Identification in the Presence of Feedback—A Discovery of New Capacity Formulas

RUDOLF AHLSWEDE AND GUNTER DUECK, MEMBER, IEEE

Abstract—**The main contribution of our earlier work, "Identification via Channels," was that $N = \exp\{\exp\{nR\}\}$ objects can be *identified* in block length n with arbitrarily small error probability via a discrete memoryless channel (DMC), if *randomization* can be used for the encoding procedure and if $R < C(W)$. Moreover, in this case the second-order identification capacity equals Shannon's transmission capacity $C(W)$, where W is the transmission matrix of the DMC. Here we study the identification problem in the presence of a *noiseless feedback channel* and determine the second-order capacity C_f (resp. C_F) for deterministic (resp. randomized) encoding strategies. We encounter several important phenomena. 1) Although feedback does not increase the transmission capacity of a DMC, it does increase the (second-order) identification capacity. We actually prove that**

$$C_f(W) = \max_{x \in \mathcal{X}} H(W(\cdot|x))$$

and

$$C_F(W) = \max_{P} H(P \cdot W), \quad \text{if} \quad C(W) > 0.$$

2) Notice that $C_f = 0$ if W is a matrix with 0 and 1 as entries only. Thus *noise increases* C_f! 3) The structure of the new capacity formulas is apparently much simpler than Shannon's familiar formula. This has the effect that proofs of converses become easier than in our previous work.

I. The Results

IN THE BEGINNING of [1], we discussed the notions of classical transmission codes and (randomized) identification codes. Since [1] appears in this issue, we refer the reader to it for definitions, and we start right away with the analogous concepts for discrete memoryless channels with feedback.

For the classical transmission problem an (n, M, λ) feedback code $\{(f_j, \mathcal{D}_j) | j = 1, \cdots, M\}$ is described as follows. There is given a finite set of messages $\mathcal{M} = \{1, \cdots, M\}$. One of these messages is to be sent over the channel. Message $j \in \mathcal{M}$ is encoded by a (vector-valued) function

$$f_j = \left[f_j^1, f_j^2, \cdots, f_j^n\right] \tag{1}$$

where, for $t \in \{2, \cdots, n\}$, f_j^t is defined on $\mathcal{Y}^{t-1}$ and takes values in $\mathcal{X}$. f_j^1 is an element of $\mathcal{X}$. It is understood that after the received elements $Y_1, \cdots, Y_{t-1}$ have been made known to the sender by the feedback channel, the sender transmits $f_j^t(Y_1, \cdots, Y_{t-1})$. At $t = 1$ the sender transmits f_j^1.

The distribution of the random variables (RV's) $Y_t (t = 1, 2, \cdots, n)$ is determined by f_j and W. We denote the probability of receiving $y^n = (y_1, \cdots, y_n) \in \mathcal{Y}^n$, if j has been encoded, by $W^n(y^n | f_j) = W(y_1 | f_j^1) \cdot W(y_2 | f_j^2(y_1)) \cdots W(y_n | f_j^n(y_1 \cdots y_n))$. Again the $\mathcal{D}_j \subset \mathcal{Y}^n$ $(j = 1, \cdots, M)$ are disjoint decoding sets and we require that

$$W^n(\mathcal{D}_j | f_j) \geq 1 - \lambda, \quad \text{for} \quad j = 1, \cdots, M. \tag{2}$$

Now let $M_f(n, \lambda)$ be the maximal integer M for which an (n, M, λ) feedback code exists.

Theorem S–K–K (Shannon–Kemperman–Kesten):

$$\lim_{n \to \infty} \frac{1}{n} \log M_f(n, \lambda) = C, \quad \text{for all} \quad \lambda \in (0, 1).$$

The proof and the apportionment of the credit for it can be found in [3] and [4].

Remark 1: It is also known that randomization in the encoding or/and decoding does not increase the capacity.

Now let us turn again to the identification problem. We consider two concepts, deterministic and randomized identification-feedback (IDF) codes, and make the following important observations:

1) Even in the deterministic case, feedback causes the maximal codelength to grow doubly exponentially in block length.

2) If, in addition, we allow randomization in the encoding, this results in a further improvement to the extent that the aforementioned double exponent increases.

3) In both cases the capacities are characterized in terms of entropy measures. Mutual information, however, plays no role!

4) The formulas for the capacities show that "noise" typically increases capacity!

We now formulate the exact results. Let $\mathcal{F}_n$ be the set of all possible encoding functions of the kind defined in (1). A (deterministic) (n, N, λ) IDF code for W is a system

$$\{(f_i, \mathcal{D}_i) | i = 1, \cdots, N\} \quad \text{with} \quad f_j \in \mathcal{F}_n, \ \mathcal{D}_i \subset \mathcal{Y}^n,$$

$$\text{for} \quad i \in \{1, \cdots, N\}$$

Manuscript received June 18, 1986. This paper was presented at the XXIII General Assembly of U.R.S.I., Tel Aviv, Israel, August 24–September 1, 1987.

R. Ahlswede is with the Fakultät für Mathematik, Universität Bielefeld, Universitätsstrasse, Postfach 8640, 4800 Bielefeld 1, Germany.

G. Dueck was with the Universität Bielefeld, Bielefeld, Germany. He is now with the IBM Scientific Center Heidelberg, Tiergartenstraße 15, 6900 Heidelberg, Germany.

IEEE Log Number 8825710.

and

$$W^n(\mathcal{D}_i^c|f_i) \le \lambda \qquad W^n(\mathcal{D}_j|f_i) \le \lambda \tag{3}$$

for all $i, j \in \{1, \cdots, N\}$ with $i \neq j$. A randomized (n, N, λ) IDF code for W is a system

$$\{(Q_F(\cdot|i), \mathcal{D}_i) | i = 1, \cdots, N\}$$

with $Q_F(\cdot|i) \in \mathcal{P}(\mathcal{F}_n)$, $\mathcal{D}_i \subset \mathcal{Y}^n$, and

$$\sum_{g \in \mathcal{F}_n} Q_F(g|i) W^n(D_i^c|g) \le \lambda, \tag{4}$$

$$\sum_{g \in \mathcal{F}_n} Q_F(g|j) W^n(\mathcal{D}_i|g) \le \lambda \tag{5}$$

for all $i, j \in \{1, \cdots, N\}$ with $i \neq j$.

Let $N_f(n, \lambda)$ (resp. $N_F(n, \lambda)$) be the maximal integer N for which a deterministic (resp. randomized) (n, N, λ) IDF code exists. (We add f (resp. F) to the notation to indicate the model with which we are working).

Theorem 1 (Coding Theorem and Strong Converse): If the transmission capacity C of W is positive, then we have for all $\lambda \in (0, 1/2)$:

$$a) \quad \lim_{n \to \infty} \inf \frac{1}{n} \log\log N_f(n, \lambda) \le \max_{x \in \mathcal{X}} H(W(\cdot|x))$$

$$b) \quad \lim_{n \to \infty} \sup \frac{1}{n} \log\log N_f(n, \lambda) \le \max_{x \in \mathcal{X}} H(W(\cdot|x)).$$

In particular, for deterministic feedback strategies the second-order identification capacity $C_f(W)$ equals $\max H(W(\cdot|x))$ provided that $C(W) > 0$. $C_f(W) = 0$ if and only if $C(W) = 0$ or W is a noiseless channel, i.e., $W(y|x) \in \{0, 1\}$ for all x, y. This result says that $C_f(W)$ depends solely on the maximal per letter "output entropy" $H(W(\cdot|x^*)) = \max_{x \in \mathcal{X}} H(W(\cdot|x))$.

Also, C_f increases if $H(W(\cdot|x^*))$, "the measure of noise caused by x^*," increases. Indeed, for noiseless channels, C_f is zero.

This behavior is in surprising contrast to the familiar properties of the transmission capacity. The reader will gain a complete understanding in the course of the proof of part a) of Theorem 1; here we give some of the underlying ideas.

In [1] we showed that a large amount of randomization in the encoding is necessary to achieve a positive doubly exponential rate. In case of feedback, the sender has another way of performing a random experiment, namely, to send (possibly repeatedly) a letter x with $H(W(\cdot|x)) > 0$. Its outcome is known to the sender via the feedback link. The maximal amount of randomness is achieved if one uses a letter $x^* \in \mathcal{X}$ with

$$H(W(\cdot|x^*)) = \max_x H(W(\cdot|x)).^1$$

The proof of Theorem 1 shows that all good deterministic encoding strategies use such letters x^* most of the time. The situation here is quite different from what we are used to in classical coding problems. As a consequence there is almost no connection between the capacities C_f and C.

However, if we allow randomized feedback strategies, then by [1, theorem 1] we know that $C_F \ge C$. Actually, strict inequality holds here except for those cases which are specified in Remark 2.

Theorem 2 (Coding Theorem and Strong Converse): If the transmission capacity C of W is positive, then, for all $\lambda \in (0, 1/2)$,

$$a) \quad \lim_{n \to \infty} \inf \frac{1}{n} \log\log N_F(n, \lambda) \ge \max_{P \in \mathcal{P}(\mathcal{X})} H(P \cdot W)$$

$$b) \quad \lim_{n \to \infty} \sup \frac{1}{n} \log\log N_F(n, \lambda) \le \max_{P \in \mathcal{P}(\mathcal{X})} H(P \cdot W).$$

Remark 2: We call W essentially noiseless if there exist subsets $\mathcal{X}^* \subset \mathcal{X}$, $\mathcal{Y}^* \subset \mathcal{Y}$ and a bijection $g: \mathcal{X}^* \to \mathcal{Y}^*$ such that

$$W(g(x)|x) = 1, \qquad \text{for all} \quad x \in \mathcal{X}^* \tag{6}$$

$$W(\cdot|x') \in \text{convex hull} \{W(\cdot|x): x \in \mathcal{X}^*\}, \quad \text{for} \quad x' \in \mathcal{X}. \tag{7}$$

We claim that $C_F = C$ if and only if $C = 0$ or W is essentially noiseless.

If W is essentially noiseless and $C > 0$ then $C = \log|\mathcal{X}^*|$ and $C_F = \log|\mathcal{Y}^*| = C$. Conversely, if $C = C_F$ and $C > 0$, then for every P' satisfying

$$I(P', W) = H(P' \cdot W) - H(W|P) = C,$$

we have $H(W|P') = 0$. This and the optimality of P' imply that W is essentially noiseless.

Remark 3: We make some comments concerning the proofs. In [1] we built ID codes from large subsets of a given channel code (for transmission). In this paper (Section III), we show that [1, theorem 1a)] can be proved in another simple manner. the ID code is "combined" from two ordinary transmission channel codes. The first one has the sole purpose of *providing sender and receiver with the (common) knowledge of the outcome of a random experiment*. Its entropy per time unit determines the second-order rate of the ID code. This important observation also makes the role of feedback for identification transparent. Feedback makes it possible to provide sender and receiver with the knowledge of the outcome of other random experiments. In the deterministic case it is the experiment obtained by sending the letter x^* n times and in the case of randomized feedback strategies it is the experiment $(\mathcal{Y}^n, \prod_1^n P \cdot W)$, which can be performed by sending the outcome of $(\mathcal{X}^n, P^n)$ over the channel. Notice that $H(P \cdot W) = I(P, W) + H(W|P)$. Theorem 2 says that the doubly exponential rates $I(P, W)$ (which are achievable with randomized encoding and no feedback) and $H(W|P)$ (which is achievable with feedback and no randomization) add up to the rate $H(P \cdot W)$ (which is achievable with both feedback and randomization). We choose of course a P, which maximizes $H(P \cdot W)$. The proofs of the converses essentially say that common random experiments of higher

[1] The results explain why, to identify the state of the world in a universal philosophical system, one has to proceed as follows: first choose your position and then create a lot of noise.

per-letter entropies do not exist under the respective circumstances.

For the second code used in the proofs of the direct part, it is only essential that its rate be positive. Thus the condition $C > 0$ enters. It can be seen by inspection of the proof that, as long as $C > 0$, an infinite identification capacity can be achieved, if sender and receiver have knowledge of the outcome of the same random experiment of an infinite entropy. It is well-known that such random experiments (also with finite entropy) can be used to increase the transmission capacity of systems of channels such as arbitrarily varying channels [5]. Their effect on the identification capacity is dramatic!

II. Notation and Known Facts

For the basic notation we again refer the reader to [1, sect. I-D]. We state here only two additional simple lemmas. For channels $V, V' \in \mathcal{W}$ let

$$\|V - V'\| = \max_{x,y} |V(y|x) - V'(y|x)|.$$

Lemma 1: For every $\epsilon > 0$, there is a $\delta' = \delta'(\epsilon) > 0$ such that

$$W^n\big(\{y^n \in \mathcal{Y}^n | y^n \in \mathcal{T}_V^n(x^n) \text{ for a } V \text{ with } \|V - W\| \le \epsilon\} | x^n\big) \le 1 - 2^{-n\delta'}$$

for $n \ge n_0(\epsilon)$.

Lemma 2: For every $\epsilon > 0$ there is a $c(\epsilon) > 0$ such that for $n \ge n_0(\epsilon)$

a) $$\left| \bigcup_{V: \|V-W\| \le \epsilon} \mathcal{T}_V^n(x^n) \right| \ge 2^{n(H(W|P_{x^n}) - c(\epsilon))}$$

b) $$\left| \bigcup_{V: \|V-W\| \le \epsilon} \mathcal{T}_V^n(x^n) \right| \le 2^{n(H(W|P_{x^n}) + c(\epsilon))}$$

c) $$|\mathcal{T}_V^n(x^n)| \ge 2^{n(H(W|P_{x^n}) - c(\epsilon))}, \quad \text{if } \|V - W\| \le \epsilon \text{ and } \mathcal{T}_V^n(x^n) \ne 0,$$

and $c(\epsilon) \to 0$ if $\epsilon \to 0$.

III. A New Proof of the Direct Part in [1, Theorem 1]

The proof in [1] uses in the encoding procedure probability distributions which are uniform distributions on the sets of codewords in some classical channel codes (as defined in [1]). There is a lot of freedom in selecting systems of such codes (see Remark 4 below). Here we choose a system consisting of codes, which are extensions of a single channel code. This system is designed so that, with some modifications, it can be used for the feedback case as well. It is again produced by a random selection and allows a fairly simple analysis.

We begin with two fundamental codes $\mathcal{C}'$ and $\mathcal{C}$. By Shannon's coding theorem (stated in [1]) we know that for every $\epsilon > 0$, $\epsilon < C$, there is a $\delta = \delta(\epsilon) > 0$ and an $n_0(\epsilon)$ such that for $n \ge n_0(\epsilon)$ an $(n, M', 2^{-n\delta})$ code

$$\mathcal{C}' = \{(u_j', \mathcal{D}_j') | j = 1, \cdots, M'\} \tag{8}$$

and an $(\lceil \sqrt{n} \rceil, M'', 2^{-\sqrt{n}^{\delta}})$ code

$$\mathcal{C}'' = \{(u_k'', \mathcal{D}_k'') | k = 1, \cdots, M''\} \tag{9}$$

exist with $M' = \lfloor 2^{n(C-\epsilon)} \rfloor$ and $M'' = \lfloor 2^{\epsilon\sqrt{n}} \rfloor$.

We use the abbreviation $m = n + \lceil \sqrt{n} \rceil$. Now any family $\{T_i | i = 1, \cdots, N\}$ of maps

$$T_i: \{1, \cdots, M'\} \to \{1, \cdots, M''\}$$

can be used to build an ID code $\{(Q(\cdot|i), \mathcal{D}_i) | i = 1, 2, \cdots, N\}$ from $\mathcal{C}'$ and $\mathcal{C}''$. Here $Q(\cdot|i)$ is the uniform distribution on the set of codewords

$$\mathcal{U}_i = \{u_j' \cdot u_{T_i(j)}'' | j = 1, \cdots, M'\} \subset \mathcal{X}^m$$

and

$$\mathcal{D}_i = \bigcup_{j=1}^{M'} \mathcal{D}_j' \times \mathcal{D}_{T_i(j)}''.$$

We choose at random an ID code of such a structure in the following way.

For $i \in \{1, 2, \cdots, N\}$ and $j \in \{1, \cdots, M'\}$ let U_{ij} be independent RV's such that U_{ij} takes the value $u_j' \cdot u_k''$ with probability $1/M''$ for $k \in \{1, \cdots, M''\}$. We consider the random sets

$$\bar{\mathcal{U}}_i = \{U_{i1}, \cdots, U_{iM'}\} \quad \text{for } i = 1, \cdots, N. \tag{10}$$

The uniform distributions $\bar{Q}(\cdot|i)$ on these sets become random distributions. The random decoding sets are

$$\mathcal{D}(\bar{\mathcal{U}}_i) = \bigcup_{j=1}^{M'} \mathcal{D}(U_{ij}) \tag{11}$$

where

$$\mathcal{D}(U_{ij}) = \mathcal{D}_j' \times \mathcal{D}_k'', \quad \text{if } U_{ij} = u_j' \cdot u_k''. \tag{12}$$

We now analyze the maximal error performances of $\{(\bar{Q}(\cdot|i), \mathcal{D}(\bar{\mathcal{U}}_i)) | i = 1, \cdots, N\}$. It is clear from the definitions (8)–(12) that for every realization $\mathcal{U}_i$ of $\bar{\mathcal{U}}_i$

$$\frac{1}{M'} \sum_{u \in \mathcal{U}_i} W^n\big(\mathcal{D}(\mathcal{U}_i)^c | u\big) \le 2^{-n\delta} + 2^{-\sqrt{n}\delta}. \tag{13}$$

Thus only errors of the second kind remain to be considered. For this analysis we again use a large deviational approach to bound the probability that there does not exist a realization with a prescribed error of the second kind λ for two indices, without loss of generality say $i = 1, 2$. That bound yields the final result for all indices i since the probability for the union of events does not exceed the sum of the probabilities of these events. Actually, it suffices to compare the random set $\bar{\mathcal{U}}_2$ with any realization $\mathcal{U}_1$ of $\bar{\mathcal{U}}_1$. Fix $\mathcal{U}_1$ and define for $j = 1, \cdots, M'$

$$\psi_j = \psi_j(\bar{\mathcal{U}}_2) = \begin{cases} 1, & \text{if } \mathcal{U}_{2j} \in \mathcal{U}_1 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Since the random variables U_{2j} are independent, $\psi_1, \cdots, \psi_{M'}$ are also independent. Furthermore, by our definitions

$$\mathbb{E}\psi_j = \frac{1}{M''} \quad \text{for } j = 1, \cdots, M'. \tag{15}$$

An elementary calculation shows that, for $M'' = \lfloor 2^{\sqrt{n}\epsilon} \rfloor$,

$$D(\lambda \| 1/M'') \geq \lambda \cdot \sqrt{n} \cdot \epsilon - 1. \tag{16}$$

Therefore, [1, lemma LD] implies the following.

Corollary 1: For $\lambda \in (0,1)$ and $1/M'' < \lambda$

$$\Pr\left(\sum_{j=1}^{M'} \psi_j > M' \cdot \lambda \right) \leq 2^{-M'(\lambda\sqrt{n}\epsilon - 1)}.$$

We need one other elementary fact. Suppose that $\bar{\mathcal{U}}_2 = \mathcal{U}_2$ and $u \in \mathcal{U}_1 - \mathcal{U}_2$; then

$$W^m(\mathcal{D}(\mathcal{U}_2)|u) \leq 2^{-n\delta} + 2^{-\sqrt{n}\delta}. \tag{17}$$

To see this, let $u = u_j' \cdot u_k''$. Notice that for $u \notin \mathcal{U}_2$, $\mathcal{D}(\mathcal{U}_2) \cap (\mathcal{D}_j' \times \mathcal{D}_k'') = \emptyset$ and that therefore

$$W^m(\mathcal{D}(\mathcal{U}_2)|u) \leq W^m\left((\mathcal{D}_j' \times \mathcal{D}_k'')^c | u \right).$$

Equation (17) follows, because the definitions (8) and (9) imply that

$$W^m\left((\mathcal{D}_j' \times \mathcal{D}_k'')^c | u \right) \leq 2^{-n\delta} + 2^{-\sqrt{n}\delta}.$$

An upper bound on the error of the second kind is now readily established:

$$\begin{aligned} &\sum_{u \in \mathcal{U}_1} W^m(\mathcal{D}(\bar{\mathcal{U}}_2)|u) \\ &\quad = \sum_{u \in \mathcal{U}_1 \cap \mathcal{U}_2} W^m(\mathcal{D}(\bar{\mathcal{U}}_2)|u) \\ &\qquad + \sum_{u \in \mathcal{U}_1 - \bar{\mathcal{U}}_2} W^m(\mathcal{D}(\bar{\mathcal{U}}_2)|u) \\ &\quad \leq |\mathcal{U}_1 \cap \bar{\mathcal{U}}_2| + |\mathcal{U}_1 - \bar{\mathcal{U}}_2| \cdot (2^{-n\delta} + 2^{-\sqrt{n}\delta}), \end{aligned}$$

where we have used (17). Since $|\mathcal{U}_1 \cap \bar{\mathcal{U}}_2| = \sum_{j=1}^{M'} \psi_j(\bar{\mathcal{U}}_2)$,

$$\frac{1}{M'} \sum_{u \in \mathcal{U}_1} W^m(\mathcal{D}(\bar{\mathcal{U}}_2)|u) \leq \frac{1}{M'} \sum_{j=1}^{M'} \psi_j(\bar{\mathcal{U}}_2) + 2 \cdot 2^{-\sqrt{n}\delta}. \tag{18}$$

Now fix $\lambda \in (0,1)$. By Corollary 1 for large n we have that with *positive* probability

$$\frac{1}{M'} \sum_{u \in \mathcal{U}_1} W^m(\mathcal{D}(\bar{\mathcal{U}}_2)|u) \leq \lambda + 2 \cdot 2^{-\sqrt{n}\delta} \tag{19}$$

and similarly

$$\frac{1}{M'} \sum_{u \in \bar{\mathcal{U}}_2} W^m(\mathcal{D}(\mathcal{U}_1)|u) \leq \lambda + 2 \cdot 2^{-\sqrt{n}\delta}. \tag{20}$$

Hence there is a realization $\bar{\mathcal{U}}_2 = \mathcal{U}_2$ for which (19) and (20) hold. We use this argument repeatedly for $i = 3,4,\cdots,N$ (as in [1]). An $(n, N, \lambda + 2 \cdot 2^{-\sqrt{n}\delta})$ ID code exists, if

$$(N-1)\Pr\left(\sum_{j=1}^{M'} \psi_j > M'\lambda \right) < 1. \tag{21}$$

From Corollary 1 and $M' = \lfloor 2^{n(C-\epsilon)} \rfloor$ (21) holds for every N with

$$N \leq 2^{1/2 \cdot (\lambda\sqrt{n}\epsilon - 1)2^{n(C-\epsilon)}}.$$

This proves the result.

Remark 4: Instead of extending the code $\mathcal{C}'$, one can prove the same result by making a random selection of subcodes of $\mathcal{C}'$ whose lengths are small but proportional to $|\mathcal{C}'|$.

IV. Proof of the Direct Part of Theorem 1

We know already from [1] that randomization in the encoding causes $N(n,\lambda)$ to grow doubly exponentially in n. In the preceding proof we gained additional insight. The amount of "correlated randomization," that is, the size of a random experiment, whose outcomes are known to the sender and to the receiver (with very small error probability), is the decisive quantity determining the growth of $N(n,\lambda)$.

As our random experiment we used the uniform distribution on the set of codewords of the code $\mathcal{C}'$. The outcome $u_j' \in \{u_1', \cdots, u_{M'}'\}$ is known to the sender. Then the outcome is transmitted over the channel and made known to the receiver with high probability. The parameter $M' = \lfloor 2^{n(C-\epsilon)} \rfloor$ is the size of this random experiment.

The presence of feedback allows the design of another random experiment. Feedback is used here solely for this purpose. Otherwise the coding scheme is essentially the same as previously. We now describe this random experiment and the coding scheme. Let $x^* \in \mathcal{X}$ be a letter with

$$H(W(\cdot|x^*)) = \max_{x \in \mathcal{X}} H(W(\cdot|x)). \tag{22}$$

Choose again as total block length

$$m = n + \lceil \sqrt{n} \rceil \tag{23}$$

and define $\mathcal{C}''$ as in (9). We now describe the substitute for $\mathcal{C}'$.

Regardless which object $i \in \{1,\cdots,N\}$ is presented to the sender, he first sends $x^{*n} = (x^*,\cdots,x^*) \in \mathcal{X}^n$. The received sequence $y^n \in \mathcal{Y}^n$ becomes known to the sender by the feedback channel.

The resulting correlated random experiment $(\mathcal{Y}^n, W^n(\cdot|x^{*n}))$ needs a modification, because $W^n(\cdot|x^{*n})$ is far from being uniform on $\mathcal{Y}^n$. However, $W^n(\cdot|x^{*n})$ is essentially uniform on the set

$$\mathcal{D}^* = \bigcup_{V:\, \|V - W\| \leq \epsilon} \mathcal{T}_V^n(x^{*n}), \tag{24}$$

which carries essentially all its probability. We lump the small-probability set $\mathcal{Y}^n - \mathcal{D}^*$ together in an erasure symbol e with the understanding that

$$W^n(e|x^{*n}) = W^n(\mathcal{Y}^n - \mathcal{D}^*|x^{*n}). \tag{25}$$

We choose as our random experiment $(\mathcal{D}^* \cup \{e\}, W^n(\cdot|x^{*n}))$. The price paid for more uniformity is a small error probability, if e occurs. However, previously we still had to deal in $\mathcal{C}'$ with small error probabilities. By Lemma 2, $|\mathcal{D}^*| \sim 2^{nH(W(\cdot|x^*))}$, and this quantity now takes the role of M'.

Instead of the maps T_i: $\{1,\cdots,M'\} \to \{1,\cdots,M''\}$, we now use maps F_i: $\mathcal{D}^* \to \{1,\cdots,M''\}$ in the block $[n+$

$1,\cdots,m]$. This means that after $y^n \in \mathscr{D}^*$ has been received, the sender sends $\mu''_{F_i(y^n)}$, if $i \in \{1,2,\cdots,N\}$ is given to him.

In case $y^n \notin \mathscr{D}^*$ an error is declared and the sender can fill the $\lceil n^{1/2} \rceil$ positions in any way, for instance by sending $x^*\lceil n^{1/2} \rceil$ times again. Clearly, for each F_1, we have defined an encoding function $f_i \in \mathscr{F}_m$ as introduced in Section I. For the decoding we define the sets

$$\mathscr{D}(F_i) = \bigcup_{y^n \in \mathscr{D}^*} \{y^n\} \times \mathscr{D}''_{F_i(y^n)}, \qquad \text{for } i = 1,\cdots,N. \quad (26)$$

The astute reader can avoid the following formal analysis, which is necessary only because our random experiment is not exactly uniform.

With respect to the error of the first kind notice that $W^m(\mathscr{D}(F_i)^c|f_i) \le W^n((\mathscr{D}^*)^c|x^{*n}) + 2^{-\sqrt{n}\delta}$, and thus by Lemma 2,

$$W^m\big(\mathscr{D}(F_i)^c|f_i\big) \le 2^{-n\delta'} + 2^{-\sqrt{n}\delta}. \quad (27)$$

To achieve a small maximal error probability of the second kind we find suitable maps F_i again by random selection.

For $i \in \{1,2,\cdots,N\}$ and $y^n \in \mathscr{D}^*$ let $\bar{F}_i(y^n)$ be independent random variables such that $\bar{F}_i(y^n)$ takes every value $k \in \{1,\cdots,M''\}$ with probability $1/M''$. Let F_1 be any realization of $\bar{F}_1$.

In analogy to the ψ_j in Section III, we define random variables $\psi_{y^n} = \psi_{y^n}(\bar{F}_2)$ for every $y^n \in \mathscr{D}^*$ by

$$\psi_{y^n} = \begin{cases} 1, & \text{if } F_1(y^n) = \bar{F}_2(y^n) \\ 0, & \text{otherwise.} \end{cases} \quad (28)$$

These random variables are independent and have expected value $1/M''$. Application of [1, lemma LD] in conjunction with Lemma 2 yields the following.

Corollary 2: For $\lambda \in (0,1)$, $1/M'' < \lambda$, and for a channel V with $\|V - W\| \le \epsilon$,

$$\Pr\Big(\sum_{y^n \in \mathscr{T}^n_V(x^{*n})} \psi_{y^n} > |\mathscr{T}^n_V(x^{*n})| \cdot \lambda\Big) \le 2^{-2^{nH(W(\cdot|x^{*n}))-c(\epsilon)\cdot(\lambda\sqrt{n}\epsilon-1)}},$$

if $n \ge n_0(\epsilon)$. Consequently, with probability at least

$$1 - (n+1)^{|\mathscr{X}|\cdot|\mathscr{Y}|} \cdot 2^{-2^{n(H(W(\cdot|x^{*n}))-c(\epsilon))\cdot(\lambda\sqrt{n}\epsilon-1)}}$$

$\bar{F}_2$ satisfies, for *all* V with $\|V - W\| \le \epsilon$,

$$\sum_{y^n \in \mathscr{T}^n_V(x^{*n})} \psi_{y^n} \le |\mathscr{T}^n_V(x^{*n})| \cdot \lambda. \quad (29)$$

We now derive an upper bound on $W^m(\mathscr{D}(\bar{F}_2)|f_1)$ for those values of $\bar{F}_2$:

$$\begin{aligned} W^m\big(\mathscr{D}(\bar{F}_2)|f_1\big) &\le W^m\Big(\big(\mathscr{D}^* \times \mathscr{Y}^{\lceil\sqrt{n}\rceil}\big)^c|f_1\Big) \\ &\quad + W^m\Big(\big(\mathscr{D}^* \times \mathscr{Y}^{\lceil\sqrt{n}\rceil}\big) \cap \mathscr{D}\big(\bar{F}_2\big)|f_1\Big) \\ &\le W^n\big((\mathscr{D}^*)^c|x^{*n}\big) \\ &\quad + \sum_{\substack{y^n \in \mathscr{D}^* \\ F_1(y^n) \ne \bar{F}_2(y^n)}} W^n(y^n|x^{*n}) \cdot 2^{-\sqrt{n}\delta} \\ &\quad + \sum_{\substack{y^n \in \mathscr{D}^* \\ F_1(y^n) = \bar{F}_2(y^n)}} W^n(y^n|x^{*n}). \end{aligned}$$

By Lemma 1 we have $W^n(\mathscr{D}^*|x^{*n}) \ge 1 - 2^{-n\delta'}$.

The second summand is obviously not larger than $2^{-\sqrt{n}\delta}$. For an upper bound on the third summand we use (29). We get

$$\begin{aligned} W^m\big(\mathscr{D}(\bar{F}_2)|f_1\big) &\le 2^{-n\delta'} + 2^{-\sqrt{n}\delta} \\ &\quad + \sum_{V:\,\|V-W\|\le\epsilon} W^n(\mathscr{T}^n_V(x^{*n})|x^{*n}) \\ &\quad \cdot \frac{\sum_{y^n \in \mathscr{T}^n_V(x^{*n})} \psi_{y^n}}{|\mathscr{T}^n_V(x^{*n})|} \\ &\le 2^{-n\delta'} + 2^{-\sqrt{n}\delta} + \lambda. \end{aligned}$$

The same arguments yield the same bound for

$$W^m\big(\mathscr{D}(F_1)|\tilde{f}_2\big),$$

if $\tilde{f}_2$ denotes the encoding function defined by the map $\bar{F}_2$. We repeatedly use this argument as in Section III and construct a code length N satisfying

$$N \ge (n+1)^{-2|\mathscr{X}|\cdot|\mathscr{Y}|} \cdot 2^{2^{nH(W(\cdot|x^{*n}))-c(\epsilon)(\lambda\sqrt{n}\epsilon-1)}}$$

and an error of the second kind less than $2^{-n\delta'} + 2^{-\sqrt{n}\delta} + \lambda$.

VI. Proof of the Direct Part of Theorem 2

Since now randomization in the encoding and feedback are available, we can combine the two kinds of random experiments used for the proofs of the direct parts in [1, theorem 1] and Theorem 1, respectively. Of course such a combination imposes restrictions to the effect that now the doubly exponential capacities $\max_P I(P,W)$ and $\max_x H(W(\cdot|x)) = \max_P H(W|P)$ do not simply add. Instead, the capacity is now given by

$$\max_P \big(I(P,W) + H(W|P)\big) = \max_P H(P\cdot W). \quad (30)$$

To show this, choose a P^* such that for $Q^* = P^* \cdot W$, $H(Q^*) = \max_P H(P\cdot W)$ and define as random experiment

$$\Big(\Big(\bigcup_{Q:\,\|Q-Q^*\|\le\epsilon} \mathscr{T}^n_Q\Big) \cup \{e\}, Q^{*n}\Big).$$

This can be realized as follows. The sender chooses a sequence x^n according to the random experiment $(\mathscr{X}^n, P^{*n})$ and sends it over the channel. $Q^{*n}(y^n)$ is the probability for receiving y^n. This sequence is also known to the sender via feedback. We can therefore substitute in the previous proof $\mathscr{D}^*$ by

$$\mathscr{D}^{**} = \bigcup_{Q:\,\|Q-Q^*\|\le\epsilon} \mathscr{T}^n_Q \quad (31)$$

and get Theorem 2-a).

VII. Proof of the Converse Part of Theorem 1

We have already mentioned that in the case of feedback the proofs of the converses become much simpler than the proofs in [1]. We need here only one auxiliary result.

Lemma 3 (Image Size for a Deterministic Feedback Strategy): For any n-length feedback strategy f and any $\nu \in$

$(0,1)$,

$$\min_{\mathscr{E} \subset \mathscr{Y}^n:\ W^n(\mathscr{E}|f) \geq 1-\nu} |\mathscr{E}| \leq K = 2^{nH(W(\cdot|x^*))+\alpha\sqrt{n}} \quad (32)$$

where $H(W(\cdot|x^*)) = \max_{x \in \mathscr{X}} H(W(\cdot|x))$, $\alpha = \sqrt{\beta/\nu}$, and $\beta = \max(\log^2 3, \log^2 |\mathscr{Y}|)$. Before we prove Lemma 3 we show that it implies Theorem 1-b).

Let $\{(f_i, \mathscr{D}_i): 1 \leq i \leq N\}$ be an (n, N, λ) IDF code with $\lambda \in (0,1/2)$. We can choose ν such that $1-\nu-\lambda > 1/2$. For f_i let $\mathscr{E}_i$ be a set for which the minimum is assumed in (32). Thus we have $W^n(\mathscr{D}_i \cap \mathscr{E}_i|f_i) > 1/2$ and the sets $\mathscr{D}_i \cap \mathscr{E}_i$ $(i = 1,2,\cdots,N)$ are necessarily distinct because the errors of the second kind are smaller than $\lambda < 1/2$. Therefore, by Lemma 3, $N \leq \sum_{k=0}^{K} \binom{|\mathscr{Y}|^n}{k} \leq 2^{n \log |\mathscr{Y}| \cdot K}$ and Theorem 1-b) follows.

Proof of Lemma 3: The cardinality of the set

$$\mathscr{E}^* = \{y^n | -\log W^n(y^n|f) \leq \log K\}$$

is clearly smaller than, K, and it suffices to show that $W^n(\mathscr{E}^*|f) \geq 1-\nu$. For this we first give another description of $W^n(\mathscr{E}^*|f)$. Strategy f induces the random variables $Y^s = (Y_1, \cdots, Y_s)$; $s = 1, \cdots, n$; with distributions

$$\Pr(Y^s = y^s) = W^s(y^s|f), \qquad y^s \in \mathscr{Y}^s.$$

Defining $Z_t = -\log W(Y_t|f(Y^{t-1}))$, we can write

$$W^n(\mathscr{E}^*|f) = \Pr\left(\sum_{t=1}^{n} Z_t \leq \log K\right). \quad (33)$$

We now analyze this expression by considering the conditional expectations $\mathbb{E}(Z_t|Y^{t-1})$.

Since

$$\Pr(Y_t = y_t|Y^{t-1} = y^{t-1}) = W(y_t|f(y^{t-1})),$$

we have for $y^{t-1} \in \mathscr{Y}^{t-1}$,

$$\mathbb{E}(Z_t|y^{t-1}) = -\sum_{y_t \in \mathscr{Y}} W(y_t|f(y^{t-1}) \log W(y_t|f(y^{t-1}) \leq H(W(\cdot|x^*)),$$

and therefore

$$\mathbb{E}(Z_t|y^{t-1}) \leq H(W(\cdot|x^*)). \quad (34)$$

Finally, we introduce the RV's

$$U_t = Z_t - \mathbb{E}(Z_t|Y^{t-1}), \quad (35)$$

which obviously satisfy

$$\mathbb{E}(U_t|Y^{t-1}) = 0, \qquad \mathbb{E}U_t = 0. \quad (36)$$

Moreover, since U_s is a function of $Y_1, \cdots, Y_s$, this implies for $s < t$ $\mathbb{E}(U_t|U_s) = 0$. Therefore, the RV's $U_1, \cdots, U_n$ are uncorrelated, i.e.,

$$\mathbb{E}U_sU_t = 0, \qquad \text{for } s \neq t. \quad (37)$$

Notice that (33)–(36) and the definition of K imply

$$W^n(\mathscr{E}^*|f) \geq \Pr\left(\sum_{t=1}^{n} U_t \leq \alpha\sqrt{n}\right). \quad (38)$$

By Chebyshev's inequality,

$$\Pr\left(\sum_{t=1}^{n} U_t \leq \alpha\sqrt{n}\right) \geq 1-\nu$$

provided that

$$\operatorname{var} U_t \leq \beta, \qquad \text{for } t = 1,2,\cdots,n. \quad (39)$$

Verification of (39) completes the proof.

Using (36) we can write

$$\operatorname{var} U_t = \mathbb{E}U_t^2 = \mathbb{E}\left(U_t - \mathbb{E}(U_t|Y^{t-1})\right)^2 = \sum_{y^t} \Pr(Y^{t-1} = y^{t-1}) \cdot \mathbb{E}\left(\left(U_t - \mathbb{E}(U_t|Y^{t-1})\right)^2 | Y^{t-1} = y^{t-1}\right)$$

and by the well-known minimality property of the expected value this can be upper-bounded by

$$\sum_{y^{t-1}} \Pr(Y^{t-1} = y^{t-1}) \mathbb{E}\left(\left(U_t - \mathbb{E}(Z_t|Y^{t-1})\right)^2 | Y^{t-1} = y^{t-1}\right) = \sum_{y^{t-1}} \Pr(Y^{t-1} = y^{t-1}) \mathbb{E}(Z_t^2|Y^{t-1} = y^{t-1}).$$

By the definition of Z_t

$$\mathbb{E}(Z_t^2|Y^{t-1} = y^{t-1}) = \sum_{y_t \in \mathscr{Y}} W(y_t|f(y^{t-1})) \cdot \log^2 W(y_t|f(y^{t-1})).$$

Since $x \log^2 x$ is bounded in $[0,1]$, this quantity is bounded by a function of $|\mathscr{Y}|$ uniformly in t and y^{t-1}. A Lagrange multiplier argument gives the bound

$$\beta = \max(\log^2 3, \log^2 |\mathscr{Y}|).$$

Thus, $\operatorname{var} U_t \leq \beta$.

VIII. Proof of the Converse Part of Theorem 2

The proof is based on the same ideas as the previous one. Here we need the following auxiliary result.

Lemma 4 (Image Size for a Randomized Feedback Strategy): For any n-length randomized feedback strategy F and any $\nu \in (0,1)$,

$$\min_{\mathscr{E}' \in \mathscr{Y}^n:\ W^n(\mathscr{E}'|F) \geq 1-\nu} |\mathscr{E}'| \leq K' = 2^{nH(Q')+\alpha\sqrt{n}} \quad (40)$$

where $H(Q') = \max_P H(P \cdot W)$, $\alpha = \sqrt{\beta/\nu}$, and $\beta = \max(\log^2 3, \log^2 |\mathscr{Y}|)$.

Replacing Lemma 3, $\mathscr{E}_i$, and K in the derivation of Theorem 1-b) by Lemma 4 and the corresponding quantities $\mathscr{E}_i'$, K' we get Theorem 2-b).

Proof of Lemma 4: The randomized strategy F can be viewed as a probability distribution Q_F on the set $\mathscr{F}_n$ of n-length deterministic feedback strategies. Therefore,

$$W^n(\mathscr{E}'|F) = \sum_{g \in \mathscr{F}_n} Q_F(g) W^n(\mathscr{E}'|g). \quad (41)$$

Q_F induces the RV Y^n with distribution

$$\Pr(Y^n = y^n) = \sum_{g \in \mathscr{F}_n} Q_F(g) W^n(y^n|g).$$

We write $Q(y^n) = \Pr(Y^n = y^n)$. The cardinality of the set

$$\mathscr{E}'^* = \{y^n | -\log Q(y^n) \leq \log K'\}$$

is clearly smaller than K', and it suffices to show now that $Q(\mathscr{E}'^*) \geq 1-\eta$. Defining $Z_t' = -\log Q(Y_t|Y^{t-1})$, we can write

$$Q(\mathscr{E}'^*) = \Pr\left\{\sum_{t=1}^{n} Z_t' \leq \log K'\right\}. \tag{42}$$

For its analysis, we consider now $\mathbf{E}(Z_t'|Y^{t-1})$.

Notice that

$$\mathbf{E}(Z_t'|y^{t-1}) = -\sum_{y_t \in \mathscr{Y}} Q(y_t|y^{t-1}) \log Q(y_t|y^{t-1})$$

and that $Q(\cdot|y^{t-1})$ is a distribution of the form $P \cdot W$, because

$$Q(y_t|y^{t-1}) = \sum_{g \in \mathscr{F}_n} Q_F(g) \frac{\prod_{i=1}^{t-1} W(y_i|g(y^{i-1}))}{\sum_g Q_F(g) \prod_{i=1}^{t-1} W(y_i|g(y^{i-1}))} \cdot W(y_t|g(y^{t-1})).$$

Therefore we have

$$\mathbf{E}(Z_t'|y^{t-1}) \leq H(Q'). \tag{43}$$

This is the substitute for (39). Otherwise, we continue exactly as before. We define functions

$$U_t' = Z_t' - \mathbf{E}(Z_t'|Y^{t-1}),$$

which again have the desired properties $\mathbf{E}U_t' = 0$, $\mathbf{E}U_s'U_t' = 0$ for $s \neq t$, and $\operatorname{var} U_t' \leq \beta$. Application of Chebyshev's inequality again establishes the result.

Remark 5: The method for proving the converse parts of Theorems 1 and 2 resembles the approach of Kemperman [4] for proving the strong converse of the coding theorem for memoryless channels with feedback. This "analytical" approach turns out to be better suited for coding problems involving feedback than the "typical sequences" approach. Other such instances are the coding theory for nonstationary and infinite alphabet channels. In fact, we have alternative proofs for the converse parts of Theorems 1 and 2 via typical sequences, but they are much more complicated.

References

[1] R. Ahlswede and G. Dueck, "Identification via channels," *IEEE Trans. Inform. Theory*, this issue, pp. 15–29.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[3] J. Wolfowitz, "Coding theorems of information theory," 3rd ed. New York: Springer Verlag, 1978.

[4] J. H. B. Kemperman, "Strong converses for a general memoryless channel with feedback," in *Trans. 6th Prague Conf. Information Theory, Stat. Dec. Fct's and Rand. Proc.* Czech. Acad. Sci., 1973, pp. 375–409.

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