $V$ . Another difficulty is created by the fact that each equitype channel requires a different  $M(\sim \exp(nI(P, V)))$  for the approximation by  $M$ -type distributions. Supremizing this over all possible equitype channels would result in  $M \sim \min$

$\{|A|^n, |B|^n\}$ , which is above the required  $M \sim \exp(nC)$  unless the channel is noiseless. This problem is created by the fact that some of the equitype channels are much less "noisy" than  $W^n$ , and therefore are not representative of the average. The elimination from consideration of those atypical equitype channels is the purpose of the next step.

*Step 3) Channel Clipping:* The canonical decomposition into equitype channels allows us to easily modify the channel in order to suit our needs. Essentially, we only retain those equitype channels which are close to the original DMC in the sense

$$\Lambda_\delta^P = \{V \in \Lambda^P : D(V \| W | P) \leq \delta\}. \quad (14)$$

To that end, we define a new stochastic matrix  $W_\delta^* : A^n \rightarrow B^n$  by the canonical decomposition

$$W_\delta^*(y^n | x^n) = \sum_{P \in \Gamma} \sum_{V \in \Lambda^P} \bar{c}_V^P W_V^P(y^n | x^n), \quad (15)$$

with

$$\bar{c}_V^P = \begin{cases} c_V^P \left[ \sum_{U \in \Lambda_\delta^P} c_U^P \right]^{-1}, & \text{if } V \in \Lambda_\delta^P, \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

Since by clipping the channel in this way we have only eliminated rare transitions,  $W_\delta^*$  approximates  $W^n$  closely.

*Lemma 3:* For every  $n$ ,  $x^n \in A^n$ ,  $D \subset B^n$ , and  $\delta > 0$ ,

$$\begin{aligned} W^n(D | x^n) &\geq (1 - \exp(-n\delta)(n+1)^{|A||B|}) W_\delta^*(D | x^n), \quad (17) \\ W^n(D | x^n) &\leq W_\delta^*(D | x^n) + \exp(-n\delta)(n+1)^{|A||B|}. \quad (18) \end{aligned}$$

*Proof:* Fix  $x^n \in A^n$ , and let its empirical distribution be  $P$ . Denote

$$\alpha_\delta^P = \sum_{U \in \Lambda_\delta^P} c_U^P. \quad (19)$$

Then (7) implies that

$$W^n(D | x^n) = \sum_{V \in \Lambda^P} c_V^P W_V^P(D | x^n), \quad (20)$$

$$W_\delta^*(D | x^n) = \sum_{V \in \Lambda^P} \bar{c}_V^P W_V^P(D | x^n); \quad (21)$$

but, since  $\alpha_\delta^P \bar{c}_V^P \leq c_V^P$ , we obtain

$$\alpha_\delta^P W_\delta^*(D | x^n) \leq W^n(D | x^n). \quad (22)$$

Moreover,

$$\begin{aligned} W^n(D | x^n) &\leq \sum_{V \in \Lambda_\delta^P} \bar{c}_V^P W_V^P(D | x^n) \\ &\quad + \sum_{V \in \Lambda^P - \Lambda_\delta^P} c_V^P W_V^P(D | x^n) \\ &\leq W_\delta^*(D | x^n) + 1 - \alpha_\delta^P. \end{aligned}$$

In order to conclude the proof of Lemma 3, consider for every  $P \in \Gamma$

$$\begin{aligned} 1 - \alpha_\delta^P &= \sum_{V \in \Lambda^P - \Lambda_\delta^P} c_V^P \\ &= \sum_{V \in \Lambda^P - \Lambda_\delta^P} W^n(T_V(x^n) | x^n) \\ &\leq \sum_{V \in \Lambda^P - \Lambda_\delta^P} \exp(-nD(V \| W | P)) \\ &\leq (n+1)^{|A||B|} \exp(-n\delta), \end{aligned}$$

where the first inequality follows from [4, p. 32] and the second inequality results from (14) and  $|\Lambda^P| \leq (n+1)^{|A||B|}$ .  $\square$

It is easy to check that Lemma 3 reduces the proof of Lemma 1 to the verification of the existence of an  $\exp(nC + n\gamma)$ -type  $\tilde{Q}$  that satisfies, for all sufficiently large  $n$ ,

$$\tilde{Q} W_\delta^*(D) \leq (1 + \epsilon) Q W_\delta^*(D) + \exp(-n\delta), \quad (23)$$

$$\tilde{Q} W_\delta^*(D) \geq (1 - \epsilon) Q W_\delta^*(D) - \exp(-n\delta). \quad (24)$$

In order to ascertain that fact, use (17), (18), and (23) to show that for all sufficiently large  $n$  and every  $\delta' < \delta$ ,

$$\begin{aligned} \tilde{Q} W^n(D) &\leq (1 + \epsilon)(1 - \exp(-n\delta)(n+1)^{|A||B|})^{-1} \\ &\quad \cdot Q W^n(D) + \exp(-n\delta)(n+2)^{|A||B|} \\ &\leq (1 + \epsilon)(1 - \exp(-n\delta'))^{-1} Q W^n(D) \\ &\quad + \exp(-n\delta'). \quad (25) \end{aligned}$$

Conversely, (17), (18), and (24) lead to

$$\begin{aligned} \tilde{Q} W^n(D) &\geq (1 - \epsilon)(1 - \exp(-n\delta)(n+1)^{|A||B|}) Q W^n(D) \\ &\quad - (1 - \epsilon) \exp(-n\delta)(n-1)^{|A||B|} - \exp(-n\delta) \\ &\geq (1 - \epsilon)(1 - \exp(-n\delta')) Q W^n(D) - \exp(-n\delta'). \quad (26) \end{aligned}$$

But since in the statement of Lemma 1 the choice of  $\delta$  is arbitrary and the function  $\rho$  is continuous,  $\delta'$  may be substituted by  $\delta$  in (25) and (26). The remainder of the proof is devoted to showing the approximations (23) and (24) for the clipped channel.

*Step 4) Required Finesse of Approximations for the Clipped Channel:* We have anticipated that the finesse of the approximation required for equitype channel  $W_V^P$  is  $M \sim \exp(nI(P, V))$ . After clipping the channel, we can upper bound  $I(P, V)$  in terms of the channel capacity because of the following result.

*Lemma 4:* If  $D(V \| W | P)^{1/2} < \min\{\frac{1}{8} \log e, 1\}$ , then

$$\begin{aligned} |I(P, V) - I(P, W)| &\leq 2g(D(V \| W | P)) \\ &\quad + D(V \| W | P)^{1/2} \log |B|, \quad (27) \end{aligned}$$

where  $g$  is the concave function

$$g(x) = \begin{cases} \sqrt{\frac{2x}{\log e}} \log \frac{1}{\sqrt{\frac{2x}{\log e}}}, & x > 0, \\ 0, & x = 0. \end{cases}$$

*Proof:* We can upper bound the left-hand side in (27) by

$$|I(P, V) - I(P, W)| \leq |H(PV) - H(PW)| + |H(V|P) - H(W|P)|. \quad (28)$$

From [4, Lemma 1.2.7 and Problem 1.3.17] we have for any arbitrary probability distributions  $Q, R$  such that  $D(Q\|R) \leq \log(e^{1/8})$

$$|H(Q) - H(R)| \leq g(D(Q\|R)). \quad (29)$$

Let  $L = \{x: D(V(\cdot|x)\|W(\cdot|x)) \leq D(V\|W|P)^{1/2}\}$  by the reverse Markov inequality:  $P[L^c] \leq D(V\|W|P)^{1/2}$ ;

$$\begin{aligned} & |H(V|P) - H(W|P)| \\ & \leq \sum_x P(x) |H(V(\cdot|x)) - H(W(\cdot|x))| \\ & \leq \sum_{x \in L} P(x) g(D(V(\cdot|x)\|W(\cdot|x))) \\ & \quad + \sum_{x \in L^c} P(x) \log |B| \\ & \leq \sum_x P(x) g(D(V(\cdot|x)\|W(\cdot|x))) \\ & \quad + D(V\|W|P)^{1/2} \log |B| \\ & \leq g(D(V\|W|P)) + D(V\|W|P)^{1/2} \log |B|, \end{aligned}$$

where the last inequality is a consequence of the concavity of  $g$ . Furthermore, we can apply (29) to the case  $Q = PV$ ,  $R = PW$  because

$$\begin{aligned} D(PV\|PW) & \leq D(V\|W|P) \\ & \leq D(V\|W|P)^{1/2} \leq \frac{1}{8} \log e, \end{aligned}$$

thus yielding the sought-after result upon use of (28) and (30).  $\square$

Letting

$$\rho(\delta) = 2\delta + 2g(\delta) + \sqrt{\delta} \log |B|, \quad (31)$$

we see that there is an interval  $(0, \delta_0)$  on which  $\rho$  is continuous and strictly increasing. If  $V \in \Lambda_\delta^P$ , then

$$I(P, V) + \delta \leq I(P, W) + \rho(\delta).$$

Therefore,

$$\sup_{P \in \Gamma} \sup_{V \in \Lambda_\delta^P} I(P, V) + \delta \leq C + \rho(\delta). \quad (32)$$

Henceforth, we will let  $\gamma = \rho(\rho)$  and

$$M = \exp(nC + n\gamma).$$

Note that, if  $y^n \in G_V^P$  and  $V \in \Lambda_\delta^P$ , (32) and the definition of  $G_V^P$  imply

$$Q(H_V^P(y^n)) \geq \frac{1}{M} \exp(n\delta). \quad (33)$$

*Step 5) The  $M$ -type Approximating Distribution  $\tilde{Q}$ :*

The distribution  $\tilde{Q}$  will be chosen as the empirical distribution of a realization  $(\tilde{u}_1, \dots, \tilde{u}_M)$  of the i.i.d. random variables  $(U_1, \dots, U_M)$  with common distribution  $Q$ , chosen according to the following result.

*Lemma 5:* There exist  $\tilde{u}_i \in T_P$ ,  $i = 1, \dots, M$ , such that for all  $V \in \Lambda_\delta^P$

$$\frac{1}{M} \sum_{i=1}^M 1\{\tilde{u}_i \in H_V^P(y^n)\} \leq (1 + \epsilon)Q(H_V^P(y^n)), \quad (34)$$

for every  $y^n \in G_V^P$

$$\frac{1}{M} \sum_{i=1}^M 1\{\tilde{u}_i \in H_V^P(y^n)\} \geq (1 - \epsilon)Q(H_V^P(y^n)), \quad (35)$$

for every  $y^n \in G_V^P$

$$\frac{1}{M} \sum_{i=1}^M W_V^P((G_V^P)^c | \tilde{u}_i) \leq \exp\left(-\frac{n\delta}{3}\right). \quad (36)$$

*Proof:* First note that it follows from Lemma 2 that, if  $\delta' < \delta$ ,

$$\begin{aligned} & P\left[\frac{1}{M} \sum_{i=1}^M U_i W_V^P((G_V^P)^c) > \exp\left(-\frac{n\delta'}{2}\right)\right] \\ & < \exp\left(-\frac{n\delta'}{2}\right). \end{aligned} \quad (37)$$

Thus,

$$\begin{aligned} & P\left[\frac{1}{M} \sum_{i=1}^M U_i W_V^P((G_V^P)^c) > \exp\left(-\frac{n\delta'}{2}\right) \text{ for some } V \in \Lambda_\delta^P\right] \\ & < (n+1)^{|A|+|B|} \exp\left(-\frac{n\delta'}{2}\right) \\ & \leq \exp\left(-\frac{n\delta}{3}\right), \end{aligned} \quad (38)$$

for sufficiently large  $n$ . Next we show that

$$\begin{aligned} & P\left[\frac{1}{M} \sum_{i=1}^M 1\{U_i \in H_V^P(y^n)\} > (1 + \epsilon)Q(H_V^P(y^n))\right] \\ & \leq \exp_e\left(-\frac{\epsilon^2}{3} \exp(n\delta)\right), \end{aligned} \quad (39)$$

$$\begin{aligned} & P\left[\frac{1}{M} \sum_{i=1}^M 1\{U_i \in H_V^P(y^n)\} < (1 - \epsilon)Q(H_V^P(y^n))\right] \\ & \leq \exp_e\left(-\frac{\epsilon^2}{3} \exp(n\delta)\right). \end{aligned} \quad (40)$$

by using the following corollary to Sanov's lemma [5, p. 117].  $\square$

*Lemma 6:* Let  $(Z_1, \dots, Z_M)$  be a Bernoulli sequence taking values in  $\{0, 1\}$  with  $P[Z_i = 1] = \mu$ . Then, for every  $0 < \epsilon < 1$ ,

$$\begin{aligned} & P\left[\frac{1}{M} \sum_{i=1}^M Z_i > (1 + \epsilon)\mu\right] \\ & \leq \exp(-MD((1 + \epsilon)\mu\|\mu)), \end{aligned}$$

$$P\left[\frac{1}{M}\sum_{i=1}^M Z_i < (1-\epsilon)\mu\right] \leq \exp(-MD((1-\epsilon)\mu\|\mu)),$$

where  $D(\alpha\|\beta)$  denotes the divergence between the distributions  $(\alpha, 1-\alpha)$  and  $(\beta, 1-\beta)$ .

Now we apply Lemma 6 to the case where

$$Z_i = 1\{U_i \in H_V^P(y^n)\}, \\ \mu = E[Z_i] = Q(H_V^P(y^n)),$$

and we use the following fact.

*Lemma 7:* There exists  $\epsilon_0 > 0$  such that if  $-\epsilon_0 \leq t \leq \delta_0$ , then

$$D(\mu + t\mu\|\mu) \geq \frac{1}{2}t^2\mu \log e.$$

*Proof:* Let us assume the nontrivial case  $\mu > 0$ . Applying the Kullback-Leibler result on the local quadratic behavior of divergence [4, p. 27], we get

$$\lim_{t \rightarrow 0} \frac{1}{t^2} D(\mu + t\mu\|\mu) = \frac{1}{2} \log e I_0(\mu),$$

where  $I_0(\mu)$  is equal to the Fisher information of the family of binary distributions  $(\mu + t\mu, 1 - \mu - t\mu)$  that is readily computed as

$$I_0(\mu) = \frac{\mu}{(1+t)(1-\mu-t\mu)}.$$

Therefore,  $\mu > 0$  implies that  $I_0(\mu) > \mu$  and Lemma 7 follows.  $\square$

Lemma 6 and 7 upper bound the probabilities in (39) and (40) by

$$\exp\left(-M\epsilon^2 \frac{\mu}{3}\right) \leq \exp_e\left(-\frac{\epsilon^2}{2} \exp(n\delta)\right), \quad (41)$$

where the inequality follows from (33). This shows (39) and (40) for  $0 < \epsilon < \epsilon_0$ . Now, using the union bound, we obtain

$$P\left[\frac{1}{M}\sum_{i=1}^M 1\{U_i \in H_V^P(y^n)\} > (1+\epsilon)Q(H_V^P(y^n)) \text{ for some } y^n \in G_V^P \text{ and } V \in \Lambda_\delta^P\right] \\ \leq |G_V^P| |\Lambda_\delta^P| \exp_e\left(-\frac{\epsilon^2}{2} \exp(n\delta)\right) \\ \leq |B|^n (n+1)^{|A|+|B|} \exp_e\left(-\frac{\epsilon^2}{2} \exp(n\delta)\right) \\ \leq \exp_e\left(-\frac{\delta^2}{3} \exp(n\delta)\right) \quad (42)$$

for sufficiently large  $n$ , and analogously

$$P\left[\frac{1}{M}\sum_{i=1}^M 1\{U_i \in H_V^P(y^n)\} < (1-\epsilon)Q(H_V^P(y^n)) \text{ for some } y^n \in G_V^P \text{ and } V \in \Lambda_\delta^P\right] \\ \leq \exp_e\left(-\frac{\epsilon^2}{2} \exp(n\delta)\right). \quad (43)$$

Since the sum of the three upper bounds in (38), (42), and (43) is strictly less than 1, for sufficiently large  $n$ , we must conclude that the probability of the union of the three events is strictly less than 1, and, therefore, there exists a realization  $(\tilde{u}_1, \dots, \tilde{u}_M)$  of  $(U_1, \dots, U_M)$  with the sought-after properties.  $\square$

*Step 6) Approximation of  $QW_\delta^*$  by  $\tilde{Q}W_\delta^*$ :* In this final step, we show that Lemma 5 is sufficient to show the approximation formulas (23) and (24). For every  $y^n \in G_V^P$ , we can write (cf. (10))

$$\tilde{Q}W_V^P(y^n) = \tilde{Q}(H_V^P(y^n))/L_V^P \\ \leq (1+\epsilon)Q(H_V^P(y^n))/L_V^P \\ = (1+\epsilon)QW_V^P(y^n), \quad (44)$$

where the inequality follows from (34) because by definition of  $\tilde{Q}$ , for any  $T \subset A^n$ ,

$$\tilde{Q}(T) = \frac{1}{M} \sum_{i=1}^M 1\{\tilde{u}_i \in T\}.$$

Now, for any arbitrary  $D \subset B^n$ , we can write for every  $V \in \Lambda_\delta^P$

$$\tilde{Q}W_V^P(D) = \tilde{Q}W_V^P(D \cap G_V^P) + \tilde{Q}W_V^P(D \cap (G_V^P)^c) \\ \leq (1+\epsilon)QW_V^P(D) + \tilde{Q}W_V^P((G_V^P)^c) \\ \leq (1+\epsilon)QW_V^P(D) + \exp(-n\delta/3). \quad (45)$$

Proceeding similarly, we get

$$(1-\epsilon)QW_V^P(D) \leq \tilde{Q}W_V^P(D) + (1-\epsilon)QW_V^P((G_V^P)^c) \\ \leq \tilde{Q}W_V^P(D) + \exp(-n\delta/3), \quad (46)$$

where the second inequality follows from (11). Averaging (45) and (46) with respect to  $\{\tilde{c}_V^P, V \in \Lambda^P\}$ , and replacing  $\delta$  by  $\delta/3$  in the definition of  $W_\delta^*$ , we obtain the sought-after approximation formulas (23) and (24). This concludes the proof of Theorem 2.  $\square$

Lemma 1 is the germ of a very general result recently obtained by the authors: The restriction to discrete memoryless channels can be dropped in Lemma 1 and arbitrary input distributions  $Q$  can be allowed. Arbitrarily accurate approximation of the output finite-dimensional distributions is achieved in the sense of variational distance. Such a result (and its converse) are proved in [6] in a way which is completely different from the proof of Lemma 1. Those results form the basis of an approximation theory of output statistics that finds a variety of applications in information theory.

#### IV. IDENTIFICATION PLUS TRANSMISSION

In this section, we present a technical application of the theory of identification via noisy channels that provides a natural setting to an engineering problem of potential practical importance in multiuser communication. An additional advantage of this new communication problem is that it lends

itself to a completely general coding theorem whose proof is much easier than that of the identification coding theorem.

Consider the communication problem depicted in Fig. 1, where there is one transmitter and  $N$  receivers. The transmitter wishes to transmit information reliably to *one* of the receivers, whose identity is not predetermined. Every potential receiver listens to the noisy channel and must decide whether it is indeed the intended recipient of the message, and if so, it decodes the message sent by the transmitter. This is a very common situation in practical applications such as local-area networks, radio networks, and downlink satellite communication where a common broadcast channel is provided in order for the central station to communicate with the terminals. The central station is required to deliver a sequence of messages, each intended for one of the terminals. If the various messages are generated at different times, a conceptually sensible strategy is to decouple the transmission of different messages and consider every one of them in isolation. A more general problem, which we do not consider here, would involve the *simultaneous* transmission of a collection of messages (and their respective addresses) to their intended receivers.

The straightforward solution to the problem posed here is to encode the address of the destination  $a \in \{1, \dots, N\}$  in a header followed by a codeword representing the message  $m \in \{1, \dots, M\}$ . If  $\alpha n$  and  $(1 - \alpha)n$  symbols are devoted to the transmission of address and message, respectively, then the Shannon theorem ensures that reliable communication is possible if  $M$  and  $N$  grow with  $n$  as

$$\frac{1}{n} \log N \rightarrow \alpha C, \quad (47)$$

$$\frac{1}{n} \log M \rightarrow (1 - \alpha)C, \quad (48)$$

but not faster, where  $C$  is the capacity of the channel. In many applications  $N$  is negligible with respect to  $M$ , and the header is devoted to a very small fraction  $\alpha$  of the transmitted symbols, thereby achieving a growth of  $M$  that is essentially given by  $\exp(nC)$ . Somewhat surprisingly, it turns out that such a rate of information transmission can be sustained even if the amount of information contained in the address is not negligible with respect to the information contained in the message. In fact, it is possible to transmit information at the channel capacity rate as long as the number of bits in the address does not exceed  $M$ .

It is indeed possible to do better than the straightforward strategy of using a separate transmission code for address and message. To see this, first note that each station is interested in finding out whether it is the intended recipient or not; if it is not, then it is not interested in estimating which of the other stations is the intended recipient. Naturally, we immediately recognize from this that an identification code can be used in order to transmit the address. Therefore, Theorem 1 implies that we can transmit a number of addresses that grows as

$$\frac{1}{n} \log \log N \rightarrow \alpha C. \quad (49)$$

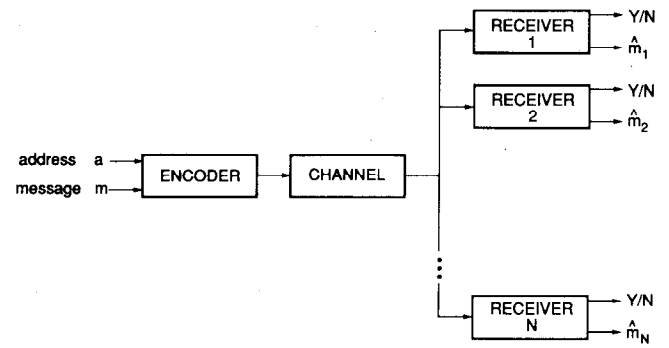


Fig. 1. Identification plus transmission through a noisy channel.

We see that the strategy of juxtaposing an identification code and a transmission code to send address and message respectively, achieves (48) and (49). Therefore, we can achieve

$$\frac{1}{n} \log \log N \rightarrow C, \quad (50)$$

if the message transmission rate goes to zero, and

$$\frac{1}{n} \log M \rightarrow C, \quad (51)$$

if the address identification rate goes to zero. However, it is actually possible to achieve (50) and (51) *simultaneously* by using an Identification + Transmission (IT) code (defined below) where, unlike the aforementioned, address and message are not encoded separately. A decoupled coding strategy is far from optimum because reliable transmission of the message is only required by the intended receiver. Actually, in certain applications, a privacy feature may be desirable whereby any other receiver is unable to decode the transmitted message reliably.

*Definition:* An  $(n, N, M, \lambda_1, \lambda_2)$  IT code is a mapping  $f: \{1, \dots, N\} \times \{1, \dots, M\} \rightarrow A^n$  and a collection of subsets  $\{D_{a,m} \subset B^n, a \in \{1, \dots, N\}, m \in \{1, \dots, M\}\}$  such that for all  $a = 1, \dots, N$ ,

- 1)  $D_{a,m} \cap D_{a,l} = \emptyset, \quad \text{if } l \neq m,$
- 2)  $\frac{1}{M} \sum_{m=1}^M W^{(n)}(D_{a,m} | f(a, m)) \geq 1 - \lambda_1,$
- 3)  $\frac{1}{M} \sum_{m=1}^M W^{(n)}(D_b | f(a, m)) \leq \lambda_2, \quad \text{if } b \neq a,$

where we have used the notation  $D_a \triangleq \bigcup_{m=1}^M D_{a,m}$ .

The *rate-pair* of an  $(n, N, M, \lambda_1, \lambda_2)$  IT code is  $(\frac{1}{n} \log M, \frac{1}{n} \log \log N)$ .

Upon reception of the channel output  $y^n$ , station  $a$  checks whether  $y^n \in D_a$ . If  $y^n \notin D_a$ , then the station declares that it is not the intended recipient of the message. If  $y^n \in D_a$ , then station  $a$  searches for the unique (cf. Condition 1)  $m \in \{1, \dots, M\}$  such that  $y^n \in D_{a,m}$  and outputs message  $m$ . Let us now address the reliability afforded by the IT code. First, the reliability with which the address is received is

equal to the reliability achievable had it been separately encoded with an ID code:

$$\begin{aligned} & P[\text{Station } a \text{ declares } Y \mid a \text{ transmitted}] \\ &= \sum_{m=1}^M \frac{1}{M} W^{(n)}(D_a \mid (a, m) \text{ transmitted}) \\ &\geq \sum_{m=1}^M \frac{1}{M} W^{(n)}(D_{a,m} \mid (a, m) \text{ transmitted}) \geq 1 - \lambda_1, \end{aligned}$$

where we have used Condition 2 and the fact that the message  $m \in \{1, \dots, M\}$  is chosen equiprobably and independently of the choice of address. (Needless to say, the *contents* or meaning of the messages need not coincide from station to station.) Similarly, if  $b \neq a$ ,

$$\begin{aligned} & P[\text{Station } b \text{ declares } Y \mid a \text{ transmitted}] \\ &= \sum_{m=1}^M \frac{1}{M} W^{(n)}(D_b \mid f(a, m)) \leq \lambda_2, \end{aligned}$$

because of Condition 3. Regarding the reliability with which the message is received, Condition 2 states that the intended recipient decodes the correct message with average probability of error (over the set of equiprobable messages) better than  $\lambda_1$ . Note that for every  $a \in \{1, \dots, N\}$ ,  $\{(f(a, m), D_{a,m}) \in A^n \times 2^{B^n}, m = 1, \dots, M\}$  is an  $(n, M, \lambda_1)$  transmission code (in the average probability of error sense). The IT code puts no constraints on the reliability with which unintended stations decode the message. Since it is always possible to randomize the labeling of messages for each address (i.e., we can apply a different permutation to  $\{f(a, m), D_{a,m}, m = 1, \dots, M\}$  for every  $a \in \{1, \dots, N\}$ ), no information about a message transmitted to Station  $a$  is obtained by a station that does not have access to Station  $a$ 's decoder:  $\{D_{a,m}, m = 1, \dots, M\}$ . Therefore, it is feasible to incorporate the aforementioned privacy feature into an IT code.

Let us now turn our attention to characterize the set of achievable rate-pairs of IT codes. We say that  $(R, \bar{R})$  is a  $(\lambda_1, \lambda_2)$ -achievable IT rate-pair if for every  $\gamma > 0$  and for all sufficiently large  $n$ , there exist  $(n, N, M, \lambda_1, \lambda_2)$  IT codes such that

$$\frac{1}{n} \log \log N > \bar{R} - \gamma, \quad (52)$$

$$\frac{1}{n} \log M > R - \gamma. \quad (53)$$

We will show a very general result which states that for any finite-input channel  $\{W^{(n)}: A^n \rightarrow B^n\}_{n=1}^{\infty}$  (not necessarily memoryless) with capacity  $C$ ,  $(C, C)$  is an *optimal* IT rate-pair in the sense that any other  $(\lambda_1, \lambda_2)$ -achievable rate pair  $(R, \bar{R})$  cannot achieve either  $\bar{R} > C$  or  $R > C$ . The direct part of this result can be proved essentially in the same way as the direct part of the identification coding theorem [1, Theorem 1a].

**Theorem 3:** For any arbitrary channel  $\{W^{(n)}: A^n \rightarrow B^n\}_{n=1}^{\infty}$ , denote its  $\lambda$ -capacity by  $C$ . Then  $(C, C)$  is a  $(\lambda_1, \lambda_2)$ -achievable IT rate pair, provided that  $0 < \lambda \leq \lambda_1$  and  $\lambda < \lambda_2$ .

*Proof:* We will construct IT codes from arbitrary transmission codes using Proposition 1 in [1] which is rephrased for our purposes as follows.

**Lemma 8:** There exists a function  $\rho: (0, 1) \rightarrow (1, +\infty)$  such that for every integer  $S$  and  $0 < \mu < 1$ , there exists an  $N \times M$  matrix  $s(a, m) \in \{1, \dots, S\}$ , which satisfies (with

$$A_i \triangleq \bigcup_{m=1}^M \{s(i, m)\})$$

- 1)  $|A_i| = M$ ,
- 2)  $|A_i \cap A_j| < \mu M$ , if  $i \neq j$ ,
- 3)  $M = \lfloor S \log_2 \rho(\mu) \rfloor$ ,
- 4)  $N > \frac{1}{S} [\rho(\mu)^S - 1]$ .

For every  $(n, S, \lambda)$  transmission code  $\{(\phi(s), E_s) \in A^n \times 2^{B^n}, s = 1, \dots, S\}$  (where the error probability is understood in the maximal sense), we will choose an IT code using the matrix guaranteed by Lemma 8 with  $\mu = \lambda_2 - \lambda$  as follows:

$$f(a, m) = \phi(s(a, m)), \quad a = 1, \dots, N, \quad m = 1, \dots, M,$$

$$D_{a,m} = E_{s(a,m)}, \quad a = 1, \dots, N, \quad m = 1, \dots, M.$$

Because of Condition 1 in Lemma 8,  $D_{a,m} \cap D_{a,l} = \emptyset$  if  $l \neq m$ . Furthermore, for every  $(a, m)$ ,

$$\begin{aligned} W^{(n)}(D_{a,m} \mid f(a, m)) &= W^{(n)}(E_{s(a,m)} \mid \phi(s(a, m))) \\ &\geq 1 - \lambda \geq 1 - \lambda_1 \end{aligned} \quad (54)$$

and for every  $a \neq b$ ,

$$\begin{aligned} & \frac{1}{M} \sum_{m=1}^M W^{(n)}(D_b \mid f(a, m)) \\ &= \frac{1}{M} \sum_{m: s(a,m) \in A_a \cap A_b} W^{(n)}(D_b \mid f(a, m)) \\ & \quad + \frac{1}{M} \sum_{m: s(a,m) \in A_a \cap A_b^c} W^{(n)}(D_b \mid f(a, m)) \\ &\leq \frac{|A_a \cap A_b|}{M} + \frac{1}{M} \sum_{m: s(a,m) \in A_a \cap A_b^c} \\ & \quad \cdot W^{(n)}(D_b \mid f(a, m)) \\ &\leq \lambda_2 - \lambda + \frac{1}{M} \sum_{m: s(a,m) \in A_a \cap A_b^c} \\ & \quad \cdot W^{(n)}(E_{s(a,m)}^c \mid \phi(s(a, m))) \\ &\leq \lambda_2 - \lambda + \frac{|A_a|}{M} \lambda = \lambda_2, \end{aligned} \quad (55)$$

where the second inequality follows from Condition 2 and the fact that  $D_b \subset E_{s(a,m)}^c$  if  $s(a, m) \notin A_b$ . The last two relations follow from the assumed reliability of the code and Condition 1, respectively. Equations (54) and (55) show that the IT code so constructed is an  $(n, N, M, \lambda_1, \lambda_2)$  IT code. Now using Conditions 3 and 4, we get that for any  $\gamma > 0$  and sufficiently large  $n$ ,

$$\begin{aligned} \frac{1}{n} \log M &> \frac{1}{n} \log S - \frac{\gamma}{2}, \\ \frac{1}{n} \log \log N &> \frac{1}{n} \log S - \frac{\gamma}{2}. \end{aligned}$$

But, since  $C$  is the  $\lambda$ -capacity of the channel, there exist  $(n, S, \lambda)$  codes for sufficiently large  $n$  such that

$$\frac{1}{n} \log S > C - \frac{\gamma}{2}. \quad \square$$

We can view the  $N \times M$  matrix on  $\{1, \dots, S\}$  whose existence is guaranteed by Lemma 8 as describing a binary constant-weight code with blocklength  $S$ , size  $N$ , weight  $M$  and pairwise overlap bounded by  $\mu M$ . Each row of  $\{s(a, m)\}$  corresponds to a codeword in the constant-weight code whose  $M$  1's occur at the indices  $[s(a, 1), \dots, s(a, M)]$ . Then, the construction of IT codes in the proof of Theorem 3 can be viewed as the concatenation of a transmission code (inner code) and a binary constant-weight code (outer code). This viewpoint is adopted in [7] where an explicit construction of *optimal* binary constant-weight codes for identification (and identification plus transmission) is shown.

In order to show the converse result, i.e., that  $(C, C)$  is indeed optimal, we could invoke the strong converse of the Shannon coding theorem [8] and the identification coding theorem (Section III), because identification plus transmission cannot be any easier than either identification or transmission separately. However, this would enable us to prove the result for DMC's only. Rather than invoking the rather difficult result obtained in Section III, it is possible to prove the converse to identification plus transmission coding for any finite-input channel and in a most straightforward fashion.

**Theorem 4:** If  $\lambda_1 + \lambda_2 < 1$  and  $(R, \bar{R})$  is a  $(\lambda_1, \lambda_2)$ -achievable IT rate pair, then

$$R \leq C \quad (56)$$

and

$$\bar{R} \leq C, \quad (57)$$

where  $C$  is the  $\lambda_1$ -capacity<sup>4</sup> of the channel.

*Proof:* The bound in (56) readily follows from the definitions of IT code, achievable IT rate pair and  $\lambda_1$ -capacity.

<sup>4</sup> In the average probability sense.

Because  $\lambda_1 + \lambda_2 < 1$ , if  $a \neq b$ , then any  $(n, N, M, \lambda_1, \lambda_2)$  IT code satisfies  $[f(a, 1), \dots, f(a, M)] \neq [f(b, 1), \dots, f(b, M)]$ . For, otherwise,

$$\begin{aligned} \lambda_2 &\geq \frac{1}{M} \sum_{m=1}^M W^{(n)}(D_b | f(a, m)) \\ &= \frac{1}{M} \sum_{m=1}^M W^{(n)}(D_b | f(b, m)) \\ &\geq \frac{1}{M} \sum_{m=1}^M W^{(n)}(D_{b,m} | f(b, m)) \geq 1 - \lambda_1. \end{aligned}$$

Consequently, the maximum number of addresses is bounded by

$$N \leq |A^n|^M, \quad (58)$$

which implies that for any  $\gamma > 0$  and sufficiently large  $n$ ,

$$\frac{1}{n} \log \log N \leq \frac{1}{n} \log M + \gamma. \quad (59)$$

Finally, from the definitions of achievable IT rate pair we get, for sufficiently large  $n$ ,

$$\bar{R} - \gamma \leq \frac{1}{n} \log \log N \quad (60)$$

and, since for every  $a$ ,  $\{(f(a, m), D_{a,m}), m = 1, \dots, M\}$  is an  $(n, M, \lambda_1)$  code (in the average error probability sense), we have

$$\frac{1}{n} \log M \leq C + \gamma \quad (61)$$

infinitely often. Joining (59), (60), and (61), we obtain (57) since the choice of  $\gamma$  was arbitrary.  $\square$

To conclude, let us remark that the strong converse of the identification coding theorem (Theorem 2) does not follow from Theorem 4. Whereas, for example, the single user coding theorem is a corollary of the multiple-access coding theorem, Theorem 2 does not follow from the fact that  $(0, C)$  is in the boundary of the region of achievable IT rate pairs (Theorem 4) because although every  $(n, N, M, \lambda_1, \lambda_2)$  IT code (trivially) determines an  $(n, N, \lambda_1, \lambda_2)$  ID code, the reverse is not true.

#### REFERENCES

- [1] R. Ahlswede and G. Dueck, "Identification via channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 15–29, Jan. 1989.
- [2] —, "Identification in the presence of feedback—A discovery of new capacity formulas," *IEEE Trans. Inform. Theory*, vol. 35, pp. 30–36, Jan. 1989.
- [3] J. Wolfowitz, *Coding Theorems of Information Theory*, third ed. New York: Springer, 1978.
- [4] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [5] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.
- [6] T. S. Han and S. Verdú, "Approximation theory of output statistics," 1991, manuscript in preparation.
- [7] S. Verdú and V. K. Wei, "Explicit construction of optimal codes for identification via channels," presented at *1991 IEEE Int. Symp. Inform. Theory*, Budapest, Hungary, June 24–28, 1991.
- [8] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math.*, vol. 1, pp. 591–606, Dec. 1957.