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

RUDOLF AHLSWDE AND GUNTER DUECK, MEMBER, IEEE

Abstract—The main contribution of our earlier work, "Identification via Channels," was that \( N = \exp (n R) \) 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. \( C_{fr} \)) 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

\[ C_f(W) = \max_{x \in I} H(W(x)) \]

and

\[ C_{fr}(W) = \max_{p} H(p, W) \quad \text{if} \quad C(W) > 0. \]

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 \((n, M, \lambda)\) feedback code \((\langle f_j, D \rangle)_{j=1, \ldots, M}\) is described as follows. There is given a finite set of messages \( M = \{1, \ldots, M\} \). One of these messages is to be sent over the channel. Message \( j \in M \) is encoded by a (vector-valued) function

\[ f_j = [f_j^1, f_j^2, \ldots, f_j^n] \quad (1) \]

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

The distribution of the random variables \((Y_t)_{t=1, \ldots, n}\) is determined by \( f_j \) and \( W \). We denote the probability of receiving \( y^n = (y_1, \ldots, y_n) \in \mathbb{F}_q^n \) if \( j \) has been encoded by \( W^n(y^n|f_j) = W(y_1|f_j^1) \cdots W(y_n|f_j^n) \). Again the \( \mathcal{D}_j \subset \mathbb{F}_q^n \) \((j=1, \ldots, M)\) are disjoint decoding sets and we require that

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

Now let \( M_f(n, \lambda) \) be the maximal integer \( M \) for which an \((n, M, \lambda)\) 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 appportionment 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 codewords 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 exponential 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) \((n, N, \lambda)\) IDF code for \( W \) is a system

\[ \{ (f_j, \mathcal{D}_j)_{i=1, \ldots, N} \} \quad \text{with} \quad f_j \in \mathcal{F}_n, \mathcal{D}_j \subset \mathbb{F}_q^n, \]

for \( i \in \{1, \ldots, N\} \).
and 
\[ W^n(\mathcal{D}|f_i) \leq \lambda \]  \[ W^n(\mathcal{D}|f_j) \leq \lambda \]  
(3) for all \( i, f \in \{ 1, \cdots, N \} \) with \( i \neq j \). A randomized \((n, N, \lambda)\) IDP code for \( W \) is a system 
\[ \{ (Q_f(i), \mathcal{D}_i) | i = 1, \cdots, N \} \]  
with \( Q_f(\cdot | i) \in \mathcal{P}(\mathcal{X}) \), \( \mathcal{D}_1 \subset \mathcal{Y}^n \), and 
\[ \sum_{g \in \mathcal{Y}^n} Q_f(g|i) W^n(D,1|g) \leq \lambda, \]  
(4) 
\[ \sum_{g \in \mathcal{Y}^n} Q_f(g|j) W^n(\mathcal{D},j|g) \leq \lambda \]  
(5) for all \( i, j \in \{ 1, \cdots, N \} \) with \( i \neq j \).

Let \( N_f(n, \lambda) \) (resp. \( N_r(n, \lambda) \)) be the maximal integer \( N \) for which a deterministic (resp. randomized) \((n, N, \lambda)\) IDP 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 \( \lambda \in (0, 1/2) \):

\[ a) \lim_{n \to \infty} \frac{1}{n} \log \log N_f(n, \lambda) \leq \max_{x \in \mathcal{X}} H(W(\cdot|x)). \]

\[ b) \lim_{n \to \infty} \sup_{x \in \mathcal{X}} \frac{1}{n} \log \log N_r(n, \lambda) \leq \max_{x \in \mathcal{X}} H(W(\cdot|x)). \]

In particular, for deterministic feedback strategies the second-order identification capacity \( C_f(W) \) equals \( \max_{x \in \mathcal{X}} H(W(\cdot|x)) \) 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 
\[ H(W(\cdot|x^*)) = \max_{x \in \mathcal{X}} H(W(\cdot|x)). \]

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 \( C_r \geq C \). 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 \( \lambda \in (0, 1/2) \):

\[ a) \lim_{n \to \infty} \frac{1}{n} \log \log N_f(n, \lambda) \leq \max_{P \in \mathcal{P}(\mathcal{X})} H(P;W). \]

\[ b) \lim_{n \to \infty} \sup_{P \in \mathcal{P}(\mathcal{X})} \frac{1}{n} \log \log N_r(n, \lambda) \leq \max_{P \in \mathcal{P}(\mathcal{X})} H(P;W). \]

**Remark 2:** We call \( W \) essentially noiseless if there exist subsets \( \mathcal{X} \subset \mathcal{X} \), \( \mathcal{Y} \subset \mathcal{Y} \) and a bijection \( g: \mathcal{X} \rightarrow \mathcal{Y} \) such that 
\[ W(g(x)|x) = 1, \]  \( \forall x \in \mathcal{X} \) \]  \( W(\cdot|x) \in \text{ convex hull } \{ W(\cdot|x) : x \in \mathcal{X} \}, \]

for \( x^* \in \mathcal{X} \).

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 
\[ I(P',W) = H(P';W) - H(W|P), \]  \( P \), 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 \( (\mathcal{X}^{*},\mathcal{Y}^{*})\mathcal{P};W \), which can be performed by sending the outcome of \( (\mathcal{X}^{*},\mathcal{Y}^{*})\mathcal{P} \) over the channel. Notice that \( H(P;W) = I(P;W) = H(W|P) \). 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 \( H(P;W) \) (which is achievable with both feedback and randomization). We choose of course a \( P \) which maximizes \( H(P;W) \). The proofs of the converses essentially say that common random experiments of higher
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
\[
W^n(\{ y^n \in \mathcal{Y}| y^n \in \mathcal{T}_p'(x^n) \})
\]
for a \( V \) with \( ||V - W|| \leq \epsilon \) \( |x^n| \leq 1 - 2^{-n \delta'} \) for \( n \geq n_0(\epsilon) \).

Lemma 2: For every \( \epsilon > 0 \) there is a \( c(\epsilon) > 0 \) such that for \( n \geq n_0(\epsilon) \)
\[
\begin{align*}
a) & \quad \bigcup_{V: ||V - W|| \leq \epsilon} \mathcal{T}_p'(x^n) \bigg| 2^n(||W_P|| - c(\epsilon)) \\
b) & \quad \bigcup_{V: ||V - W|| \leq \epsilon} \mathcal{T}_p'(x^n) \bigg| 2^n(||W_P|| + c(\epsilon)) \\
c) & \quad \mathcal{T}_p'(x^n) \bigg| 2^n(||W_P|| - c(\epsilon)) \leq 2^n(||W_P|| + c(\epsilon)) \text{ if } ||V - W|| \leq \epsilon \text{ and } \mathcal{T}_p'(x^n) \neq 0,
\end{align*}
\]
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 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 \( \mathcal{Q}' \) and \( \mathcal{Q} \). 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 \( n \geq n_0(\epsilon) \) an \( (n, M', 2^{-n\delta}) \) code
\[
\mathcal{Q}' = \{ (u'_j, \mathcal{Q}'_j) | j = 1, \cdots, M' \} \tag{8}
\]
and an \(( \lfloor \sqrt{n} \rfloor, M'', 2^{-n\delta'} \) code
\[
\mathcal{Q}'' = \{ (u''_k, \mathcal{Q}''_k) | k = 1, \cdots, M'' \} \tag{9}
\]
exist with \( M' = \lfloor 2^{n(C-\epsilon)} \rfloor \) and \( M'' = \lfloor 2^{n\delta'} \rfloor \).

We use the abbreviation \( m = n + \lfloor \sqrt{n} \rfloor \). Now any family \( \{ T_i | i = 1, \cdots, N \} \) of maps
\[
T_i: \{1, \cdots, M'\} \rightarrow \{1, \cdots, M''\}
\]
can be used to build an ID code \(( (Q(\cdot | i), \mathcal{Q}(\cdot)) | i = 1, 2, \cdots, N \) \) from \( \mathcal{Q}' \) and \( \mathcal{Q}'' \). Here \( Q(\cdot | i) \) is the uniform distribution on the set of codewords
\[
\mathcal{U}_i = \{ u'_j : u''_{T_i(j)} | j = 1, \cdots, M' \} \subset \mathcal{X}^m
\]
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 \( u'_j : u''_{k,j} \) with probability \( 1/ M'' \) for \( k \in \{1, \cdots, M''\} \). We consider the random sets
\[
\mathcal{S}_i = \{ U_{1,i}, U_{2,i}, \cdots, U_{M,i} \} \quad \text{for } i = 1, \cdots, N. \tag{10}
\]

The uniform distributions \( \mathcal{Q}(\cdot | i) \) on these sets become random distributions. The random decoding sets are
\[
\mathcal{D}(\mathcal{S}_i) = \bigcup_{j=1}^{M'} \mathcal{D}(U_{ij}) \tag{11}
\]
where
\[
\mathcal{D}(U_{ij}) = \mathcal{D}' \times \mathcal{D}'' \quad \text{if } U_{ij} = u'_j \cdot u''_{k,j}. \tag{12}
\]

We now analyze the maximal error performances of \(( (Q(\cdot | i), \mathcal{Q}(\cdot)) | i = 1, \cdots, N \) \). It is clear from the definitions (8)–(12) that for every realization \( \mathcal{S}_i \) of \( \mathcal{S}_i \),
\[
1_{M''} \sum_{u'' \in \mathcal{Q}''} W^n((Q(\cdot | i)), \mathcal{Q}(\cdot)) | u'' | \leq 2^{-n_d} + 2^{-n\delta} \tag{13}
\]
Thus only errors of the second kind remain to be considered. For this analysis we again use a large devolational approach to bound the probability that there does not exist a realization with a prescribed error of the second kind \( \lambda \) 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 \( \mathcal{S}_2 \) with any realization \( \mathcal{S}_1 \) of \( \mathcal{S}_1 \). Fix \( \mathcal{S}_1 \) and define for \( j = 1, \cdots, M' \)
\[
\psi_j = \psi_j(\mathcal{S}_2) = \begin{cases} 1, & \text{if } \mathcal{S}_2 \in \mathcal{S}_1 \\ 0, & \text{otherwise.} \end{cases} \tag{14}
\]
Since the random variables \( U_{ij} \) 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 $M'' = 2^{|\varphi|}$,
\begin{equation}
D(1/|\mathcal{M}'|) \geq \lambda \cdot n \cdot e - 1. \tag{16}
\end{equation}

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

**Corollary 1:** For $\lambda \in (0,1)$ and $1/M'' < \lambda$,
\[ \Pr \left( \sum_{j=1}^{M'} C_j > M' \cdot \lambda \right) \leq 2^{-M'(\lambda \epsilon^2 / 2)}. \]

We need one other elementary fact. Suppose that $\varphi_2 = \varphi_2'$ and $u \in \Phi_1 \cap \varphi_2'$, then
\begin{equation}
W^m(\varphi(\varphi_2')|u) \leq 2^{-n+2} - \epsilon^2. \tag{17}
\end{equation}

To see this, let $u = u': u''$. Notice that for $u \in \Phi_1$, $\varphi(\varphi_2') \cap \varphi(\varphi_2'') = \emptyset$ and that therefore
\[ W^m(\varphi(\varphi_2')|u) \leq W^m(\varphi(\varphi_2') \cap \varphi(\varphi_2'')|u). \]

Equation (17) follows, because the definitions (8) and (9) imply that
\[ W^m(\varphi(\varphi_2') \cap \varphi(\varphi_2'')|u) \leq 2^{-n+2} - \epsilon^2. \]

An upper bound on the error of the second kind is now readily established:
\begin{align*}
\sum_{u \in \Phi_1} W^m(\varphi(\varphi_2')|u) &= \sum_{u \in \Phi_1 \cap \varphi_2'} W^m(\varphi(\varphi_2')|u) \\
&\quad + \sum_{u \in \Phi_1 \setminus \varphi_2'} W^m(\varphi(\varphi_2')|u) \\
&\leq |\Phi_1 \cap \varphi_2'| |\Phi_1 \setminus \varphi_2'| 2^{-n+2} - \epsilon^2.
\end{align*}

where we have used (17). Since $|\Phi_1 \cap \varphi_2'| - \sum_{j=1}^{M'} C_j \varphi_j(\varphi_2')$.
\begin{equation}
\frac{1}{M'} \sum_{u \in \Phi_1} W^m(\varphi(\varphi_2')|u) \leq \lambda + 2^{-n+2} - \epsilon^2. \tag{18}
\end{equation}

Now fix $\lambda \in (0,1)$. By Corollary 1 for large $n$ we have that with positive probability
\begin{equation}
\frac{1}{M'} \sum_{u \in \Phi_1} W^m(\varphi(\varphi_2')|u) \leq \lambda + 2^{-n+2}. \tag{19}
\end{equation}

and similarly
\begin{equation}
\frac{1}{M'} \sum_{u \in \Phi_2} W^m(\varphi(\varphi_2')|u) \leq \lambda + 2^{-n+2}. \tag{20}
\end{equation}

Hence there is a realization $\varphi_2 = \varphi_2'$ for which (19) and (20) hold. We use this argument repeatedly for $i = 3,4,\ldots,N$ (as in [1]). An $(n,N,\lambda + 2^{-n+2})$ ID code exists, if
\begin{equation}
(N-1) \Pr \left( \sum_{j=1}^{M} C_j > M' \lambda \right) < 1. \tag{21}
\end{equation}

From Corollary 1 and $M'' = 2^{n(\epsilon^2 / 2)}$ (21) holds for every $N$ with
\[ N \leq 2^{1/2 - (\lambda \epsilon^2 / 2) - 2} \].

This proves the result.

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

**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 $\varphi'$. The outcome $u_j \in \{u_j', \ldots, u_j''\}$ is known to the sender. Then the outcome is transmitted over the channel and made known to the receiver with high probability. The parameter $M' = [2^{n(\epsilon^2)}}$ 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 $\varphi''$ as in (9). We now describe the substitute for $\varphi'$.

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

The resulting correlated random experiment ($\Phi^n, W^n(\cdot|x^*)$) needs a modification, because $W^n(\cdot|x^*)$ is far from being uniform on $\mathcal{Y}^n$. However, $W^n(\cdot|x^*)$ is essentially uniform on the set
\[ \mathcal{D}^* = \bigcup_{i \neq \epsilon} \mathcal{T}^i_\epsilon(x^*), \]
which carries essentially all its probability. We lump the small-probability set $\mathcal{D}^* \setminus \mathcal{D}^*$ together in an erasure symbol $\epsilon$ with the understanding that
\[ W^n(\epsilon|x^*) = W^n(\varphi^n - \mathcal{D}^*|x^*). \tag{25}
\]

We choose as our random experiment $(\mathcal{D}^* \cup \{\epsilon\}, W^n(\cdot|x^*))$. The price paid for more uniformity is a small error probability, if $\epsilon$ occurs. However, previously we still had to deal in $\varphi'$ with small error probabilities. By Lemma 2, $|\mathcal{D}^*| \leq c^H(W(1\epsilon))$, and this quantity now takes the role of $M'$.

Instead of the maps $T_i: \{1,\ldots,M'\} \rightarrow \{1,\ldots,M''\}$, we now use maps $F_i: \mathcal{D}^* \rightarrow \{1,\ldots,M''\}$ in the block $[n + \sqrt{n}].$
This means that after \( y^n \in \mathcal{D} \) has been received, the sender sends \( \mu_{y^n}(F_i) \), if \( i \in \{1, 2, \ldots, N\} \) is given to him.

In case \( y^n \in \mathcal{D} \) an error is declared and the sender can fill the \( |n^{1/2}| \) positions in any way, for instance by sending \( x^n[n^{1/2}] \) times again. Clearly, for each \( F_i \), we have defined an encoding function \( F_i \in \mathcal{F}_{y^n} \) as introduced in Section I.

For the decoding we define the sets
\[
\mathcal{D}(F_i) = \bigcup_{y^n \in \mathcal{D}^*} \{ y^n \times \mathcal{D}(F_i(y^n)) \}, \quad \text{for } i = 1, \ldots, N.
\] (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
\[
W^c(\mathcal{D}(F_i) | f_i) \leq W^c(\mathcal{D}^*(x^n)) + 2^{-\beta \delta},
\]
and thus by Lemma 2,
\[
W^c(\mathcal{D}(F_i) | f_i) \leq 2^{-n \delta} + 2^{-\beta \delta}.
\] (27)

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

For \( i \in \{1, 2, \ldots, N\} \) and \( y^n \in \mathcal{D}^* \) let \( F_i(y^n) \) be independent random variables such that \( F_i(y^n) \) takes every value \( k \in \{1, \ldots, M''\} \) with probability \( 1/M'' \). Let \( F_i \) be any realization of \( F_i \).

In analogy to \( \psi_j \) in Section III, we define random variables \( \psi_j = \psi_j(F_i) \) for every \( y^n \in \mathcal{D}^* \) by
\[
\psi_j = \begin{cases} 1, & \text{if } F_i(y^n) = F_i(y^n) \\ 0, & \text{otherwise} \end{cases}
\] (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 \( \|V - W\| \leq \epsilon \),
\[
\Pr \left( \sum_{y^n \in \mathcal{D}(x^n)} \psi_j > |\mathcal{F}_j(\{x^n\})| \right) \leq 2^{-n \delta},
\]
if \( n \geq n_0(\epsilon) \). Consequently, with probability at least
\[
1 - \left( n + 1 \right)^{-|\mathcal{D}|}, 2^{-n \delta}, \text{for } V \text{ with } \|V - W\| \leq \epsilon,
\]
\[
\sum_{y^n \in \mathcal{D}(x^n)} \psi_j \leq |\mathcal{F}_j(\{x^n\})| \lambda. \tag{29}
\]

We now derive an upper bound on \( W^c(\mathcal{D}(\overline{F}_2) | f_1) \) for those values of \( \overline{F}_2 \):
\[
W^c(\mathcal{D}(\overline{F}_2) | f_1) \leq W^c(\mathcal{D} \times \mathcal{D}(\overline{F}_1)) | f_1 \]
\[
+ W^c(\mathcal{D} \times \mathcal{D}(\overline{F}_1) \cap \mathcal{D}(\overline{F}_2)) | f_1 \]
\[
\leq W^c(\mathcal{D} \times \mathcal{D}(\overline{F}_1)) | x^n \]
\[
+ \sum_{y^n \in \mathcal{D}^*} W^c(y^n | x^n). 2^{-\beta \delta}, \quad \text{if } F_i(y^n) = \overline{F}_i(y^n) \]
\[
+ \sum_{y^n \in \mathcal{D}^*} W^c(y^n | x^n). \quad \text{if } F_i(y^n) \neq \overline{F}_i(y^n).
\]

By Lemma 1 we have \( W^c(\mathcal{D}(x^n)) \geq 1 - 2^{-\beta \delta} \).

The second summand is obviously not larger than \( 2^{-\beta \delta} \).

For an upper bound on the third summand we use (29).

We get
\[
W^c(\mathcal{D}(\overline{F}_2) | f_1) \leq 2^{-n \delta} + 2^{-\beta \delta}.
\]

The same arguments yield the same bound for
\[
W^c(\mathcal{D}(F_1) | \overline{f}_2),
\]
if \( \overline{f}_2 \) denotes the encoding function defined by the map \( \overline{F}_2 \). We repeatedly use this argument as in Section III and construct a code length \( N \) satisfying
\[
N \geq (n + 1)^{-|\mathcal{D}|}, 2^{-n \delta}, \text{and an error of the second kind less than } 2^{-n \delta} + 2^{-\beta \delta} + \lambda.
\]

**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 I] 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 \( I(W | P) \) do not simply add. Instead, the capacity is now given by
\[
\max P I(P, W) + H(W | P) = \max P I(P, W).
\] (30)

To show this, choose a \( P^* \) such that for \( Q^* = P^* W, \)
\[\]
\[H(Q^*) = \max P H(P, W) \] and define a random experiment
\[
\left( \bigcup Q : \|Q - Q^*\| \leq \epsilon, \mathcal{F}_j \right), \quad Q^* \in \mathcal{D}^*, \]
\[
\text{This can be realized as follows. The sender chooses a sequence } x^n \text{ according to the random experiment } (\mathcal{D}^*, Q^*) \text{ and sends it over the channel. } Q^*(y^n) \text{ is the probability for receiving } x^n. \text{This sequence is also known to the sender via feedback. We can therefore substitute in the previous proof } Q^* \text{ by }
\]
\[
\mathcal{D}^* = \bigcup Q : \|Q - Q^*\| \leq \epsilon \]
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 \( \nu \in \mathcal{D}^* \).
\[
\min_{\delta \in \mathcal{Q}^r \cap \mathcal{D} \mid \delta} |\delta^r| \leq K = 2^{nH(W(\cdot|x^*))} + a \delta^r
\]  
(32)

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

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

Proof of Lemma 3: The cardinality of the set
\[\delta^* = \{ y^* | \log W^{y}\{y^* | f\} \leq \log K \}\]
is clearly smaller than \(K\), and it suffices to show that \(W^r(\delta^*) \geq 1 - \nu\). For this we first give another description of \(W^r(\delta^*)\). Strategy \(f\) induces the random variables \(Y = (Y_1, \ldots, Y_s)\), \(s = 1, \ldots, n\), with distributions
\[
\Pr(Y = y^*) = W^r(\delta^*) f(y^*), \quad y^* \in \mathcal{Y}.
\]
Defining \(Z = -\log W(Y_j/f(Y_j^{y^*}))\), we can write
\[
W^{y}\{y^* | f\} = \Pr\left(\sum_{i=1}^{n} Z_i \leq \log K\right).
\]
(33)

We now analyze this expression by considering the conditional expectations \(E(Z_i|Y^{y^*})\).

Since
\[
\Pr(Y = y_j|Y^{y^*}) = W\{y_j | f(y^*)\},
\]
we have for \(y^{y^*} \in \mathcal{Y}\),
\[
E(Z_i|y^{y^*}) = - \sum_{y \in \mathcal{Y}} W\{y | f(y^{y^*})\} \log W\{y | f(y^{y^*})\} \leq H(W(\cdot|x^*))
\]
and therefore
\[
E(Z_i|y^{y^*}) \leq H(W(\cdot|x^*))
\]
(34)

Finally, we introduce the RVs
\[
U_i = Z_i - E(Z_i|Y^{y^*}),
\]
which obviously satisfy
\[
E(U_i|Y^{y^*}) = 0, \quad E(U_i) = 0.
\]
(36)

Moreover, since \(U_i\) is a function of \(Y_{s_i}, \ldots, Y_s\), this implies for \(s \leq t \ E(U_i|U_t) = 0\). Therefore, the RVs \(U_1, \ldots, U_n\) are uncorrelated, i.e.,
\[
E(U_i U_j) = 0, \quad \text{for} \ s \neq t.
\]
(37)

Notice that (33)-(36) and the definition of \(K\) imply
\[
W^y(\delta^*) \geq \Pr\left(\sum_{i=1}^{n} U_i \leq \alpha \sqrt{n}\right).
\]
(38)

By Chebyshev's inequality,
\[
\Pr\left(\sum_{i=1}^{n} U_i \leq \alpha \sqrt{n}\right) \geq 1 - \nu
\]
provided that \(\var U_i \leq \beta, \quad \text{for} \ t = 1, 2, \ldots, n\).

Verification of (39) completes the proof.

Using (36) we can write
\[
\var U_i = E(U_i^2) - E(\var U_i Y^{y^*})^2 - \sum_{y^{y^*}} \Pr(Y^{y^*} = y^{y^*}) \cdot \{U_i - E(U_i Y^{y^*})\}^2 Y^{y^*} = y^{y^*}
\]
and by the well-known minimality property of the expected value this can be upper-bounded by
\[
\sum_{y^{y^*}} \Pr(Y^{y^*} = y^{y^*}) \cdot \{U_i - E(U_i Z^{y^*})\}^2 Z^{y^*} = y^{y^*}
\]
By the definition of \(Z_i\)
\[
E(Z_i^2) = E(Z_i Y^{y^*}) = \sum_{y \in \mathcal{Y}} W^y(\delta^*) f(y^{y^*})
\]
Since \(x \log^2 x\) is bounded in \([0,1]\), this quantity is bounded by a function of \(\|\vartheta\|\) uniformly in \(t\) and \(y^{y^*}\). A Lagrange multiplier argument gives the bound
\[
\beta = \max(\log^2(3), \log^2(2/\|\vartheta\|))
\]
Thus, \(\var U_i \leq \beta\).

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 \(\nu \in (0,1)\),
\[
\min_{\delta \in \mathcal{Q}^r \cap \mathcal{D} \mid \delta} |\delta^r| \leq K = 2^{nH(Q^r)} + a \delta^r
\]
(40)

where \(H(Q^r) = \max_{P \in \mathcal{P}_n} H(P, W), \quad a = \sqrt{\beta / \nu}\), and \(\beta = \max(\log^2(3), \log^2(2/\|\vartheta\|))\).

Replacing Lemma 3, \(\delta^r\), and \(K\) in the derivation of Theorem 1-b) by Lemma 4 and the corresponding quantities \(\delta^*, K^*\) we get Theorem 2-b).

Proof of Lemma 4: The randomized strategy \(F\) can be viewed as a probability distribution \(Q_\mathcal{F}\) on the set \(\mathcal{P}_n\) of \(n\)-length deterministic feedback strategies. Therefore,
\[
W^y(\delta^* F) = \sum_{F \in \mathcal{P}_n} Q_\mathcal{F}(g) W^y(\delta^*|g).
\]
(41)

\(Q_\mathcal{F}\) induces the RV \(Y^*\) with distribution
\[
\Pr(Y^* = y^*) = \sum_{F \in \mathcal{P}_n} Q_\mathcal{F}(g) W^y(y^*|g).
\]
We write \(Q(y^*) = \Pr(Y^* = y^*)\). The cardinality of the set
\[
\delta^* = \{ y^* | \log Q(y^*) \leq \log K^* \}
\]
is clearly smaller than $K'$, and it suffices to show now that $Q(\varepsilon^{(s)}) \geq 1 - \eta$. Defining $Z'_t = -\log Q(Y_t | Y^{t-1})$, we can write

$$Q(\varepsilon^{(s)}) = \Pr \left( \sum_{i=1}^{n} Z'_i \leq \log K' \right).$$

(42)

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

Notice that

$$E(Z'_t | Y^{t-1}) = - \sum_{y_t} Q(y_t | Y^{t-1}) \log Q(y_t | Y^{t-1})$$

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

$$Q(y_t | Y^{t-1}) = \sum_{g \in \mathcal{X}} Q_f(g) \frac{\prod_{i=1}^{t-1} W(y_i | g(y_i^{t-1}))}{\sum_{g' \in \mathcal{X}} Q_f(g') \prod_{i=1}^{t-1} W(y_i | g'(y_i^{t-1}))} \cdot W(y_t | g(y_t^{t-1})).$$

Therefore we have

$$E(Z'_t | Y^{t-1}) \leq H(Q').$$

(43)

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

$$U'_t = Z'_t - E(Z'_t | Y^{t-1}),$$

which again have the desired properties $E(U'_t) = 0$, $E(U'_t U'_s) = 0$ for $t \neq s$, 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