On Identification

Christian Kleinewächter

Contents

1 Introduction

In Shannon's classical model of transmitting a message over a noisy channel we have the following situation:

There are two persons called *sender* and *receiver*. Sender and receiver can communicate via a *channel*. In the simplest case the sender just puts some *input letters* into the channel and the receiver gets some *output letters*. Usually the channel is *noisy*, i.e. the channel output is a random variable whose distribution is governed by the input letters. This model can be extended in several ways: Channels with *passive feedback* for example give the output letters back to the sender. Multiuser channels like *multiple access channels* or *broadcast channels* (which will not be considered in this paper) have several senders or receivers which want to communicate simultaneously. Common to all these models of *transmission* is the task that sender and receiver have to perform: Both have a common *message set* M and the sender is given a *message* $i \in M$. He has to *encode* the message (i.e. transform it into a sequence of input letters for the channel) in such a way, that the receiver can *decode* the sequence of output letters so that he can decide with a small probability of error what the message i was. The procedures for encoding and decoding are called a *code* for the channel and the number of times the channel is used to transmit one message is called the *blocklength* of the code.

While this model of transmission is probably the most obvious model, there are also other ones. Maybe the receiver does not need to know what the exact message is, he is satisfied to get only partial information. Depending on the type of questions the receiver has about the messages, transmission codes can be rather inefficient. If one uses special codes suited to answer only these questions, then there might be more efficient codes with respect to the number of possible messages for given blocklength.

One of these models is identification. In the theory of identification, the receiver has a message of special interest to him, and he only wants to know whether the sender was given this special message or not. This special message is unknown to the sender (otherwise this problem would be trivial, the sender just has to encode one of two alternatives). As an application consider the following situation: Sender and receiver both have a large amount of text (like a book). The receiver now wants to know if both texts are equal (otherwise the receiver may want to order a copy of the sender's text). The trivial solution using a transmission scheme would be that the sender simply encodes the whole text, sends it over the channel, the receiver decodes the text and compares it to his text, but of course this is very inefficient.

Usually the number of messages that one can transmit grows exponentially in the blocklength, so this method would require codes with blocklengths proportional to the length of the text. The ratio between the logarithm of the number of messages and the blocklength is called the *rate* of the code and the maximal rate that is asymptotically achievable is called the capacity of the channel.

In [1] and [2] Ahlswede and Dueck showed that the maximal sizes of an identification code (*ID-code*) for a stationary discrete memoryless channel without and with noiseless feedback grow double exponentially fast in the blocklength and they determined the second order rate for these channels, if we allow an arbitrary small but positive probability of error that the answer we get is false. (For channels without feedback this holds only for randomized codes.) The rate of these codes is therefore defined as the ratio of the iterated logarithm of the number of messages and the blocklength.

Applied to our previous example this means that we need only codes of blocklength proportional to the logarithm of the length of the text for identification.

If we deal with ID-codes, we have to consider two kinds of errors, namely if message i is given to the sender the receiver, who is interested in message i, can decide that his message was not sent (missed identification) and a receiver who is interested in $j \neq i$ can decide that j was sent (false identification). These two errors will turn out to be of different nature. While missed identifications are caused by errors in the transmission, false identifications are also inherent to the codes.

In the case of a channel W without feedback the second order identification capacity is the classical Shannon capacity

$$C = \max_P I(P, W)$$

Here P is a distribution on the input letters and $I(P, W)$ is the *mutual information* of two random variables with joint distribution $P_{XY}(x, y) = P(x)W(y|x)$, that means these random variables behave like channel input respectively output.

The upper bound for channels without feedback was proved under the assumption that the allowed error probability vanishes exponentially fast. In [4] Han and Verdú proved the upper bound for constant error probability smaller than 1/2. It is obvious, that if we allow error probabilities larger than 1/2, then trivial codes, where the receiver simply guesses, achieve infinite rates. There are also other conditions on the error probabilities to avoid these trivialities, and mostly they lead to the same capacity, for example, if the sum of the maximal probability of accepting a false message plus the maximal probability of falsely rejecting a message should be less than 1. For the proof of Theorem 1 in Section 3 we show that an even slightly weaker condition suffices, namely that for each message the sum of error probabilities should be less than 1.

For a channel W with noiseless feedback (i.e. the sender immediately gets each letter that is received) the capacity is

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

for deterministic and

$$\max_{P} H(PW)$$

for randomized codes, if there are two different rows in W, otherwise they are 0. Identification codes for channels with feedback will be called *IDF-codes*.

The proof of the direct parts in [2] works in two stages: In the first stage, sender and receiver communicate over the channel with the goal to generate two random variables which are equal with high probability and have large entropy. These random variables are used to select a function which maps messages to a rather small set. In the second step (which contributes only negligible to the blocklength) the sender evaluates this function at the message and transmits it to the receiver. The receiver evaluates the function at the message he is interested in and compares the result with the value he got from the sender. He will vote for a positive identification if and only if they are equal.

This leads to the following interpretation of these results: If the identification capacity is positive, then it equals the common randomness capacity as defined in [8], i.e. the maximal entropy of an almost uniform random experiment which can be set up by communicating over the channel and is known with high probability to both sender and receiver. In the absence of feedback, one possibility to achieve the maximal possible rate of such a common random experiment is that the sender performs a uniform random experiment and transmits the result to the receiver using an ordinary transmission code. If noiseless feedback is available, the sender sends letters in such a way, that the entropy of the channel output (which he knows from the feedback channel) is maximized, where he can either use a deterministic or randomized input strategy, depending on the kind of code he may use. This interpretation proved to be the right one also for other kind of channels like the multiple access channel (see [3]).

Thus the question arises if this equality is valid in general.

The answer is negative. In [7] Ahlswede gives an example of a non stationary memoryless channel with double exponentially growing input alphabet with identification capacity 1 and common randomness capacity 0. The structure of this channel has some similarities to the structure of ID-codes used in most of the achievability proofs for ID-coding theorems, thus it can be viewed as a channel with "built–in" ID-encoder.

In Section 3 we give a counterexample for the other direction. For given real numbers C_{ID} and C_{CR} with $0 < C_{ID} < C_{CR}$, we will explicitly construct a discrete channel with memory and noiseless passive feedback with identification capacity C_{ID} and common randomness capacity C_{CR}. This channel is constructed in such a way that it can be used in two ways. In one respect, the channel is good for the generation of common randomness, in the other it is suitable for identification.

It is quite reasonable to consider channels with memory. One may think for example of a system where data are transmitted by different voltage levels at high frequency. Because of the electrical capacity of the system it can be difficult to switch from a low voltage level to a high one and vice versa. There are also certain types of magnetic recording devices have problems with long sequences of the same letter. These examples for instance lead to the notion of run length limited codes. A third example are systems requiring the use of binary codewords which have approximately the same number of zeroes and ones. This limitation arises if the system can only transmit an unbiased alternating current, therefore these codes are called DC-free.

In [6] Ahlswede and Zhang gave bounds on the maximal rate for ID-codes for discrete memoryless channels with noisy feedback. Channels without feedback can in this model be described by a feedback channel that maps all input letters with probability 1 on one special feedback letter. The case with noiseless feedback is described by a feedback channel that has positive probability for a pair (y, z) of output respectively feedback letters only if $y = z$. In the cases of either no feedback or noiseless feedback, the upper and lower bound coincide and therefore this can be viewed as a generalization and unification of the results of [1], [2] and [4], where the identification capacities for those channels were determined. Unfortunately in general there is a gap between the upper and lower bound.

In Section 4 we improve the upper bound on the size of *deterministic* codes given in [6] for the channels where the *main channel is noiseless*. The known bound states that the second order rate of deterministic ID codes is independently bounded by the maximal mutual information that channel input and feedback give about the channel output and the maximal entropy of the feedback. We improve this bound by showing that these two quantities cannot be maximized independently, instead one has to choose an *input distribution that is good for both*, transmission and the generation of randomness. We will show for the channels considered the new upper bound equals the known lower bound, therefore the deterministic identification capacity for these channels is now known.

2 Auxiliary results

In this section we will introduce some basic definitions and well know facts. All logarithms in this paper are assumed to be binary and we define $0 \log 0$ to be 0.

Definition 2.1 *For an arbitrary set $\mathcal{S}$ we will denote by $\mathcal{P}(\mathcal{S})$ the set of all probability distributions on $\mathcal{S}$.*

Definition 2.2 (Entropy) *Let P be a probability distribution on a finite set $\mathcal{X}$. The entropy of P is defined as*

$$H(P) \triangleq -\sum_{x \in \mathcal{X}} P(x) \log P(x).$$

Also if X is a random variable with distribution P, we define the entropy of X by

$$H(X) \triangleq H(P).$$

Entropy is a concave function, because the function $f(t) = -t \log t$ is concave. Furthermore, it is *Schur-concave*:

Lemma 2.1 (Schur-concavity of entropy) *Let $n \in \mathbb{N}$. Let P, Q be two probability distributions on $\{1, \ldots, n\}$ which are non increasing. If P and Q satisfy*

$$\sum_{i=1}^{k} P(i) \leq \sum_{i=1}^{k} Q(i), \tag{2.1}$$

then $H(P) \geq H(Q)$.

Entropy can be viewed as a measure of uncertainty of a random variable. For example, if one has to guess the value of a random variable X using yes-/no-questions about its value, an optimal strategy takes on the average approximately $H(X)$ questions. (They differ by at most 1 question.) To catch the dependency between random variables, we need the concept of conditional entropy:

Definition 2.3 (Conditional Entropy) *Let X, Y be random variables on finite sets $\mathcal{X}, \mathcal{Y}$ with distributions P resp. Q. $Q(\cdot|x)$ is the distribution of Y under the condition $X = x$. The conditional entropy of Y given X is defined by*

$$H(Y|X) \triangleq -\sum_{x\in\mathcal{X}} P(x) \sum_{y\in\mathcal{Y}} Q(y|x) \log Q(y|x).$$

The conditional entropy $H(Y|X)$ is the uncertainty about Y after having observed X, so the difference between $H(Y)$ and $H(Y|X)$ is the decrease in uncertainty about Y by observing X or loosely speaking the amount of information we "learn" about Y through observation of X. This quantity is called *mutual information.*

Definition 2.4 (Mutual information) *Let X and Y be random variables on finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then we define the mutual information between X and Y by*

$$I(X \wedge Y) \triangleq H(Y) - H(Y|X).$$

The following lemma gives an upperbound on the difference of entropies of two distributions as a function of the l_1-distance (and the cardinality of $\mathcal{X}$, which is usually fixed):

Lemma 2.2 *Let P and Q be two distributions on a set $\mathcal{X}$ with*

$$\sum_{x\in\mathcal{X}} |P(x) - Q(x)| \leq \delta \leq \frac{1}{2}. \tag{2.2}$$

Then

$$|H(P) - H(Q)| \leq -\delta \log \frac{\delta}{|\mathcal{X}|}. \tag{2.3}$$

Proof: Set $f(t) = -t \log t$. f is concave in $[0, 1]$. Thus for $0 \leq \tau \leq \frac{1}{2}$, $0 \leq t \leq 1 - \tau$ we have

$$f(t) - f(t + \tau) \leq f(0) - f(\tau) \tag{2.4}$$

and

$$f(t + \tau) - f(t) \leq f(1 - \tau) - f(1) \tag{2.5}$$

which gives

$$|f(t) - f(t+\tau)| \leq \max\{f(\tau), f(1-\tau)\} = f(\tau). \tag{2.6}$$

Thus

$$\begin{aligned}
|H(P) - H(Q)| &\leq \sum_{x \in \mathcal{X}} |f(P(x) - f(Q(x)))| \\
&\leq \sum_{x \in \mathcal{X}} f(|P(x) - Q(x)|) \\
&\leq \delta \left(\sum_{x \in \mathcal{X}} f\left(\frac{|P(x) - Q(x)|}{\delta} \right) - \log \delta \right) \\
&\leq \delta(\log |\mathcal{X}| - \log \delta)
\end{aligned}$$

where the last step follows from Jensen's inequality. □

Another important application of entropy is counting sequences that have a given empirical distribution:

Definition 2.5 (n-type) *Let $n \in \mathbb{N}$. A probability distribution P on a finite set $\mathcal{X}$ is called n-type, if*

$$P(x) \in \left\{ \frac{0}{n}, \frac{1}{n}, \ldots, \frac{n}{n} \right\} \text{ for all } x \in \mathcal{X}.$$

The set of all sequences in $\mathcal{X}^n$ that have empirical distribution P will be denoted by T_P. Vice versa if a sequence x^n is given, we will denote its empirical distribution by P_x.

n-types can be viewed as empirical distribution of sequences of length n. For fixed n-type P their number can be estimated as follows:

Lemma 2.3 *Let $n \in \mathbb{N}$. For an n-type P*

$$\frac{2^{nH(P)}}{(n+1)^{|\mathcal{X}|}} \leq |T_P| \leq 2^{nH(P)}. \tag{2.7}$$

A proof can be found for example in [10].

The following definition expands the notion of types to pairs of sequences:

Definition 2.6 *Let $\mathcal{X}$, $\mathcal{Z}$ be finite sets, let $W : \mathcal{X} \to \mathcal{Z}$ be a stochastic matrix, i.e. a matrix $W = (W(z|x))_{x\in\mathcal{X},z\in\mathcal{Z}}$. Let $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$. Then we say that $z^n = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ has conditional type W given x^n if for all pairs $(x, z) \in \mathcal{X} \times \mathcal{Z}$*

$$< x, z|x^n, z^n >=< x|x^n > W(z|x) \tag{2.8}$$

where $< x|x^n >= |\{i|x_i = x\}|$ is the number of occurences of letter x in x^n and $< x, z|x^n, z^n >= |\{i|x_i = x \wedge z_i = z\}|$ is the number of occurences of the pair (x, z) in the sequence $((x_1, z_1), \ldots, (x_n, z_n))$. The set of these sequences is denoted by $T_W(x^n)$.

As a corollary from Lemma (2.3) we get

Corollary 2.1 *Let $n \in \mathbb{N}$, let $\mathcal{X}$, $\mathcal{Z}$ be finite sets. For $x^n \in \mathcal{X}^n$ and a stochastic matrix $W : \mathcal{X} \to \mathcal{Z}$ (i.e. a matrix $W = (W(z|x))_{x\in\mathcal{X},z\in\mathcal{Z}}$ with $\sum_{z\in\mathcal{Z}} W(z|x) = 1$ for all $x \in \mathcal{X}$) such that $T_V(X^n) \neq \emptyset$ we have*

$$\frac{2^{nH(Z|X)}}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} \le |T_W(x^n)| \le 2^{nH(Z|X)}, \tag{2.9}$$

where X, Z are random variables with joint distribution $P_{XZ}(x, z) = P(x)W(z|x)$.

3 Common randomness capacity can exceed identification capacity

In this section we will construct a channel where the *identification capacity is positive but smaller than the common randomness capacity.*

We will extend the definitions of [2] from memoryless channels with feedback to channels with memory. These channels will not be considered in general, but only a very specific subclass of them.

Definition 3.1 (Discrete channel with infinite memory) *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets. Let*

$$\mathcal{W} = \{W|W : \mathcal{X} \to \mathcal{Y}\}$$

be the set of all stochastic matrices with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. A discrete channel V with infinite memory (here abbreviated as DC) with respect to the input letters *is a series $V = (V_i)_{i\in I\!N}$ of maps, where $V_i : \mathcal{X}^{i-1} \to \mathcal{W}$. V_1 is understood to be an element of $\mathcal{W}$. Given an input sequence $x^n = (x_1, \dots, x_n)$ of length n the transmission probability for output sequence $y^n = (y_1, \dots, y_n)$ is given by*

$$V^n(y^n|x^n) = \prod_{i=1}^{n} V_i(x_1, \dots, x_{i-1})(y_i|x_i). \tag{3.1}$$

In our example the input alphabet $\mathcal{X}$ will be the set $\{1, 2\}$ and the channel will in fact "remember" only the first input letter.

We consider DCs with immediate noiseless feedback, i.e. after each transmitted letter the sender gets the output letter on the receivers side and his further encodings may depend on all previous output letters.

Formally, if we have N messages $1, \dots, N$, then the encoding of message $j \in \{1, \dots N\}$ is a vector $f_j = [f_j^1, \dots f_j^n]$, where $f_j^i : \mathcal{Y}^{i-1} \to \mathcal{X}$. f_j^1 is understood to be an element of $\mathcal{X}$. The set of all functions of this kind is denoted by $\mathcal{F}_n$.

For a given DC V and an encoding f the probability $V^n(y^n|f)$ of receiving an output sequence $y^n = (y_1, \dots, y_n)$ is given by

$$V^n(y^n|f) = V_1(y_1|f^1) \cdot V_2(f^1)(y_1|f^2(y_1)) \cdot \ldots \cdot \\ V_n(f^1, f^2(y_1), \ldots, f^{n-1}(y_1, \ldots, y_{n-2}))(y_n|f^1(y_1, \ldots, y_{n-1})). \tag{3.2}$$

Definition 3.2 (Deterministic identification code for a DC) *Let n, $N \in I\!N$, $0 \leq \lambda \leq 1$. A deterministic (n, N, λ) identification code for a DC V with feedback is a family of pairs*

$$\{(f_j, \mathcal{D}_j)\}_{j \in \{1,\ldots,N\}}, \mathcal{D}_j \subset \mathcal{Y}^n \tag{3.3}$$

with the following properties

$$V^n(\mathcal{D}_i|f_i) > 1 - \lambda \textit{ for all } i \in \{1, \ldots, M\} \tag{3.4}$$

$$V^n(\mathcal{D}_j|f_i) < \lambda \textit{ for all } i, j \in \{1, \ldots, M\}, i \neq j \tag{3.5}$$

Definition 3.3 (Randomized identification code for a DC) *Let n, $N \in I\!N$, $0 \leq \lambda \leq 1$. A randomized (n, N, λ) identification code for a DC V with feedback is a family of pairs*

$$\{(Q(\cdot|j), \mathcal{D}_j)\}_{j \in \{1,\ldots,N\}}, \quad Q(\cdot|j) \in \mathcal{P}(\mathcal{F}_n), \quad \mathcal{D}_j \subset \mathcal{Y}^n \tag{3.6}$$

with the following properties

$$\sum_{f \in \mathcal{F}_n} Q(f|i) W^n(\mathcal{D}_i|f) > 1 - \lambda \textit{ for all } i \in \{1, \ldots, M\} \tag{3.7}$$

$$\sum_{f \in \mathcal{F}_n} Q(f|i) W^n(\mathcal{D}_j|f) < \lambda \textit{ for all } i, j \in \{1, \ldots, M\}, i \neq j \tag{3.8}$$

The definition of the common randomness capacity for a specific model is affected by the resources which are given to the sender and the receiver and this dependency lies in the definition of a permissible pair of random variables. Generally speaking, a pair of random variables is permissible (for length n), if it can be realized by a communication over the channel with blocklength n.

For our model it is the following (according to the notation and definitions of [8]):

A pair (K, L) of random variables is permissible, if K and L are functions of the channel output Y^n. If we allow randomization in the encoding, K may also depend on a random variable M which is generated in advance by the sender.

Definition 3.4 (Common randomness capacity) *The common randomness capacity C_{CR} is the maximal number ν such, that for a constant $c > 0$ and for all $\epsilon > 0$, $\delta > 0$ and for all n sufficiently large there exists a permissible pair (K, L) of random variables for length n on a set $\mathcal{K}$ with $|\mathcal{K}| < e^{cn}$ with*

$$Pr\{K \neq L\} < \epsilon$$

and

$$\frac{H(K)}{n} > \nu - \delta.$$

The upper bound on the cardinality $|\mathcal{K}|$ of the ground set guarantees that the small amount of probability where K and L may differ can give only small contribution to the entropy. It also ensures that $H(L) \approx H(K)$. Obviously, for a channel with noiseless passive feedback the common randomness capacity equals the maximal output entropy.

Theorem 1 *Let α, β be real numbers with $0 < \alpha < \beta$. There exists a DC V with common randomness capacity β and identification capacity α, that is if we denote by $N_f(n, \lambda)$ the maximal number N for which a deterministic (n, N, λ)-code for the DC V exists, then*

$$\lim_{n\to\infty} \frac{\log\log N_f(n, \lambda)}{n} = \alpha. \tag{3.9}$$

Analogously if we denote by $N_F(n, \lambda)$ the maximal number N for which a randomized (n, N, λ)-code exists, then

$$\lim_{n\to\infty} \frac{\log\log N_F(n, \lambda)}{n} = \alpha. \tag{3.10}$$

For the proof of Theorem 1 we give a construction for the channel. We show that from an (n, N, λ) identification code for the discrete memoryless channel with feedback induced by a certain matrix W_B we can construct an (n, N, λ) identification code for our DC V. Conversely we will show that a deterministic (n, N, λ) identification code for V induces a deterministic $(n, N-1, \lambda)$ identification code for the DMC induced by W_B. For randomized codes, we will show that a code for the DC V leads to a code for the DMC W_B with a number of messages that is smaller by at most a constant factor that depends on the error probability λ.

Set $\mathcal{X} = \{1, 2\}$, $\mathcal{Y} = \{1, \ldots, \lceil 2^\beta \rceil\}$. Let P_0 be the probability distribution on $\mathcal{Y}$ with $P_0(1) = 1$, fix any two probability distributions P_1, P_2 on $\mathcal{Y}$ with

$$\begin{aligned} H(P_1) &= \alpha && (3.11)\\ H(P_2) &= \beta && (3.12)\\ P_1(i) &\geq P_1(i+1) \text{ for all } i = 1, \ldots |\mathcal{Y}| - 1. && (3.13) \end{aligned}$$

Now define three stochastic matrices $W_A, W_B, W_C : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ by

$$\begin{aligned} W_A(\cdot|x) &= P_0 \text{ for all } x \in \mathcal{X} && (3.14)\\ W_B(\cdot|x) &= P_{x-1} && (3.15)\\ W_C(\cdot|x) &= P_2 \text{ for all } x \in \mathcal{X}. && (3.16) \end{aligned}$$

Finally we will define our DC V: Set $V_0 \triangleq W_A$ and

$$W_i(x_1, \ldots, x_{i-1}) = \begin{cases} W_B & \text{if } x_1 = 1 \\ W_C & \text{otherwise.} \end{cases} \qquad (3.17)$$

By this construction the sender decides with his first letter whether he wants to use the channel for the maximal possible generation of common randomness (and giving up the possibility to transmit messages), or whether he also wants a channel with positive transmission capacity.

For this channel and an encoding $f_j = [f_j^1, \ldots, f_j^n]$ Equation (3.2) yields

$$V^n(y^n|f_j) = \begin{cases} P_0(y_1) \cdot W_B(y_2|f_j^2(y_1)) \cdot \ldots \cdot W_B(y_n|f_j^n(y_1, \ldots, y_{n-1})) & \text{if } f_j^1 = 1 \\ P_0(y_1) \cdot P_2(y_2) \cdot \ldots \cdot P_2(y_n) & \text{otherwise.} \end{cases} \qquad (3.18)$$

It is obvious that this channel has a common randomness capacity of β, since the maximal output entropy is achieved, if the sender sends a 2 as first letter and arbitrary letters afterwards, which gives $(n-1)\cdot\beta$ bits of common randomness for blocks of length n.

The main tools in the calculation of the identification capacities of this channels will be Theorem 1 and Theorem 2 from [2] which give the capacities for memoryless channels:

Theorem 2 *Let W_i be a stochastic matrix. Let $0 < \lambda < 1/2$ be given. If there are two different rows in W_i, then for the maximal sizes $N_{f,W_i}(n,\lambda)$ of deterministic identification codes for W_i with error probability smaller than λ*

$$\liminf_{n\to\infty} n^{-1}\log\log N_{f,W_i}(n,\lambda) = \max_{x\in\mathcal{X}} H(W_i(\cdot|x)). \tag{3.19}$$

For the maximal sizes $N_{F,W_i}(n,\lambda)$ of randomized identification codes for W_i with error probability smaller than λ we have

$$\liminf_{n\to\infty} n^{-1}\log\log N_{F,W_i}(n,\lambda) = \max_{P\in\mathcal{P}(\mathcal{X})} H(PW_i). \tag{3.20}$$

Lemma 3.1 *Let $\lambda < 1/2$, $n \in I\!N$, let $\{(f_j,\mathcal{D}_j)\}_{j\in\{1,\ldots,N\}}$ be an (n,N,λ)-code for the DC defined before, then there exists at most one index j with $f_j^1 = 2$.*

Proof:

Suppose we have $j \neq j'$ with $f_j^1 \neq 1$ and $f_{j'}^1 \neq 1$. Then by Equations (3.4), (3.5) and (3.18)

$$1/2 < 1-\lambda < W^n(\mathcal{D}_j|f_j) = W^n(\mathcal{D}_j|f_{j'}) < \lambda < 1/2. \quad \square \tag{3.21}$$

Lemma 3.2 *Let $\lambda < 1/2$, $n \in I\!N$. Then*

$$N_f(n,\lambda) \leq N_{f,W_B}(n-1,\lambda) + 1.$$

Proof:

Let $\{(f_j,\mathcal{D}_j)\}_{j\in\{1,\ldots,N\}}$ be a deterministic (n,N,λ) ID-code for the DC V. By Lemma (3.1) we can assume without loss of generality that $f_j^1 = 1$ for all $j \in \{1,\ldots,N-1\}$. Define a code $\{(f'_j,\mathcal{D}'_j)\}_{j\in\{1,\ldots,N-1\}}$ of length $n-1$ with $N-1$ codewords by the following rules

$$f'^i_j(y_1,\dots,y_{i-1}) = f^{i+1}_j(1,y_1,\dots,y_{i-1}) \text{ for all } i \tag{3.22}$$

$$\mathcal{D}'_j = \{(y_1,\dots,y_{n-1}) \in \mathcal{Y}^{n-1} | (1,y_1,\dots,y_{n-1}) \in \mathcal{D}_j\} \tag{3.23}$$

By Equation (3.18)

$$\begin{aligned}
V^n((1,y_1,\dots,y_{n-1})|f_j) &= W_B(y_1|f^2_j(1)) \cdot \ldots \cdot W_B(y_{n-1}|f^n_j(1,y_1,\dots,y_{n-2})) \\
&= W_B(y_1|f'^1_j) \cdot \ldots \cdot W_B(y_{n-1}|f'^{n-1}_j(y_1,\dots,y_{n-2})) \\
&= W^n_B(y^{n-1}|f'^j)
\end{aligned}$$

and

$$V^n((0,y_1,\dots,y_{n-1})|f_j) = 0. \tag{3.24}$$

Thus

$$W^n_B(\mathcal{D}'_i|f'^j) = V^n(\mathcal{D}_i|f^j) \text{ for all } i,j \in \{1,\dots,N-1\} \tag{3.25}$$

i.e. $\{(f'_j,\mathcal{D}'_j)\}_{j\in\{1,\dots,N-1\}}$ is an $(n-1,N-1,\lambda)$ ID-code for the DMC defined by matrix W_B. $\square$

Lemma 3.3 *Let* $\lambda < 1/2$, $n \in I\!N$. *Then*

$$N_f(n+1,\lambda) \geq N_{f,W_B}(n,\lambda).$$

Proof:

Let $\{(f_j,\mathcal{D}_j)\}_{j\in\{1,\dots,N\}}$ be an (n,N,λ) ID-code for the DMC W_B.

Define a code $\{(f'_j,\mathcal{D}'_j)\}_{j\in\{1,\dots,N\}}$ of length $n+1$ with N codewords by

$$\begin{aligned}
f'^1_j &= 1 && (3.26)\\
f'^i_j(y_1,\dots,y_{i-1}) &= f^{i-1}_j(y_1,\dots,y_{i-2}) \text{ for all } i \in \{2,\dots,n+1\} && (3.27)\\
\mathcal{D}'_j &= \{(1,y_1,\dots,y_n) \in \mathcal{Y}^{n+1} | (y_1,\dots,y_n) \in \mathcal{D}_j\} && (3.28)
\end{aligned}$$

Therefore

$$V^{n+1}(\mathcal{D}'_i|f'^j) = V^n(\mathcal{D}_i|f^j) \text{ for all } i,j \in \{1,\dots,N\} \tag{3.29}$$

i.e. $\{(f'_j, \mathcal{D}'_j)\}_{j\in\{1,\dots,N\}}$ is an $(n+1, N, \lambda)$ ID-code for the DC W. $\Box$

Now we will return to the proof of Theorem 1. By Theorem (2), Lemma 3.2 and Lemma 3.3

$$\limsup_{n\to\infty} \frac{\log\log N_f(n,\lambda)}{n} \leq \limsup_{n\to\infty} \frac{\log\log N_{f,W_B}(n-1,\lambda)+1}{n} \leq \max_{x\in\mathcal{X}} H(W_B(\cdot|x)) = \alpha$$

and

$$\liminf_{n\to\infty} \frac{\log\log N_f(n,\lambda)}{n} \geq \liminf_{n\to\infty} \frac{\log\log N_{f,W_B}(n-1,\lambda)}{n} \geq \max_{x\in\mathcal{X}} H(W_B(\cdot|x)) = \alpha.$$

Thus

$$\lim_{n\to\infty} \frac{\log\log N_f(n,\lambda)}{n} = \alpha, \tag{3.30}$$

proving Theorem 1 for deterministic codes.

Since a deterministic code can also be viewed as a randomized code, this also gives the direct part of Theorem 1 for randomized codes. For the proof of the converse we need some further observations.

The second statement in Theorem (2) needs the following modification. The proof is rather lengthy, so instead of copying it we will only point out the modifications needed to adapt it for our purposes. An investigation of the proof of [2, Theorem 2 b] shows that the converse still holds if the bounds for the error probabilities are slightly weakened, namely replace Inequalities (4) and (5) in [2] by

$$\sum_{g\in\mathcal{F}_n} Q_F(g|i)\left(W^n(\mathcal{D}_i^C|g) + W^n(\mathcal{D}_j|g)\right) \leq 2\cdot\lambda \tag{3.31}$$

for all $i, j \in \{1, \ldots, N\}$, $i \neq j$. The second order rate of a series IDF-code satisfying (3.31) with $\lambda < 1/2$ is upperbounded by $\max_{P \in \mathcal{P}(\mathcal{X})} H(P \cdot W)$.

To verify this, choose $\nu > 0$ with $2 \cdot \lambda + \nu < 1$ and define $\mathcal{E}_i^*$ as in [2, p. 35]. Equation (3.31) implies that the sets $\mathcal{D}_i \cap \mathcal{E}_i^*$ are pairwise different: Suppose we have $i \neq j$ with $\mathcal{D}_i \cap \mathcal{E}_i^* = \mathcal{D}_j \cap \mathcal{E}_j^*$. This would lead to the contradiction

$$\begin{aligned} 1 &= \sum_{g \in \mathcal{F}_n} Q_F(g|i) \left(W^n((\mathcal{D}_i \cap \mathcal{E}_i^*)^C | g) + W^n((\mathcal{D}_i \cap \mathcal{E}_i^*) | g) \right) \\ &= \sum_{g \in \mathcal{F}_n} Q_F(g|i) \left(W^n((\mathcal{D}_i \cap \mathcal{E}_i^*)^C | g) + W^n((\mathcal{D}_j \cap \mathcal{E}_j^*) | g) \right) \\ &\leq \sum_{g \in \mathcal{F}_n} Q_F(g|i) \left(W^n((\mathcal{D}_i)^C | g) + W^n((\mathcal{E}_i^*)^C | g) + W^n((\mathcal{D}_j) | g) \right) \\ &< 2\lambda + \nu < 1. \end{aligned}$$

The rest of the proof can be applied as it is.

Lemma 3.4 *Let $\lambda < 1/2$, $n \in \mathbb{N}$. Then*

$$N_F(n, \lambda) \leq N'_{F,W_B}(n-1, \lambda) \left\lceil \frac{1}{1 - 2 \cdot \lambda} \right\rceil.$$

where $N'_{F,W_B}(n-1, \lambda)$ is the maximal size of a randomized identification code with the modified error probability defined by Equation (3.31) being less than λ.

Proof:

Let $\{(Q(\cdot|j), \mathcal{D}_j)\}_{j \in \{1,\ldots,N\}}$ be a randomized (n, N, λ) IDF-code for the DC V. Let for all i $p_i = W^n(\mathcal{D}_i|g')$ and $q_i = \sum_{g \in \mathcal{F}_n | g^1 = 2} Q(g|i)$, where g' is any encoding function with $g'^1 = 2$. There exists a subset $I \subset \{1, \ldots, N\}$ of cardinality $|I| \geq \frac{N}{\lceil \frac{1}{1-2\cdot\lambda} \rceil}$ with $|p_i - p_j| \leq 1 - 2 \cdot \lambda$ for all $i, j \in I$. This leads to the inequality

$$1 - p_i + p_j \geq 2 \cdot \lambda \tag{3.32}$$

and thus

$$\sum_{G\in\mathcal{F}_N:g^1=2}\big(V^n(\mathcal{D}_i^C)+V^n(\mathcal{D}_j)\big)\geq 2\cdot\lambda\cdot q_i \tag{3.33}$$

and

$$\sum_{G\in\mathcal{F}_N:g^1=1}\big(V^n(\mathcal{D}_i^C)+V^n(\mathcal{D}_j)\big)< 2\cdot\lambda\cdot(1-q_i) \text{ for all } i,j\in I. \tag{3.34}$$

(Especially we have $q_i<1$ for all $i\in I$.) Analogue to Lemma 3.2, we define the code $\{(Q'(\cdot|_j),\mathcal{D}'_j)\}_{j\in I}$ of blocklength $n-1$ with $|I|$ codewords by

$$Q'(g'|i)=\frac{Q(g|i)}{1-q_i} \tag{3.35}$$

where $g'\in\mathcal{F}^{n-1}$ and $g\in\mathcal{F}^n$ are encoding function with $g^1=1$ and

$$g'^i(y_1,\ldots,y_{i-1})=g^{i+1}(1,y_1,\ldots,y_{i-1}) \tag{3.36}$$

$$\mathcal{D}'_j=\{(y_1,\ldots,y_{n-1})\in\mathcal{Y}^{n-1}|(1,y_1,\ldots,y_{n-1})\in\mathcal{D}_j\} \tag{3.37}$$

Since

$$\sum_{G\in\mathcal{F}_N:g^1=1}\frac{Q(\cdot|i)}{1-q_i}\big(V^n(\mathcal{D}_i^C)+V^n(\mathcal{D}_j)\big)<2\lambda \tag{3.38}$$

this code satisfies the modified error condition of Equation (3.31) for the DMC W_B. $\square$ [1]

[1] Actually one may expect that also randomized identification codes for the DC V could only have one more message than the corresponding codes for the DMC W_B (as we could show for deterministic codes), but at least the conversion of codes we have used is not sufficient to show this. Here is an an informal recipe how to construct counterexamples (a formal construction is left out because this is neither relevant for the proof of our theorem nor very instructive): Suppose we have an IDF code for the DMC W_B with the property that all distributions of input letters are equal and the decoding sets consist only of sequences that are "typical" (there will be a formal definition in the next section). Then we can build a code for the DC V with twice as many code words as follows: The first half will be the original encoding functions and decoding sets both prefixed by "1" as

Lemma 3.5 *Let W_B, P_1 and α be defined as before. Then the maximal output entropy for the DMC W_B is*

$$\max_{P \in \mathcal{P}(X)} H(PW_B) = H(P_1) = \alpha.$$

Proof: Consider a convex combination

$$P_\rho = (1-\rho)P_0 + \rho P_1, \quad \rho \in [0,1].$$

P_ρ is a non increasing probability distribution and

$$\sum_{i=1}^{k} P_\rho(i) = (1-\rho) + \rho \sum_{i=1}^{k} P_1(i) \geq \sum_{i=1}^{k} P_1(i)$$

for all $k \in \mathcal{Y}$, i.e. P_ρ majorizes P_1, thus by Schur-concavity of entropy $H(P_\rho) < H(P_1)$. □

Using our modified Version [2, Theorem 2b] we get therefore the desired result for randomized identification codes for our specific channel:

$$\limsup_{n\to\infty} \frac{\log\log N_F(n,\lambda)}{n} \leq \limsup_{n\to\infty} \frac{\log\log N'_{F,W_B}(n-1,\lambda)\left\lceil \frac{1}{1-2\cdot\lambda}\right\rceil}{n} \leq \max_{P\in\mathcal{P}(X)} H(PW_B) = \alpha$$

and

$$\liminf_{n\to\infty} \frac{\log\log N_F(n,\lambda)}{n} \geq \liminf_{n\to\infty} \frac{\log\log N_f(n,\lambda)}{n} \geq \max_{x\in\mathcal{X}} H(W_B(\cdot|x)) = \alpha.$$

in the previous lemma. For the other half, choose a suitable distribution on the original encoding function and any function that sends "2" as the first letter in such a way, that the errors of second kind with respect to decoding sets of the first half are little smaller than λ. For the decoding sets take the sets of the first half and remove some sequences to get the error of second kind for encoding functions of the first half are small enough. Then add some output sequences that are "typical" for encodings that start with a "2" to get the probability for errors of first kind right.

4 The deterministic identification capacity for noiseless channels with noisy feedback

In this section we improve the upper bound for deterministic ID-codes from [6] for the case of noiseless channels with noisy feedback. We show that our improved bound coincides with the lower bound of [6] for this type of channels.

Definition 4.1 *A channel with noisy feedback is a quadruple*

$$\{\mathcal{X}, W, \mathcal{Y}, \mathcal{Z}\}. \tag{4.1}$$

Here $\mathcal{X}$ is the input alphabet, $\mathcal{Y}$ and $\mathcal{Z}$ are the output alphabets for the receiver respectively the feedback and

$$W = (W(y,z|x))_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} \tag{4.2}$$

is a stochastic matrix. To avoid trivialities, all alphabets are assumed to contain at least two elements. We want to consider finite alphabets, so we can also assume that all alphabets are the set of natural numbers less or equal to the given cardinality.

The channel is assumed to be memoryless, thus the transmission probabilities for sequences $x^n = (x_1, \dots, x_n)$, $y^n = (y_1, \dots, y_n)$ and $z^n = (z_1, \dots, z_n)$ of length n are given by

$$W^n(y^n, z^n|x^n) = \prod_{t=1}^{n} W(y_t, z_t|x_t) \tag{4.3}$$

Now an encoding of message $j \in \{1, \dots N\}$ is a vector-valued function $f_j = [f_j^1, \dots f_j^n]$, where $f_j^i : \mathcal{Z}^{i-1} \to \mathcal{X}$. f_j^1 is an element of $\mathcal{X}$. If the message is j and the sender has received $z_1, ..., z_{j-1}$ as feedback sequence, he sends $f_j^i(z_1, ..., z_{j-1})$ as the ith letter. The set of all functions of this kind is again denoted by $\mathcal{F}_n$.

An encoding function f determines the joint distribution of the output random variables $(Y_1, \dots, Y_n)$ and the feedback random variables $(Z_1, \dots, Z_n)$ by the following formula:

$$P(Y^n = y^n, Z^n = z^n|f) = W^n(y^n, z^n|f) = \prod_{t=1}^{n} W\left(y_t, z_t|f^t\left(z_1, \ldots, z_{t-1}\right)\right) \tag{4.4}$$

Definition 4.2 *A deterministic* (n, N, λ) *IDF code for* W *is a family of pairs*

$$\{(f_i, \mathcal{D}_i) : i \in \{1, \ldots, N\}\} \tag{4.5}$$

with

$$f_i \in \mathcal{F}_n, \quad \mathcal{D}_i \subset \mathcal{Y}^n \tag{4.6}$$

and

$$W^n(\mathcal{D}_i|f_i) \geq 1 - \lambda, \quad W^n(\mathcal{D}_j|f_i) < \lambda \text{ for all } i, j \in \{1, \ldots, N\}, i \neq j \tag{4.7}$$

Let us denote by $N_f^*(n, \lambda)$ the maximal N such that an (n, N, λ) code for W exists.

The following result of Ahlswede and Zhang in [6] gives bounds on the growth of $N_f^*(n, \lambda)$.

Theorem AZ *If the transmission capacity* C *of* W *is positive, then we have for all* $0 < \lambda < 1/2$

$$\liminf_{n \to \infty} n^{-1} \log \log N_f^*(n, \lambda) \geq \max I(Z \wedge U|X), \tag{4.8}$$

where the maximum is taken over all random variables (X, Y, Z, U) *with joint distribution* $P_{XYZU}(x, y, z, u) = p(x)W(y, z|x)q(u|x, z)$ *under the constraint*

$$I(U \wedge Z|XY) \leq I(X \wedge Y). \tag{4.9}$$

Furthermore

$$\limsup_{n \to \infty} n^{-1} \log \log N_f^*(n, \lambda) \leq \min\{\max I(XZ \wedge Y), \max H(Z|X)\}, \tag{4.10}$$

where the maximum is taken over all random variables (X, Y, Z) *with joint distribution* $P_{XYZ}(x, y, z) = p(x)W(y, z|x)$.

We want to consider channels where the channel from the sender to the receiver is noiseless, i.e. $\mathcal{X} = \mathcal{Y}$ and $W(y, z|x) = 0$ for all $y \neq x$. We will in this case identify W with the matrix with entries $W(z|x) = W(x, z|x)$.

Theorem 3 *If* W *is a noiseless channel with noisy feedback, then for all* $\lambda \in (0, 1/2)$

$$\lim_{n\to\infty} n^{-1} \log\log N_f^*(n, \lambda) = \max\{\min\{H(X), H(Z|X)\}\} \tag{4.11}$$

where the maximum is taken over all random variables X, Z *with joint distribution of the form*

$$P_{XZ}(x, z) = p(x)W(z|x) \tag{4.12}$$

For $z^n \in \mathcal{Z}^n$ and $f \in \mathcal{F}^n$ let us denote

$$f(z^n) = \big(f^1, f^2(z_1), \ldots, f^n(z_1, \ldots, z_{n-1})\big)$$

Since for a noiseless channel input and output random variables are equal with probability 1, for given $z^n \in \mathcal{Z}^n$ $y^n = f(z^n)$ is by Equation (4.4) the unique sequence with $P_{Y^n Z^n}(y^n, z^n) > 0$ and thus the output random variable Y^n has distribution

$$P_{Y^n}(y^n) = \sum_{z^n \in f^{-1}(y^n)} P_{Z^n}(z^n) = \sum_{z^n \in f^{-1}(y^n)} \prod_{t=1}^{n} W\left(z_t | f(z_1, \ldots, z_{t-1})\right) \tag{4.13}$$

Here $f^{-1}(y^n)$ is the inverse image of y^n under the encoding function f, i.e.

$$f^{-1}(y^n) = \{z^n \in \mathcal{Z}^n | f(z^n) = y^n\}. \tag{4.14}$$

Nevertheless, we want to distinguish between the input and output random variables to stress whether we are considering feedback sequences produced by certain input sequences or output sequences as a function of feedback sequences.

The following lemma states that the output distributions generated by the encoding functions of an IDF-code must differ significantly on the decoding sets. Thus we can get an upper bound on the size of the code by *upper-bounding the number of significantly different output distributions that can be generated.*

Lemma 4.1 *Let* $\{(f_i, \mathcal{D}_i) : i \in \{1, \dots, N\}\}$ *be an* (n, N, λ) *IDF-code for* W *with* $\lambda \in (0, 1/2)$. *Set* $\delta = 1/4 - \lambda/2$. *Then*

$$|W(\mathcal{D}_i|f_i) - W(\mathcal{D}_i|f_j)| > \delta \text{ for all } i, j \in \{1, \dots, N\}, i \neq j. \tag{4.15}$$

Proof: Assume Equation (4.15) is false for i, j. Then

$$\frac{1}{2} < 1 - \lambda \leq W(\mathcal{D}_i|f_i) \leq W(\mathcal{D}_i|f_j) + \delta < \lambda + \delta < \frac{1}{2} \quad . \qquad \Box \tag{4.16}$$

Now for each given $X^n = x^n$ we have to split $\mathcal{Z}^n$ into one set of negligible probability and one set which contains only sequences of almost the same probability. The second set will be the *$W - \delta$-typical sequences under the condition x^n*.

Definition 4.3 **($W - \delta$-typical sequences)** *Let* $W : \mathcal{X} \to \mathcal{Z}$ *be a stochastic matrix, let* $x^n \in \mathcal{X}^n$. $z^n \in \mathcal{Z}^n$ *is called* $W - \delta$*-typical under the condition* x^n *if for all* $x \in \mathcal{X}, z \in \mathcal{Z}$

$$\frac{1}{n} |{<} x, z | x^n, z^n {>} - {<} x | x^n {>} \cdot W(z|x)| \leq \delta \tag{4.17}$$

and additionally $< x, z|x^n, z^n >= 0$ *if* $W(z|x) = 0$. *The set of all* $W - \delta$*-typical sequences under the condition* x^n *is denoted by* $T_{W,\delta}(x^n)$. *If* z^n *is not typical, we call it* atypical.

In the sequel we will for sequences of length n choose $\delta_n = n^{-1/3}$ and refer to $W - \delta_n$-typical sequences simply as W-typical sequences and denote them by $T_W(x^n)$. The following lemma is the so–called *joint asymptotic equipartition property (joint AEP)* for our model. It is an analogue to the law of large numbers in probability theory.

Lemma 4.2 *For fixed $\mathcal{X}$ and $\mathcal{Z}$, for every stochastic matrix $W : \mathcal{X} \to \mathcal{Z}$ there exists a nullsequence $(\epsilon_n)_{n\in I\!N}$, such that for every distribution p on $\mathcal{X}$ and for all $x^n \in T_p$*

$$W^n(T_W(x^n)|x^n) \geq 1 - \epsilon_n \tag{4.18}$$

and

$$|T_W(x^n)| \leq 2^{n(H(Z|X)+\epsilon_n)} \tag{4.19}$$

Furthermore if $z^n \in T_W(x^n)$, then

$$\left|\frac{1}{n}\log W(z^n|x^n) + H(Z|X)\right| \leq \epsilon_n \tag{4.20}$$

where X, Z are random variables with joint distribution

$$P_{XZ}(x,z) = p(x)W(z|x). \tag{4.21}$$

Proof: These results are folklore in information theory, we supply some justification and references for the benefit of the non-expert reader. The first inequality is the second assertion in [10, Lemma 2.12], the second is a rewriting of the upper bound given by the second inequality in [10, Lemma 2.13]:

For $x^n = (x_1, \ldots, x_n)$ let $Z_1, \ldots, Z_n$ be independent random variables where Z_i has distribution $P_{Z_i} = W(\cdot|x_i)$. Then for all pairs $(x,z) \in \mathcal{X} \times \mathcal{Z}$ the random variable $< x, z|x^n, Z^n >$ has binomial distribution with expected value $< x|x^n > W(z|x)$. The variance is

$$<x|x^n> W(z|x)\left(1 - W(z|x)\right) \leq \frac{<x|x^n>}{4} \leq \frac{n}{4}. \tag{4.22}$$

We apply Chebyshev's inequality and get

$$P\left(n^{-1}\left|<x,z|x^n,z^n> - <x|x^n> \cdot W(z|x)\right| > \delta\right) \leq \frac{1}{4k\delta^2}. \tag{4.23}$$

Thus we get from the union bound that

$$W^n(T_W(x^n)|x^n) \geq 1 - \frac{|\mathcal{X}||\mathcal{Z}|}{4n\delta^2}. \tag{4.24}$$

The second inequality is a rather straightforward consequence of Lemma (2.2) and Corollary (2.1). For the third inequality notice that

$$W(z^n|x^n) = \prod_{(x,z)\in\mathcal{X}\times\mathcal{Z}:W(z|x)>0} W(z|x)^{<x,z|x^n,z^n>} \tag{4.25}$$

Thus

$$\begin{aligned} -\log W(z^n|x^n) &= -\sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}:W(z|x)>0} < x, z|x^n, z^n > \log W(z|x) \\ &\geq -\sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}:W(z|x)>0} (np(x)W(z|x) + \delta_n) \log W(z|x) \\ &= nH(Z|X) - n\delta_n \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}:W(z|x)>0} \log W(z|x). \end{aligned} \tag{4.26}$$

Analogously we have

$$-\log W(z^n|x^n) \leq nH(Z|X) + n\delta_n \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}:W(z|x)>0} \log W(z|x) \tag{4.27}$$

Since $\sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}:W(z|x)>0} \log W(z|x)$ is fixed, this proves the third inequality for suitable ϵ_n. $\square$

To approximate the output probability for a given sequence y^n by Equation (4.18) we can restrict ourselves to sequences z^n which are jointly typical under the condition $x^n = y^n$.

Since the probability of a decoding set is determined by the sizes of intersections of the inverse images of the decoding set with the different types of feedback sequences, we can then estimate the number of different encoding functions by the combination of sizes of intersections.

Lemma 4.3 *For every $0 < \lambda < \lambda' < 1/2$, n sufficiently large depending on λ and λ' and every (n, N, λ)-IDF-code $\{(f_i, \mathcal{D}_i) : i \in \{1, \dots, N\}\}$ there exists an $(n+1, N, \lambda')$-IDF-code $\{(f_i', \mathcal{D}_i') : i \in \{1, \dots, N\}\}$ with*

$$f'(z^{n+1}) \in \bigcup_{i=1}^{N} \mathcal{D}'_i \Rightarrow z^n \in T_W\left(f(z^n)\right) \tag{4.28}$$

where $z^{n+1} = (z^n, 1)$.

Proof: For $i \in \{1, \ldots, n\}$ and $f_i = (f_i^1, \ldots, f_i^n)$ let

$$f'_i = (f_i^1, \ldots, f_i^{n+1}) \tag{4.29}$$

where

$$f_i^{n+1}(z^n) = \begin{cases} 1 \text{ if } z^n \in T_W(f(z^n)) \\ 2 \text{ otherwise} \end{cases} \tag{4.30}$$

and

$$\mathcal{D}'_i = \{(y_1, \ldots, y_n, 1) : (y_1, \ldots, y_n) \in \mathcal{D}_i\} \quad . \tag{4.31}$$

Obviously this code satisfies Implication (4.28) and Inequality (4.18) guarantees that the error probabilities are smaller than $\lambda + \epsilon_n < \lambda'$ for n sufficiently large. □

From now, we want to consider only IDF-codes respectively encoding functions that satisfy (4.28). We will denote the maximal size of such a code by $N'_f(n, \lambda)$. We introduce a new possibility for errors of first kind by declaring an error, if the feedback sequence is not jointly typical with the first n letters of the input sequence, but by the previous lemma, there is no loss in rate and the decrease in error performance is not significant. By considering only *typical* sequences, the sets also boil down to the required sizes. The next lemma shows that the number of feedback types that "typically" occur is at most polynomial in n.

Lemma 4.4 *Let P_X be a distribution on $\mathcal{X}$. The number of types of feedback sequences z^n for which there is an input sequence x^n of type p such that z^n is $W-\delta$-typical under the condition x^n is upperbounded by $(2n\delta+1)^{|\mathcal{Z}|}$.*

Proof: The number of occurrences of each letter can take one of at most $(2n\delta + 1)$ values. $\square$

We will now consider output sequences of fixed n-type p on $\mathcal{X}$ and upperbound the number of different probabilities of decoding sets we can generate on this set under the assumption that we produce an output sequence that is outside any decoding set, if the feedback sequence is not jointly typical with the input sequence:

Lemma 4.5 *Let p be an n-type on $\mathcal{X}$. Let $B(p,n)$ denote the number of different values that can occur as the output probability of a decoding set $\mathcal{D}$ with $\mathcal{D} \subset T_p$ under an encoding function that maps atypical feedback sequences outside $\mathcal{D}$. Then*

$$B(p,n) \leq \binom{2^{nH(X)} + 2^{n(H(Z|X)+\epsilon_n)}}{2^{nH(X)}}^{(2n\delta+1)^{|\mathcal{Z}|}} \tag{4.32}$$

where (X,Z) is a pair of random variables with joint distribution $P_{XZ}(x,z) = p(x)W(z|x)$.

Proof: There are at most $2^{nH(X)}$ output sequences of type p and for each of this sequences there are at most $2^{n(H(Z|X)+\epsilon_n)}$ feedback sequences that are jointly typical with that given sequence and have the same joint type. So if we condition on the event that the feedback sequence is of that joint type, the probability of that output sequence is l-typical, $l < 2^{n(H(Z|X)+\epsilon_n)}$, because it depends only on the number of feedback sequences that are mapped on x^n. Since the number of joint types that are $W-\delta$-jointly typical is less than $(2n\delta + 1)$ this proves the lemma, because the number of l-types on a set $\mathcal{U}$ is nondecreasing in both l and the cardinality of $\mathcal{U}$. $\square$

To get a bound on the number of different decoding functions, we have to consider all possible input types, but this number is only polynomial in n, so we have to upperbound $n^{-1} \log\log B(p,n)^{(n+1)^{|\mathcal{X}|}}$:

To simplify notation, a non-integer argument for factorials or bounds of sums and products is understood to be rounded down.

Lemma 4.6

$$\limsup_{n\to\infty} n^{-1} \log\log B(p,n)^{(n+1)^{|\mathcal{X}|}} \leq H(X) \tag{4.33}$$

$$\limsup_{n\to\infty} n^{-1} \log\log B(p,n)^{(n+1)^{|\mathcal{X}|}} \leq H(Z|X) \tag{4.34}$$

Proof:

$$
\begin{aligned}
B(p,n) &\le \left(\frac{\left(2^{nH(X)}+2^{n(H(Z|X)+\epsilon_n)}\right)!}{\left(2^{nH(X)}\right)!\left(2^{n(H(Z|X)+\epsilon_n)}\right)!}\right)^{(2n\delta+1)^{|\mathcal{Z}|}} \\
&\le \left(\frac{\prod_{i=1}^{2^{n(H(Z|X)+\epsilon_n)}}\left(2^{nH(X)}+i\right)}{\left(2^{n(H(Z|X)+\epsilon_n)}\right)!}\right)^{(2n\delta+1)^{|\mathcal{Z}|}} \\
&= \left(\prod_{i=1}^{2^{n(H(Z|X)+\epsilon_n)}}\frac{2^{nH(X)}+i}{i}\right)^{(2n\delta+1)^{|\mathcal{Z}|}} \\
&= \left(\prod_{i=1}^{2^{n(H(Z|X)+\epsilon_n)}}\left(1+\frac{2^{nH(X)}}{i}\right)\right)^{(2n\delta+1)^{|\mathcal{Z}|}}
\end{aligned}
$$

Now let us use an input distribution p that maximizes $\min(H(X),H(Z|X))$.

$$
\begin{aligned}
\limsup_{n\to\infty}\frac{1}{n}\log\log B(p,n) &\le \limsup_{n\to\infty}\frac{1}{n}\log\log\left(\prod_{i=1}^{2^{n(H(Z|X)+\epsilon_n)}}\left(1+\frac{2^{nH(X)}}{i}\right)\right)^{(2n\delta+1)^{|\mathcal{Z}|}} \\
&= \limsup_{n\to\infty}\frac{1}{n}\log\sum_{i=1}^{2^{n(H(Z|X)+\epsilon_n)}}\log\left(1+\frac{2^{nH(X)}}{i}\right) \\
&\le \limsup_{n\to\infty}\frac{1}{n}\log\left(2^{n(H(Z|X)+\epsilon_n)}nH(X)\right) \\
&= H(Z|X)
\end{aligned}
$$

The other inequality can be proven analogously, since the expression on the right hand side of Equation 4.32 is symmetrical in $H(X)$ and $H(Z|X)+\epsilon_n$.

So for each type p the second order rate of a code that uses only input words of this type is bounded from above by the minimum of $H(X)$ and $H(Z|X)$. Because there are only polynomial in n many input types, this gives the desired result, namely let $B(n)=\max\{B(p,n) : p \text{ is n-type}\}$ and let $0<\lambda<\lambda'<1/2$. Then

$$
\limsup_{n\to\infty}\frac{1}{n}\log\log N_f^*(n,\lambda) = \limsup_{n\to\infty}\frac{1}{n}\log\log N_f'(n,\lambda')
$$

$$
\begin{aligned}
&\leq \limsup_{n\to\infty} \frac{1}{n} \log\log B(n)^{(n+1)^{|\mathcal{Z}|}} \\
&\leq \max_X \min\{H(X), H(Z|X)\}
\end{aligned}
$$

To complete the proof of Theorem 3, we show that this coincides with the lower bound of Theorem AZ. We consider noiseless main channels, thus Inequality (4.9) reduces to

$$I(U \wedge Z|X) \leq H(X) \tag{4.35}$$

Let U be a random variable on $\mathcal{Z}$. Let

$$q'(u|x,z) = \begin{cases} 1; & \text{if } u = z \\ 0; & \text{otherwise} \end{cases} \tag{4.36}$$

and

$$q''(u|x,z) = \begin{cases} 1; & \text{if } u = 1 \\ 0; & \text{otherwise} \end{cases} \tag{4.37}$$

If for the input distribution which maximizes $\min\{H(X), H(Z|X)\}$ (if there are several, then take any of them) it holds that $H(X) \geq H(Z|X)$, then we set $q = q'$ and get $Z = U$ with probability 1. Therefore $I(U \wedge Z|X) = H(Z|X)$. Otherwise choose a convex combination $q = \alpha q' + (1-\alpha)q''$ such that

$$I(U \wedge Z|X) = H(X). \tag{4.38}$$

On the boundary point for parameter 1 of the convex combination we have again $U = Z$, which gives $I(U \wedge Z|X) = H(Z|X) > H(X)$ by assumption, parameter 0 gives $I(U \wedge Z|X) = 0$. Thus by continuity of mutual information and mean value theorem such an α exists.

In both cases we get from Theorem AZ that

$$\liminf_{n\to\infty} n^{-1} \log\log N_f^*(n,\lambda) \geq \max\min\{H(X), H(Z|X)\} \tag{4.39}$$

which concludes the proof of Theorem 3.

Remarks:

1. The upper bound of Theorem AZ applied to our case states that

$$\limsup_{n\to\infty} \log\log N_f^*(n,\lambda) \leq \min\{\max H(X), \max H(Z|X)\} \tag{4.40}$$

To see that *Theorem 3 really improves this bound*, consider the following channel: Let $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{1,2\}$. Let

$$W(z|x) = \begin{cases} 1/2; & \text{if } x = 1 \\ 1; & \text{if } x = 2 \text{ and } z = 1 \\ 0; & \text{if } x = 2 \text{ and } z = 2 \end{cases} \tag{4.41}$$

If one uses $p = (\beta, 1-\beta)$ as input distribution, then $H(X) = h(\beta)$ and $H(Z|X) = \beta$ where $h(\beta) = -\beta\log\beta - (1-\beta)\log(1-\beta)$ is the binary entropy. So Inequality (4.40) gives

$$\limsup_{n\to\infty} n^{-1}\log\log N_f^*(n,\lambda) \leq 1, \tag{4.42}$$

while Theorem 3 gives

$$\limsup_{n\to\infty} n^{-1}\log\log N_f^*(n,\lambda) = \max_{\beta\in[0,1]} \min\{\beta, h(\beta)\} \approx 0.7729. \tag{4.43}$$

2. Our theorem can be viewed as a result about identification codes with a restriction on the input distributions that are allowed. Since the sender is not allowed to randomize in the encoding, the only possibility to generate randomness are the feedback letters. The sender now has to use input distributions that are simultaneously good for transmission and for the generation of randomness via the feedback channel. We conjecture that this is also true for feedback channels with noise:

   Conjecture 1 *Let W be an channel with noisy feedback. Then for all $0 < \lambda < 1/2$*

   $$\limsup_{n\to\infty} n^{-1}\log\log N_f^*(n,\lambda) \leq \max\min\{I(XZ\wedge Y), H(Z|X)\} \tag{4.44}$$

   where the maximum is taken over all random variables (X,Y,Z) with joint distribution $P_{XYZ}(x,y,z) = p(x)W(y,z|x)$.

However, note that this conjecture in general does not coincide with the lower bound of Theorem AZ (although it does in more cases than the old bound did), so even if it was settled, determination of the capacity would still be an open problem .

References

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

[2] Ahlswede, R. and Dueck, G., "Identification in the presence of feedback—a discovery of new capacity formulas", IEEE Trans. Inform. Theory, vol. 35, 30–36, 1989.

[3] Ahlswede, R. and Verboven, B., "On identification via multiway channels with feedback", IEEE Trans. Inform. Theory, vol. 37, 1519–1526, 1991.

[4] Han, T. S. and Verdú, S., "New results in the theory of identification via channels", IEEE Trans. Inform. Theory, vol. 38, 14–25, 1992.

[5] Han, T. S. and Verdú, S., "Approximation theory of output statistics", IEEE Trans. Inform. Theory, vol. 39, 752–772, 1993.

[6] Ahlswede, R. and Zhang, Z., "New directions in the theory of identification", SFB 343 "Diskrete Strukturen in der Mathematik", Bielefeld, Preprint 94-010, 1994.

[7] Ahlswede, R., "General theory of information transfer", SFB 343 "Diskrete Strukturen in der Mathematik", Bielefeld, Preprint 97-118, 1997, submitted to IEEE.

[8] Ahlswede, R. and Csiszár, I., "Common randomness in information theory and cryptograpy. II. CR capacity" IEEE Trans. Inform. Theory, vol. 44, 225–240, 1998.

[9] Cover, T. M. and Thomas, J. A., "Elements of information theory", Wiley Ser. Telecom., New York, 1991.

[10] Cziszar, I. and Korner, J.: "Information theory: Coding theorems for discrete memoryless systems", Academic Press, New York, 1981.