General theory of information transfer

Rudolf Ahlswede

Universität Bielefeld
Fakultät für Mathematik
Postfach 100131
33501 Bielefeld
Germany

Abstract

We report on ideas, problems and results, which occupied us during the past decade and which seem to extend the frontiers of information theory in several directions. The main contributions concern information transfer by channels. There are also new questions and some answers in new models of source coding. While many of our investigations are in an explorative state, there are also hard cores of mathematical theories. In particular we present a unified theory of information transfer, which naturally incorporates Shannon's theory of information transmission and the theory of identification in the presence of noise as extremal cases. It provides several novel coding theorems. On the source coding side we introduce the concept of identification entropy. Finally we are led beyond information theory to new concepts of solutions for probabilistic algorithms.

Footnote: Some of the ideas and results have been presented at the Tenth Symp. on Inf. Theory in the Benelux, Houthalen, Belgium May 25–26, 1989, at the Inf. Theory Meeting in Oberwolfach, April 1992, at an IEEE Workshop on Inf. Theory in San Salvador, Brazil, June 1992, at the ISIT San Antonio, Texas, Jan. 1993, the IEEE Workshop on Inf. Theory in Rydzyna, Poland, June 1995, and at the World Congress of the Bernoulli Society, Vienna, August 1997.

0. Contents

1 Introduction

We have included in the references several articles and books ([19], [20], [21], [23], [24]), which deal with information not just in a more or less technical engineering sense. They are meant to enlarge our horizon, stimulate our awareness of what is unknown about "information", and to bring us into the spirit for new adventures. Some questions from [23] give indications of the kind of thoughts which took us into their chains.

In the Appendix of [23] one finds the following definition or explication of the concept "communication":

"The establishment of a social unit from individuals, by the shared usage of language or signs. The sharing of common sets of rules, for various goal–seeking activities. (There are many *shades of opinions.*)"

Again in [23] on page 41 we read:

"Perhaps the most important technical development which has assisted in the birth of communication theory is that of telegraphy. With its introduction the speed of transmission of "intelligence" arose. When its economic value was fully realized, the problems of compressing signals exercised many minds, leading eventually to the concept of "quantity of information" and to theories of times and speed of signaling."

and on page 43:

"Hartley went further and defined information as the successive selection of signs or words from a given list, rejecting all "meaning" as a more subjective factor (it is the signs we transmit, or physical signs; we do not transmit their "meaning"). He showed that a message of N signs chosen from an "alphabet" or code book of S signs has S^N possibilities and that the "quantity of information" is most reasonably defined as the logarithm, that is, $H = N \log S$."

This concept of information is closely related to the idea of selection, or discrimination and therefore sometimes called selective–information. It is also at the very basis of Shannon's celebrated statistical theory of communication [1].

This theory has by now been developed into a sophisticated mathematical discipline with many branches and facetts. Sometimes more concrete engineering problems led to or gave the incentive to new directions of research and in other cases new discoveries were made by exploring inherent properties of the mathematical structures. Some of our views on the state of this theory, to which we also shall refer as the "Shannon Island", are expressed in [8].

The price for every good theory is simplification and its permanent challenge is reality.

"We live in a world vibrating with information" and in most cases we don't know how the information is processed or even what it is at the semantic and pragmatic levels. How does our brain deal with information? It is still worthwhile to read von Neumann's ideas about this [21].

Cherry writes on page of [23]:

"It is remarkable that human communication works at all, for so much seems to be against it; yet it does. The fact that it does depends principally upon the vast store of habits which one of us possess, the *imprints of all our past experiences*. With this, we can hear snatches of speech, the vague gestures and grimaces, and from this shreds of evidence we are able to make a continual series of inferences, guesses, with extra ordinary effectiveness."

We shall discuss the issue of "prior knowledge" later and we shall show that some aspects are accessible to a rigorous mathematical treatment.

There are various stimuli concerning the concepts of communication and information from the sciences, for instance from quantum theory in physics, the theory of learning in psychology [19], theories in linguistics [40], etc.

These hints give an idea of the size of the ocean around the Shannon Island.

We don't have the intention to drown in this ocean. However, since the ocean is large there ought to be some other islands. In fact there are.

Among those, which are fairly close to the Shannon Island we can see

1.) Mathematical Statistics

2.) Communication Networks

3.) Computer Storage and Distributive Computing

4.) Memory Cells

Since those islands are close there is hope that they can be connected by dams.

A first attempt to explore connections between Multi–user source coding and hypothesis testing was made in [10]. For interesting ideas about relations between Multiple–access channels and communication networks see Gallager [24]. A multitude of challenges to information theory comes from computer science. A proper frame for storage in memory cells is our abstract coding theory [8]. Our work on identification has led us to reconsider the basic assumptions of Shannon's Theory. It deals with "messages", which are elements of a *prescribed set of objects*, known to the communicators. The receiver wants to know the true message. This basic model occuring in all engineering work on communication channels and networks addresses a very special communication situation. More generally they are characterized by

(I) The questions of the receivers concerning the given "ensemble", to be answered by the sender(s)

(II) The prior knowledge of the receivers

(III) The senders prior knowledge.

Accordingly the paper starts with three parts.

It seems that the whole body of present day Information Theory will undergo serious revisions and some dramatic expansions. We open several directions of future research and start the mathematical description of communication models in great generality. For some specific problems we provide solutions or ideas for their solutions.

We continue in Part IV with (promised) capacity theorems for identification via multi–way channels. We also study identification in conjunction with transmission.

The proof of the "polynomial" weak converse is new even for the DMC.

In Part V we discuss a new direction of research on sources, which goes back to a problem of [15]: noiseless coding for multiple purposes. It culminates in a new concept: identification entropy.

Part VI concludes with striking results on the relation of identification and common randomness and a general discussion.

Part I: One sender answering several questions of receivers

2 A general communication model for one sender

To simplify matters we assume first that the noise is modelled by a discrete memoryless channel (DMC) with input (resp. output) alphabet $\mathcal{X}$ (resp. $\mathcal{Y}$) and transmission matrix W.

The goal in the classical Shannon communication theory is to transmit many messages reliably over this channel. This is done by coding. An (n, M, λ)–code is a system of pairs $\big\{(u_i, \mathcal{D}_i) : 1 \leq i \leq M\big\}$ with $u_i \in \mathcal{X}^n, \mathcal{D}_i \subset \mathcal{Y}^n$ and

$$\mathcal{D}_i \cap \mathcal{D}_{i'} = \varnothing \quad \text{for} \quad i \neq i', \tag{2.1}$$

$$W(\mathcal{D}_i^c | u_i) \leq \lambda \quad \text{for} \quad i = 1, \dots, M. \tag{2.2}$$

Given a set of messages $\mathcal{M} = \{1, \dots, M\}$, by assigning i to codeword u_i we can transmit a message from $\mathcal{M}$ in blocklength n over the channel with a maximal error probability less than λ. Notice that the underlying assumption in this classical transmission problem is that both, sender and receiver, know that the message is from a specified set $\mathcal{M}$. They also know the code. The receiver's goal is to get to know the message sent.

One can conceive of many situations in which the receiver has (or many receivers have) different goals.

A nice class of such situations can, abstractly, be described by a family $\Pi(\mathcal{M})$ of partitions of $\mathcal{M}$. Decoder $\pi \in \Pi(\mathcal{M})$ wants to know only which member of the partition $\pi = (A_1, \dots, A_r)$ contains m, the true message, which is known to the encoder.

We describe now some seemingly natural families of partitions.

Model 1: $\Pi_S = \{\pi_{sh}\}$, $\pi_{sh} = \big\{\{m\} : m \in \mathcal{M}\big\}$. This describes Shannon's classical transmission problem stated above.

Model 2: $\Pi_I = \{\pi_m : m \in \mathcal{M}\}$ with $\pi_m = \big\{\{m\}, \mathcal{M} \smallsetminus \{m\}\big\}$. Here decoder π_m wants to know whether m occured or not. This is the identification problem introduced in [16].

Model 3: $\Pi_K = \{\pi_S : |S| = K, S \subset \mathcal{M}\}$ with $\pi_S = \{S, \mathcal{M} \smallsetminus S\}$. This is an interesting generalisation of the identification problem. We call it K–identification.

This case also arises in several situations. For instance every person π_S may have a set S of K closest friends and the sender knows that one person $m \in \mathcal{M}$ is sick. All persons π_S want to know whether one of their friends is sick.

Model 4: $\Pi_R = \big\{\pi_r : \pi_r = \big\{\{1,\dots,r\}, \{r+1,\dots,M\}\big\}, 1 \le r \le M-1\big\}$. Here decoder π_r wants to know whether the true message exceeds r or not. We speak of the ranking problem.

Model 5: $\Pi_B = \{\Pi_A : A \subset \mathcal{M}\}$. Here $\pi_A = \{A, \mathcal{M} \smallsetminus A\}$ wants to know the answer to the binary question "Is m in A?".

Model 6: $\mathcal{M} = \{0,1\}^\ell$, $\Pi_C = \{\pi_t : 1 \le t \le \ell\}$ with $\pi_t = \big\{\{(x_1,\dots,x_\ell) \in \mathcal{M} : x_t = 1\}, \{(x_1,\dots,x_\ell) \in \mathcal{M} : x_t = 0\}\big\}$. Decoder π_t wants to know the t–th component of the vector valued message $(x_1,\dots,x_\ell)$.

In all these models we can consider the first (or second) order capacities. They are known for models 1, 2. It is shown in Section 4 that for models 4 and 5 the capacities equal Shannon's transmission capacity.

The most challenging problem is the general K–identification problem of model 3. Here an (n,N,K,λ)–code is a family of pairs $\big\{(Q(\cdot|i), \mathcal{D}_\pi) : 1 \le i \le N, \pi \in \Pi_K\big\}$, where the $Q(\cdot|i)$'s are PD's on $\mathcal{X}^n$, $\mathcal{D}_\pi \subset \mathcal{Y}^n$, and where for all $\pi = \{S, \mathcal{M} \smallsetminus S\}$ $\big(S \in \binom{\mathcal{M}}{K}\big)$

$$\begin{aligned} &\sum_{x^n} Q(x^n|i) W(\mathcal{D}_\pi^c|x^n) \le \lambda \ \text{ for all } \ i \in S, \\ &\sum_{x^n} Q(x^n|i) W(\mathcal{D}_\pi|x^n) \le \lambda \ \text{ for all } \ i \notin S. \end{aligned} \tag{2.3}$$

We also write $\mathcal{D}_S$ instead of $\mathcal{D}_\pi$. A coding theorem is established in Section 3.

Remark 1: Most models fall into the following category of regular transfer models. By this we mean that the set of partitions Π of $\mathcal{M}$ is invariant under all permutations $\sigma : \mathcal{M} \to \mathcal{M}$: $\pi = (A_1,\dots,A_r) \in \Pi$ implies $\sigma\pi = \big(\sigma(A_1),\dots,\sigma(A_r)\big) \in \Pi$.

Remark 2: Many of the models introduced concern bivariate partitions. More generally they are described by a hypergraph $\mathcal{H} = (\mathcal{M}, \mathcal{E})$, where decoder $E, E \in \mathcal{E}$, wants to know whether the m occured is in E or not.

Example 1: In a certain lottery a player can choose ℓ of the numbers $1,\dots,L$, say, $\{a_1,\dots,a_\ell\}$. A set $\{b_1,\dots,b_\ell\}$ of ℓ numbers is choosen at random.

Suppose that T players have chosen $\{a_1^1,\dots,a_\ell^1\},\dots,\{a_1^T,\dots,a_\ell^T\}$, resp. Every player wants to know whether he won, that shall mean, whether he has at least $\ell-1$ correct numbers: For the t–th player

$$|\{a_1^t,\dots,a_\ell^t\} \cap \{b_1,\dots,b_\ell\}| \ge \ell - 1.$$

How many bits have to be transmitted in a randomized encoding, so that every player knows with high probability, whether he won.

Example 2: Lets view the elements of $\{1, \dots, a\}^n$ as sequences of events. Historians (or observers of stockmarkets) have subsequence of events, say,

$$(t_1^1, \dots, t_{s_1}^1), \dots, (t_1^\ell, \dots, t_{s_\ell}^\ell).$$

The ℓ persons are to be informed with high probability correctly about the correct sequence of events. (Idea of bining, see [6], [7], [8]).

Example 3: In some countries 40% of the healthy men of an age–class are drafted by random selection. Every candidate wants to know with high probability correctly whether he is among them. This falls under model 6.

3 Analysis of a specific model: K–identification

3.1 A relation to standard identification

Recall the definition of an (n, N, K, λ)–code given in Section 2. For reasons, which become apparent soon, we assume K to grow exponentially in the blocklength n, that is,

$$K = 2^{\kappa \cdot n}, \tag{3.1}$$

where K is a first order rate.

As for the standard identification problem ($K = 1, \kappa = 0$) N can grow double exponentially, that is,

$$N = 2^{2^{Rn}}, R > 0, \tag{3.2}$$

where R is a second order rate.

The pair (R, κ) is achievable, if for any $\lambda > 0, \delta > 0$ and all sufficiently large n $\big(n, 2^{2^{(R-\delta)n}}, 2^{(\kappa-\delta)n}, \lambda\big)$ –codes exist.

Proposition 1. For every DMC the set $\mathcal{K}$ of all achievable rate pairs contains

$$\big\{(R, \kappa) : 0 \le R, \kappa; R + 2\kappa \le C_{sh}\big\},$$

where C_{sh} is Shannon's familiar capacity of the DMC.

Proof: In [16] the achievable triples (R, η_1, η_2) of second order rate R and error exponents η_1, η_2 have been investigated. Theorem 2 of [16] completely characterizes the set of achievable pairs (R, η_2) in the limit $\eta_1 \to 0$ as follows:

$$\lim_{\eta_1 \to 0} \big\{(R, \eta_2) : (R_1, \eta_1, \eta_2) \text{ is achievable}\big\} = \big\{(R, \eta_2) : R \le C_{sh} - 2\eta_2\big\}. \tag{3.3}$$

Now, any identification code $\big\{(Q_i, \mathcal{D}_i) : 1 \le i \le N\big\}$ has an associated K–identification code $\big\{\mathcal{Q}_i, \mathcal{D}_S : 1 \le i \le N, S \in \binom{\mathcal{N}}{K}\big\}$, where

$$\mathcal{D}_S = \bigcup_{i \in S} \mathcal{D}_i. \tag{3.4}$$

If the original code has parameters $N = 2^{2^{Rn}}, 2^{-\eta_1 n}, 2^{-\eta_2 n}$, then by (3.3) for $\eta_1 < \delta$ the pairs of rates (R, κ) with

$$R \leq C_{sh} - \delta - 2\kappa \tag{3.5}$$

are achievable (for any $\delta > o$) for K–identification.

Remark 3: Especially, for $\kappa = 0$, Proposition 1 gives the standard Coding Theorem for Identification.

For second order rate $R = 0$, we can have $N = K$ and the first order κ is unbounded! If not specified explicity otherwise we always assume $R > 0$.

There is a very important connection to r–cover–free families. A family of sets $\mathcal{F}$ is called *r–cover–free* if $A_0 \not\subset A_1 \cup A_2 \cup \cdots \cup A_r$ holds for all distinct $A_0, A_1, \ldots, A_r \in \mathcal{F}$. Let $M(n, r)$ denote the maximum cardinality of such an $\mathcal{F}$ over an n–element underlying set. This notion was introduced in terms of superimposed codes in [49], where for suitable constants c_1, c_2 the inequalities

$$\frac{c_1}{r^2} \leq \frac{\log M(n,r)}{n} \leq \frac{c_2}{r}$$

were proved. This result was rediscovered several times. In [50], with a rather complicated proof, the upper bound was improved to

$$\frac{\log M(n,r)}{n} \leq 2\frac{\log r + O(1)}{r^2}.$$

After the purely combinatorial proof of [52] by a simpler argument (implicitly contained in [51]) the slightly weaker bound

$$\frac{\log M(n,r)}{n} \leq 4\frac{\log r + O(1)}{r^2}$$

was obtained in [53]. Let $a = |\mathcal{X}|$. With the replacements $r \to a^{\kappa n}$, $n \to a^n$ we obtain

$$\frac{\log M(a^n, a^{\kappa n})}{a^n} \leq c \cdot \frac{\log a^{\kappa n}}{a^{2\kappa n}}$$

and thus

$$R_n \triangleq \frac{\log\log M(a^n, a^{\kappa n})}{n} \leq (1 - 2\kappa)\log a + o(1).$$

In particular, for $a = 2$, $R \leq 1 - 2\kappa$.

This raises the question of optimality of the bound in Proposition 1. For its answer one needs a suitable bound for r–cover–free uniform families $\mathcal{F}$ of subsets, each of cardinality ℓ with $\frac{1}{n}\log \ell = I$, the mutual information. However, the existing bounds are too rough!

Technically very simple is the case of K–identification for noiseless channels, if we require the error of first kind to be 0, because thus $\mathcal{D}_S$ equals the union of the support sets $\mathcal{D}_i$ for the random strategies $P_i(i \in S)$ and to just obtain error probability of second kind to be *less than* 1, necessarily $\mathcal{D}_j \not\subset \mathcal{D}_S$ for $j \notin S$. Now the bound on $a^{\kappa n}$–cover–free families is applicable.

Proposition 2. In the noiseless case and for zero error probability of first kind the bound in Proposition 1 is tight.

Notice that in our definition of achievability of a pair (R, κ) we required the existence of (n, N, K, λ)–codes for all small $\lambda > 0$ and n large. It is very convenient to introduce the concept of $\lambda(n)$–achievable pairs (R, κ) by the property that for all large n $(n, N, K, \lambda(n))$–codes exist. Moreover (R, κ) shall be called polynomially achievable, if for a polynomial q, $\lambda(n) = q^{-1}_{(n)}$, and n large $(n, N, K, \lambda(n))$–codes exist. Similarly (R, κ) is exponentially achievable, if $\varepsilon > 0$ is $\lambda(n)$–achievable for $\lambda(n) = \ell^{-\varepsilon n}$, where $\varepsilon > 0$.

Correspondingly we speak about $\mathcal{K}_{\lambda(n)}$, the region $\mathcal{K}_{\text{pol}}$ of polynomially achievable rate pairs and the region $\mathcal{K}_{\text{exp}}$ of exponentially achievable rate pairs.

This terminology is consistent with the terminology for converses, which we introduced in (12.2) – (12.5). Further qualifications for several kinds of probabilities are given when needed. Actually for many coding problems several regions coincide. However, as long as we don't know this it is convenient to have this flexible language.

3.2 An equivalence of two coding problems

Let us start with an (n, N, K, λ)–code $\left\{Q_i, \mathcal{D}_S : 1 \leq i \leq N, S \in \binom{\mathcal{N}}{K}\right\}$.

We say that S is λ^*–*decodable*, if there is a partition $\mathcal{E}_S = \{E_s : s \in S\}$ of $\mathcal{D}_S$ such that

$$\sum_{x^n} W(E_s|x^n) Q_s(x^n) \geq 1 - \lambda^* \quad \text{for all} \quad s \in S. \tag{3.6}$$

If for an (n, N, K, λ)–code every $S \in \binom{\mathcal{N}}{K}$ is λ^*–decodable, then we speak of an $(n, N, K, \lambda, \lambda^*)$–code. $\mathcal{K}^*$ denotes the set of pairs of rates for such codes with $R > 0$, which are achievable for every $\lambda > 0, \lambda^* > 0$.

Equivalence Theorem 1 *For every DMC*

$$\mathcal{K}_{\text{pol}} \subset \mathcal{K}^* \subset \mathcal{K}.$$

Proof: Obviously, $\mathcal{K}^* \subset \mathcal{K}$. Furthermore, the rate pairs in $\mathcal{K}$ are achievable for every polynomial $\lambda(n)$. We show now that an (n, N, K, λ)–code with $N = 2^{2^{Rn}}, K = 2^{\lceil \kappa n \rceil}, \lambda(n)$ can be transformed in an $\big(n, N, K, \lambda(n), \lambda^*(n)\big)$–code with

$$\lambda^*(n) \leq \lceil \kappa n \rceil \lambda(n). \tag{3.7}$$

Fix any $S \in \binom{\mathcal{N}}{K}$ and label its elements by the mapping

$$\varphi : S \to \{0, 1\}^{\lceil \kappa n \rceil}. \tag{3.8}$$

Then define for $j = 1, 2, \ldots, \lceil \kappa n \rceil$

$$\underline{S}_j = \big\{ s \in S : \varphi(s)_j = 1 \big\} \tag{3.9}$$

and

$$S_j = \underline{S}_j \cup \overline{S} \;\text{ for }\; \overline{S} \subset \mathcal{N} \smallsetminus S, |\overline{S}| = \frac{1}{2}K. \tag{3.10}$$

The S_j's are elements of $\binom{\mathcal{N}}{K}$ and the $\underline{S}_j$'s form a separating system on S : for every $s, s' \in S$ we have for some j

$$s \in \underline{S}_j \;\text{ and }\; s' \notin \underline{S}_j. \tag{3.11}$$

Introduce now the function $\varepsilon_j : S \to \{0, 1\}$ by

$$\varepsilon_j(s) = \begin{cases} 1 & \text{if } \; s \in S_j \\ 0 & \text{if } \; s \in S_j^c \end{cases}$$

and use the convention $A^1 = A$ and $A^0 = A^c$.

Then the sets

$$E_s \triangleq \bigcap_{j=1}^{\lceil \kappa n \rceil} (\mathcal{D}_{S_j})^{\varepsilon_j(s)}, s \in S, \tag{3.12}$$

are disjoint, because for $s \neq s'$ there is an S_j with $s \in S_j$ and $s' \notin S_j$ and so $\varepsilon_j(s) \neq \varepsilon_j(s')$.

Finally, we have by the properties of the original code

$$\sum_{x^n} W(E_s|x^n) Q_s(x^n) \geq 1 - \lceil n\kappa \rceil \lambda(n), s \in S. \tag{3.13}$$

The choice $\lambda(n) = \frac{1}{n^2}$ is good enough. Every S is λ'–decodable.

Remark 4: Notice that the same argument applies also in the cases of deterministic and randomized feedback strategies.

It also yields for a DMC a relation weaker than Proposition 1, namely

$$\mathcal{K} \supset \big\{ (R, \kappa) : R + 2\kappa \leq C_{er} \big\}.$$

Indeed, for an erasure code $\big\{ (u_i, \mathcal{D}_i) : 1 \leq i \leq M \big\}$ with erasure probability ε we have

$$W(\mathcal{D}_j|u_i) = 0 \;\text{ for }\; i \neq j$$

$$W(\mathcal{D}_i|u_i) \geq 1 - \varepsilon; i = 1, \ldots, M.$$

In the previous argument we can replace $\{0,1\}^n$ by $\mathcal{U} = \{u_1, \dots, u_M\}$. Subcodes of cardinalities $2^{\rho n}$ and intersecting in at most $2^{-\kappa n} 2^{\rho n}$ words give rise to identification codes (by averaging) of error probability of second kind $\lambda_2 \leq 2^{-\kappa n} 2^{\rho n}$.

The erasure probability is only relevant for the error probability of first kind.

From here on we apply Gilbert's bound with 2^n replaced by $2^{\rho n}$; $\rho \geq \kappa$, $\rho \leq C_{er}$.

Remark 5: K–identifiable, λ'–decodable codes are automatically identification codes with error probabilities smaller than $\lambda + \lambda'$. Therefore $R < C_{sh}$, and by Shannon's Coding Theorem also $\kappa \leq C_{sh}$.

Remark 6: There is another instructive relation. Let us view $\binom{\mathcal{N}}{K}$ as set $\mathcal{M}$ of objects, one of which, say S, is given to the sender for encoding. The receiver want's to know whether it equals S' (any element of $\mathcal{M}$) or not. This is a standard identification problem with $|\mathcal{M}| = \binom{N}{K}$.

Since $\frac{1}{n} \log\log|\mathcal{M}|$ cannot exceed C_{sh}, we see that for $K = 2^{\kappa n}$ and $N = 2^{2^{Rn}}$ $\binom{N}{K} \sim 2^{2^{(\kappa+R)n}} \lesssim 2^{2^{C_{sh}\cdot n}}$, or $\kappa + R \leq C_{sh}$. Thus κ cannot exceed C_{sh}. Actually, this is true even if N grows exponentially only, say like $K = 2^{\varepsilon n}$, because then

$$\binom{N}{K} \sim 2^{(\kappa+\varepsilon)n \cdot 2^{\kappa n}} \lesssim 2^{2^{C_{sh}\cdot n}} \quad \text{requires} \quad \kappa \leq C_{sh}.$$

3.3 An outer bound on the capacity region

The simple idea here is to work with a "net" $\mathcal{S} \subset \binom{\mathcal{N}}{K}$ "almost" of cardinality N^K.

View a set S as 0–1–sequence of length $N = 2^{2^{Rn}}$ with exactly $K = 2^{\kappa n}$ 1's. By Gilbert's bound we can find $\mathcal{S} = \{S_1, S_2, \dots, S_{\tilde{N}}\}$ with the properties

$$|S_i \triangle S_j| \geq (1-\alpha)2K, 0 < \alpha < 1,$$

$$\tilde{N} \geq \binom{N}{K} \left[2^K (N-K)^{(1-\alpha)K}\right]^{-1}.$$

Therefore

$$\tilde{N} \gtrsim N^{\alpha K} = 2^{2^{Rn}\cdot \alpha 2^{\kappa n}} = 2^{\alpha 2^{(R+\kappa)n}}$$

and

$$\frac{1}{n}\log\log\tilde{N} \geq R + \kappa - \frac{1}{n}|\log\alpha|.$$

We summarize this.

Lemma 1. *For every $\alpha \in (0,1)$ there is a family $\mathcal{S} = \{S_1, \dots, S_{\tilde{N}}\} \subset \binom{\mathcal{N}}{K}$ with*

(i) $|S_i \triangle S_j| \geq (1-\alpha)2K$,

(ii) $R + \kappa - \frac{1}{n}|\log\alpha| \leq \frac{1}{n}\log\log|\mathcal{S}| \leq \frac{1}{n}\log\log\binom{\mathcal{N}}{K} \leq R + \kappa.$

We can therefore upperbound $\binom{\mathcal{N}}{K}$ by upperbounding $|\mathcal{S}|$. For this we use the Equivalence Theorem and working with $\mathcal{S}$ of Lemma 1 we can relate the modified K–identification problem with a standard identification problem. For $S \in \mathcal{S}$ define $Q_S \in \mathcal{P}(\mathcal{X}^n)$ by

$$Q_S(x^n) = \frac{1}{K} \sum_{i \in S} Q(x^n|i), x^n \in \mathcal{X}^n, \tag{3.14}$$

if $Q(\cdot|i)$ is the randomized encoding for i. Now by Lemma 1 (i) and the code definition (2.1) we have for $S, S' \in \mathcal{S}, S \neq S'$,

$$\sum_{x^n} Q_S(x^n) W(\mathcal{D}_S|x^n) \geq 1 - \lambda$$

and

$$\sum_{x^n} Q_S(x^n) W(\mathcal{D}_{S'}|x^n) \leq (1-\alpha)\lambda + \alpha.$$

This is an $(n, |\mathcal{S}|, \lambda')$ identification code with

$$\lambda' = (1-\alpha)\lambda + \alpha \geq \lambda.$$

By the soft converse of [16] (or the weak converse in Section 12 or the strong converse of [28], [29]) and Lemma 1 (ii) we get the desired bound for $\mathcal{K}$. The same proof works for the K–separating codes of Section 6, if we define $\mathcal{D}_E = \bigcup_{i \in E} \mathcal{D}_{E,i}$.

So for this capacity region $\mathcal{K}^{++}$ we have the same bound.

Proposition 3.

(i) $\mathcal{K} \subset \big\{(R, \kappa) : R + \kappa \leq C_{sh}\big\}$

(ii) $\mathcal{K}^{++} \subset \big\{(R, \kappa) : R + \kappa \leq C_{sh}\big\}$.

Remark 7: There is a very simple proof for the noiseless BSC. Since the decoding sets $\mathcal{D}_S$ are distinct, it follows that

$$\left|\binom{\mathcal{N}}{K}\right| \leq 2^{2^n} \text{ and thus } \frac{1}{n}\log\log N^K = \frac{1}{n}\log\log N + \frac{1}{n}\log K = R + \kappa \leq 1.$$

Remark 8: The two Propositions 1, 2 imply for $\kappa = 0$ the standard Identification Capacity Theorem.

Remark 9: Using also the Equivalence Theorem we see that for $R = 0$ we get the converse to Shannon's Coding Theorem and only the achievable rate $\frac{1}{2}C_{sh}$!

3.4 On K–identification in case of noiseless feedback

As in [17] we assume the presence of a letter by letter noiseless feedback link. Again deterministic encoding functions for i are denoted by f_i^n and randomized encoding functions for i are denoted by F_i^n. The corresponding regions of achievable rate pairs are denoted by $\mathcal{K}_f$ and $\mathcal{K}_F$. Analogously, if all $S \in \binom{\mathcal{N}}{K}$ are λ–decodable we denote the regions by $\mathcal{K}_f^*$ and $\mathcal{K}_F^*$. We formulate now results, which are analog to those under 3.2 and 3.3. The first facts were noted already under Remark 4.

Equivalence Theorem 2. *For every DMC*

(i) $\mathcal{K}_{f\ pol} \subset \mathcal{K}_f^* \subset \mathcal{K}_f$

(ii) $\mathcal{K}_{F\ pol} \subset \mathcal{K}_F^* \subset \mathcal{K}_F$.

Proposition 4. For every DMC W

$$\mathcal{K}_F \subset \big\{(R,\kappa) : R+\kappa \leq \max_{P\in\mathcal{P}(\mathcal{X})} H(Q)\big\},$$

where $Q = P \cdot W$.

We use an entropy property for all discrete distributions.

Lemma 2 ([27]) *For* $P = (P_1, P_2, \dots) \in \mathcal{P}(\mathbb{N})$ *define*

$$\varepsilon(d,P) = \max\left\{\sum_{j\in J} P_j : J \subset \mathbb{N}, |J| = 2^{\lceil H(P)d\rceil+1}\right\},$$

and set

$$\varepsilon(d) = \min_{P\in\mathcal{P}(\mathbb{N})} \varepsilon(d,P).$$

Then

$$\varepsilon(d) = 1 - \frac{1}{d} \ \ \textit{for all} \ \ d \geq 1.$$

Proof of Proposition 4: In any (n,N,K,λ)–code with feedback $\big\{(F_i,\mathcal{D}_S) : 1 \leq i \leq N; S \in \binom{\mathcal{N}}{K}\big\}$ let Y_i^n be the output process generated by F_i via the channel. Furthermore define the process Y_S^n by the distribution

$$\text{Prob}(Y_S^n = y^n) = \frac{1}{K}\sum_{i\in S}\text{Prob}(Y_i^n = y^n).$$

By assumption

$$\text{Prob}[Y_i^n \in \mathcal{D}_S] \geq 1-\lambda, \quad \text{if} \ \ i \in S, \tag{3.15}$$

$$\text{Prob}[Y_i^n \in \mathcal{D}_{S'}] \leq \lambda, \quad \text{if} \ \ i \notin S'. \tag{3.16}$$

By Lemma 2 there are sets $\mathcal{E}_S \subset \mathcal{Y}^n$ $\left(S \in \binom{\mathcal{N}}{K}\right)$ with

$$\text{Prob}[Y_S^n \in \mathcal{E}_S] \geq 1 - \frac{1}{d}, \tag{3.17}$$

$$|\mathcal{E}_S| \leq 2^{\lceil d\ H(Y_S^n) \rceil + 1}. \tag{3.18}$$

We show later that the net $\mathcal{S} \subset \binom{\mathcal{N}}{K}$ with the properties (i), (ii) in Lemma 1 satisfies

$$\mathcal{D}_S \cap \mathcal{E}_S \neq \mathcal{D}_{S'} \cap \mathcal{E}_{S'} \ \text{ for } \ S, S' \in \mathcal{S}; S \neq S', \tag{3.19}$$

provided that λ is sufficiently small.

We know from [17], that

$$H(Y_S^n) \leq n \max_{P: Q = P \cdot W} H(Q) = \overline{H} \ \text{ (say)}. \tag{3.20}$$

Therefore by (3.18) and (3.19)

$$|\mathcal{S}| \leq 2^{2^{dn\overline{H}}} \tag{3.21}$$

and since d can be made arbitrarily close to 1 we conclude that

$$N^K \lesssim 2^{2^{g(\lambda) n \overline{H}}} \tag{3.22}$$

with $\lim\limits_{\lambda \to 0} g(\lambda) = 1$ (weak converse).

Therefore

$$\frac{1}{n} \log\log N^K = \frac{1}{n}(\log K + \log\log N) = \kappa + R \leq \overline{H}\ g(\lambda).$$

It remains to be seen that (3.19) holds.

Suppose that for $S, S' \in S$ $\mathcal{E}_S \cap \mathcal{D}_S = \mathcal{E}_{S'} \cap \mathcal{D}_{S'}$. Then by (3.15) and (3.17)

$$\text{Prob}(Y_{S'}^n \in \mathcal{E}_{S'} \cap \mathcal{D}_{S'}) = \text{Prob}(Y_{S'}^n \in \mathcal{E}_S \cap \mathcal{D}_S) \geq 1 - \frac{1}{d} - \lambda. \tag{3.23}$$

On the other hand, by (3.16)

$\text{Prob}(Y_i^n \in \mathcal{E}_S \cap \mathcal{D}_S) \leq \text{Prob}(Y_i^n \in \mathcal{D}_S) \leq \lambda$ for $i \in S' \setminus S$ and by definition of $\mathcal{S}$ $|S' \setminus S| \geq (1-\alpha)K$.

Therefore $\text{Prob}(Y_{S'}^n \in \mathcal{E}_S \cap \mathcal{D}_S) = \frac{1}{K} \sum\limits_{i \in S'} \text{Prob}(Y_i^n \in \mathcal{E}_S \cap \mathcal{D}_S) \leq (1-\alpha)\lambda + \alpha$.

This contradicts (3.23), if

$$(1-\alpha)\lambda + \alpha < 1 - \frac{1}{d} - \lambda. \tag{3.24}$$

This is equivalent with $\alpha(1-\lambda) < 1 - \frac{1}{d} - 2\lambda$.

So in order to show that for any $\varepsilon > 0$ $\kappa + R \leq \overline{H} + \varepsilon$, choose first d so that $d > 1$ and $d\overline{H} \leq \overline{H} + \varepsilon$, then choose λ smaller than $\frac{1}{3}\left(1 - \frac{1}{d}\right)$, and finally choose α smaller than $\frac{1}{3}\left(1 - \frac{1}{d}\right)$.

Remark 10: Notice that we have used that $\frac{1}{K} \sum\limits_{i \in S} f_i$ defines a *randomized* feedback strategy F_S. So this approach does not work for the case of deterministic feedback strategies!

Remark 11: We have upperbounded $\binom{N}{K}$ via upperbounding $|\mathcal{S}|$, for which we used our old idea of "distinct carriers". Instead we could also follow the approach under 3.2, in which we relate the modified K–identification problem with a standard identification problem. In case of feedback we get the upper bound for randomized strategies by the strong converse of [17].

Remark 12: For small K, say for constant K while n grows, K–identification reduces of course to K identifications and thus to identification.

Remark 13: K–identification applies whenever persons want to know whether a winner is among their favourite teams or lottery numbers or friends.

K–identification means that any person E is interested in the question whether the edge E in the hypergraph $\left(\mathcal{N}, \binom{\mathcal{N}}{K}\right)$ occured. Naturally, we can replace $\binom{\mathcal{N}}{K}$ by any set $\mathcal{E}$ of edges, if this describes the interests.

In order to motivate this model $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ let us suppose that $\mathcal{V}$ is the set of roads in a region and $\mathcal{E}$ is the set of drivers. Driver E is primarily interested in the roads of his tour. In case there has been an accident on one road $v \in \mathcal{V}$ and this road is blocked, then all E's want to know whether $v \in E$ or not (and in the affirmative case secondarily also which road it is).

There are more efficent ways of transfering the information of interest than to broadcast the complete information, which specifies the road with the accident.

The converses in case of feedback show that

$$|\mathcal{E}| < 2^{2^{\overline{H}n}}. \tag{3.25}$$

Now, if we choose $\mathcal{E} = 2^{\mathcal{N}}$, the power set, $R_1 = \text{rate}\ (\mathcal{N}) \leq \overline{H}$.

By Section 4 decoding all subsets, gives optimal rate C_{sh}. So the bound in (3.25) is not achievable.

Problem 1: Does the Equivalence Theorem hold for general hypergraphs?

3.5 A combinatorial consequence

It is remarkable that a result for K–identification (Proposition 1) has an important consequence for r–cover–free families in relation to packings. We use a result of Kuzjurin [54].

A family $\mathcal{A}$ of k–subsets of $[m]$ is called (m, k, ℓ)*–packing* iff each ℓ–subset of $[m]$ is contained in at most one member $A \in \mathcal{A}$. Therefore two members of $\mathcal{A}$ intersect in at most $\ell - 1$

elements. (In other words $\mathcal{A}$ can be viewed as a code with constant weight k, word length m and distance $d_H = 2(k-\ell)+2$.)

The *density* $d(\mathcal{A})$ of a packing $\mathcal{A}$ is the average number of k–subsets of $\mathcal{A}$ containing an ℓ–subset, that is, $d(\mathcal{A}) = \frac{|\mathcal{A}|\binom{k}{\ell}}{\binom{m}{\ell}}$. Let $k = k(m)$ and let $\ell = \ell(m) \geq 2$.

A sequence of packings $(\mathcal{A}_m)_{m \geq k}$ is called *asymptotically good* if

$$\lim_{m \to \infty} d(\mathcal{A}_m) = 1.$$

Roughly speaking the result of [54] says that $k = \sqrt{m}$ is the threshold function for the existence of asymptotically good packings. Here is the precise result.

Theorem K. *Let α be the minimum constant such that for every $\varepsilon > 0$ and sufficiently large n every interval $[n, n+n^{\alpha+\varepsilon}]$ contains a prime number. It is known that $\alpha \leq \frac{23}{43}$. The following bounds hold:*

(i) *Let $c < 1$ and $k(m) < c\sqrt{m}$, where $\lim\limits_{n \to \infty} k(m) = \infty$. Further, let for some $\varepsilon > o$ $\ell(m) = o\big(\sqrt{k(m)}\big)$ and $\ell(m) = o\left(\binom{m}{k(m)}^{1-\alpha-\varepsilon}\right)$.*

Then asymptotically good (m,k,ℓ)–packings exist.

(ii) *Let $c > 1$, $k(m) > c\sqrt{m}$ and let $\ell(m) = o\big(k(m)\big)$. Then nontrivial asymptotically good (m,k,ℓ)–packings do not exist.*

Corollary 1.

(i) *Let $m(n) = e^{\mu n}$, $k(n) = e^{\gamma n}$, and $\ell(n) = e^{\beta n}$. For $\frac{\mu}{2} > \gamma$, $\gamma/2 > \beta$ and $(\mu - \gamma)\frac{20}{43} > \beta$ we have asymptotically good (m,k,ℓ)–packings.*

(ii) *Let $m(n) = e^{\mu n}$, $k(m) = e^{\left(\frac{\mu}{2}+\varepsilon\right)n}$, and let $\ell(m) = e^{\beta n}$ with $\beta < \frac{\mu}{2} + \varepsilon$, then asymptotically good (m,k,ℓ)–packings do not exist.*

We derive from the assumptions on μ, γ, β

$$\mu > 2\gamma, \gamma > 2\beta, \mu > \gamma + \frac{43}{20}\beta. \tag{3.26}$$

We apply this and (ii) to the set of codewords $\mathcal{U} \subset \mathcal{X}^n$ of a channel code with error probability λ, $|\mathcal{U}| \sim e^{In} = m$, and $\frac{1}{n}\log K(n) = \kappa$. Then $I = \mu$, $\kappa = \gamma - \beta$ and we get for the maximal packing cardinality

$$N^*(n, I, \kappa) \lesssim \frac{\binom{e^{In}}{e^{\beta n}}}{\binom{e^{\gamma n}}{e^{\beta n}}} = \frac{\binom{e^{In}}{e^{\beta n}}}{\binom{e^{(\beta+\kappa)n}}{e^{\beta n}}}, \tag{3.27}$$

$$\frac{1}{n}\log\log N^* \lesssim \beta, \tag{3.28}$$

and for $\gamma \sim \frac{I}{2}$ the lower bound $\beta = \gamma - \kappa \sim \frac{I}{2} - \kappa$. Moreover, $\beta_{\max} \leq \min\left(\frac{I}{4}, \frac{20I}{86}\right) = \frac{10}{43}I$, $\kappa_{\min} = \frac{I}{2} - \beta_{\max} = \frac{23}{86}I$, and $R = \frac{10}{43}I$.

However, our bound $R = I - 2\kappa = \frac{20}{43}I$ in Proposition 1 is much better!

It can be seen from its derivation in 3.1 that this bound can be interpreted as a lower bound on the size $N(n, I, \kappa)$ of optimal r–cover–free families, where r has rate κ. It is known and readily verified that always

$$N(n, I, \kappa) \geq N^*(n, I, \kappa).$$

We know now that the quantities can be very different!

4 Models with capacity equal to the ordinary capacity

Some of the cases considered here were first treated by Já Já [13] for non–randomized encoding on the BSC. If randomisation is permitted, the analysis is somewhat more complicated.

4.1 The ordering problem

Suppose that one of the events $\{1, 2, \dots, N\}$ occured and is known to the sender. By proper coding he shall enable the receiver to answer the question "Is the true number less than j?" Here j is any element of $\{1, \dots, N\}$. We can also use the ordering function

$$f_0(i,j) = \begin{cases} 1 & \text{for } \ i \leq j \\ 0 & \text{otherwise.} \end{cases}$$

A (randomized) ordering code $(n, N, \lambda_1, \lambda_2)$ is a family

$$\big\{(Q(\cdot|i), \mathcal{D}_i) : i = 1, 2, \dots, N\big\}$$

of pairs with

$$Q(\cdot|i) \in \mathcal{P}(\mathcal{X}^n), \mathcal{D}_i \subset \mathcal{Y}^n \ \text{ for } \ i = 1, 2, \dots, N \tag{4.1}$$

and with errors of the first (resp. second) kind satisfying for *every* j

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_j|x^n) \geq 1 - \lambda_1 \ \text{ for } \ i = 1, \dots, j \tag{4.2}$$

and

$$\sum_{x^n \in \mathcal{X}^n} Q(x^n|i) W^n(\mathcal{D}_j|x^n) \leq \lambda_2 \ \text{ for } \ i > j. \tag{4.3}$$

Of course, we can define this way deterministic ordering codes by letting $Q(\cdot|i)$ denote point masses on points $u_i \in \mathcal{X}^n$.

Theorem 3. *Even for randomized encoding the Q_{pol}–capacity does not exceed the transmission capacity. The same holds in case of noiseless feedback.*

Proof: Suppose first that $N \leq \big(2|\mathcal{X}|\big)^n$ and that $\lambda_1, \lambda_2 \leq \frac{1}{n^2}$.

The $\mathcal{O}$–code gives rise to a transmission code as follows:

Choose first $j_1 = \lceil \frac{M}{2} \rceil$. In case of a "yes" iterate the search for the "true message" in $\{1, \dots, \lfloor \frac{M}{2} \rfloor\}$ and otherwise in $\{\lceil \frac{M}{2} \rceil, \dots, M\}$ by choosing next j_2 in the middle of these sets, resp. After $\log N$ iterations we are done. The total error probability is bounded by

$$\frac{1}{n^2} \log N \leq \frac{2|\mathcal{X}|}{n}.$$

Next, if $N > \big(2|\mathcal{X}|\big)^n$, choose *any* subset of $\{1, 2, \dots, N\}$ of a cardinality $\exp\big\{(C+\delta)n\big\}$ for some $\delta > 0$.

Apply to the subcode corresponding to this set the previous argument. This leads to a transmission code of a rate exceeding capacity and this contradiction proves that actually $N > \big(2|\mathcal{X}|\big)^n$ does not occur.

Finally, the same argument applies to the case of feedback.

Remark 14: We have shown that, generally speaking, whenever $\log N$ bits specify an event with the code concept used, its rate does not exceed C. Thus we have also the next result.

4.2 All binary questions

By proper coding the sender shall enable the receiver to answer all the questions "Is the true number in A?" Here A is any subset of $\{1, \dots, N\}$.

Theorem 4. *Even for randomized encoding the BQ–capacity does not exceed the transmission capacity. The same holds in case of noiseless feedback.*

4.3 Identification of a component

The number of components is linear in the blocklength. For exponentially small error probability words can therefore be reproduced with small error probability. (For small, but constant error probabilities, rate–distortion theory is to be used).

Theorem 5. *Even for randomized encoding the component identification capacity does not exceed the transmission capacity. The same holds in case of feedback.*

Part II: Models with prior knowledge of the receiver

The a priori structure is a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$. The encoder knows the vertex $v \in \mathcal{V}$ and the decoder D_E $(E \in \mathcal{E})$ knows beforehand whether the message to be transmitted is in E or not. In case it is, he wants to know which element of E it is.

We consider first abstract hypergraphs.

5 Zero–error decodable hypergraphs

If the decoder wants to identify $v \in E$, then any two vertices $x, y \in E$ must be separable.

Definition: The separability graph $\mathcal{G}(\mathcal{H}) = (\mathcal{V}, \mathcal{E}^*)$ is defined by

$$\{x, y\} \in \mathcal{E}^* \Leftrightarrow \exists\, F \in \mathcal{E} : \{x, y\} \subset F. \tag{5.1}$$

Let $\Psi(\mathcal{G})$ be the chromatic number of $\mathcal{G}$, then $\mathcal{H}$ is 0–error decodable iff $\Psi(\mathcal{G}) \le 2^{C_0 n}$, where C_0 is the zero–error capacity of the channel. Now $\mathcal{H}$ is λ–identifiable iff $\Psi(\mathcal{G}) \lesssim 2^{2^{Cn}}$.

Remark 15: Also if 2–separable within edges by the results of [16], [28] the answer is the same.

6 K–separating codes

Instead of zero–error decodability for hypergraphs one can consider λ–decodability, that is, an error probability not exceeding λ is permitted.

We call $\big\{(P_i, \mathcal{D}_{E,i}) : E \in \mathcal{E}, i \in E\big\}$ an $[n, N, \lambda]$–code for $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and W, if $P_i \in \mathcal{P}(\mathcal{X}^n)$ for $i \in \mathcal{V} = \{1, 2, \dots, N\}$, $\mathcal{D}_{E,i} \subset \mathcal{Y}^n$,

$$\mathcal{D}_{E,i} \cap \mathcal{D}_{E,i'} = \varnothing \quad \text{for} \quad i, i' \in E, i \neq i' \tag{6.1}$$

and for all $E \in \mathcal{E}$

$$\sum_{x^n} P_i(x^n) W(\mathcal{D}_{E,i} | x^n) \ge 1 - \lambda \quad \text{for} \quad i \in E. \tag{6.2}$$

The issue is to minimize n for given $\mathcal{H}$ (and thus N) and λ for the channel W.

For abstract hypergraphs $\mathcal{H}$ not very much can be said. The subject becomes interesting under reasonable assumptions on $\mathcal{H}$.

Example 4: $\mathcal{E} = \{\mathcal{V}\}$ describes Shannon's theory of transmission.

Example 5: $\mathcal{E} = \binom{\mathcal{V}}{K}$, the family of all K–element subsets of $\mathcal{V}$, defines the complete K–uniform hypergraph. The codes defined above are denoted here by $[n, N, K, \lambda]$ and called *K–separating codes.*

Clearly, their capacity region $\mathcal{K}^{++}$ contains $\mathcal{K}^*$ and by the Equivalence Theorem 1 also $\mathcal{K}_{\text{pol}}$.

Corollary 2.

$$\mathcal{K}^{++} \supset \mathcal{K}^* \supset \mathcal{K}_{\text{pol}}.$$

Problem 2: Determine $\mathcal{K}^{++}$!
However good is the bound in (ii) of Proposition 3 in Section 3?

6.1 Second order 2–separation capacity without and with feedback

Let us start with the first meaningfull case $K = 2$.

For $E = \{i, j\}$ we can write

$$\mathcal{D}_{E,i} = \mathcal{D}_{ij} \text{ and } \mathcal{D}_{E,j} = \mathcal{D}_{ji}.$$

We also say that any two messages are λ–decodable.

Notice that an (n, N, λ)–ID–code $\big\{(P_i, \mathcal{D}_i) : 1 \le i \le N\big\}$ satisfies $\sum\limits_{x^n} P_i(x^n) W(\mathcal{D}_i|x^n) \ge 1 - \lambda$ and $\sum\limits_{x^n} P_j(x^n) W(\mathcal{D}_i|x^n) \le \lambda (i \ne j)$. Therefore setting $\mathcal{D}_{ij} = \mathcal{D}_i \smallsetminus \mathcal{D}_j$ and $\mathcal{D}_{ji} = \mathcal{D}_j \smallsetminus \mathcal{D}_i$ we see that i and j are 2λ–separable. It immediately follows that the second order capacity for $K = 2$, say C_2, is not smaller than the ID–capacity C_{sh}. Whereas in ID-codes the decoding sets carry one index, 2–separating codes carry two indices. The decoding sets for two messages are adapted for these two and no other message. Therefore *2–separation* is a weaker notion than identification (except, perhaps, for a small shift in error probability caused by the disjointness of the two decoding sets).

Theorem 6.

(i) *The 2–separation capacity of second order C_2 equals the second order identification capacity C_{sh}.*

(ii) *The corresponding capacities for channel (deterministic and randomized) feedback strategies are also equal.*

Proof: The issues are the converses.

(i) Here we can be brief, because inspection of the strong converse proof for identification of Han/Verdu [28] shows that it is actually designed for 2–separation. The key fact, called resolvability in [29], is this:

For $P \in \mathcal{P}(\mathcal{X}^n)$ with $Q = PW^n$ and $\varepsilon > 0$ there is a $P^* \in \mathcal{P}(\mathcal{X}^n)$, which is an equidistribution over at most $\sim \exp\big\{I(P|W^n)\big\}$, not necessary distinct, members of $\mathcal{X}^n$ and such that for $Q^* = P^* W^n$

$$\|Q - Q^*\| \le \varepsilon \text{ for } n \ge n(\varepsilon). \tag{6.3}$$

(Here $\| \ \|$ denotes total–variation).

In this way to every encoding distribution $P_i (1 \leq i \leq N)$ we can find a distribution P_i^* such that the corresponding output distribution is close to that of P_i. By the code properties the Q_i's and also the Q_i^*'s are distinct. Therefore the P_i^*'s must be distinct and there number in second order rate does not exceed $\frac{1}{n} \max\limits_{P \in \mathcal{P}(\mathcal{X}^n)} I(P|W^n) = C_{sh}$.

(ii) Let us consider the deterministic case. For the randomized case we just have to replace $\underline{H}$ by $\overline{H}$.

For encoding function f_i let $\mathcal{E}_i$ be the carrier in the sence of Lemma 2 for $Q_i = W^n(\cdot|f_i)$. Then $Q_i(\mathcal{E}_i) \geq 1 - \frac{1}{d}$, $|\mathcal{E}_i| \leq 2^{d \underline{H} n}$. Omit from $\mathcal{E}_i$ the elements with smallest probability until we get a set $\mathcal{E}_i^* \subset \mathcal{E}_i$ with $Q_i(\mathcal{E}_i^*) \geq 1 - \frac{1}{d}$ and which is minimal with this property.

Set $T = \max\limits_i |\mathcal{E}_i^*|$. The number of different such sets is

$$\left| \binom{|\mathcal{Y}|^n}{T} \right| \leq 2^{(n \log |\mathcal{Y}|) 2^{d \underline{H} n}}. \tag{6.4}$$

This is the desired upper bound. However, not all $\mathcal{E}_i^*$'s are necessarily different. Therefore, we have to upperbound the multiplicity with which a set, say $\mathcal{F}$, occurs among the $\mathcal{E}_i^*$'s. W.l.o.g. we label them $\mathcal{E}_1^*, \dots, \mathcal{E}_M^*$. By our definitions

$$1 - \frac{1}{d} + \frac{1}{T} \geq Q_i(\mathcal{E}_i^*) \geq 1 - \frac{1}{d}. \tag{6.5}$$

For $i, j \in \{1, \dots, M\}$ we have for λ small

$$Q_i(\mathcal{F} \cap \mathcal{D}_{ij}) \geq 1 - \lambda - \frac{1}{d} > \lambda, Q_j(\mathcal{F} \cap \mathcal{D}_{ji}) \geq 1 - \lambda - \frac{1}{d} > \lambda,$$

$$\text{and} \quad Q_i(\mathcal{D}_{ji}), Q_j(\mathcal{D}_{ij}) \leq \lambda.$$

If we now set $\mathcal{D}'_{\ell k} = \mathcal{F} \cap \mathcal{D}_{\ell k}$ and renormalize the measure Q_i on $\mathcal{F}$ from total measure $\sim 1 - \frac{1}{d}$ (see 6.5) to 1, then we have a 2–separating code of size M with output space $\mathcal{F}$.

To this situation we apply the idea of resolvability in the following setting: We want to know how many distributions can be 2–separated on a finite set $\mathcal{T}$ with T elements which we can view as subset of $\{0,1\}^n$, $T \leq 2^m$. This is covered by Han/Verdu's result, when W is the noiseless BSC. We get the bound $M \leq 2^{2^m}$ or

$$M \leq 2^{2^{d \underline{H} n}}. \tag{6.6}$$

Together with (6.4) we get

$$N \leq 2^{(n \log |\mathcal{Y}|) 2^{d \underline{H} n}} \cdot 2^{2^{d \underline{H} n}} \leq 2^{(1 + n \log |\mathcal{Y}|) 2^{d \underline{H} n}},$$

and thus the weak converse by choosing d close to 1, λ then small enough and $n \geq n(d, \lambda)$.

6.2 Strong converses by the method of [17] for 2–separation in case of feedback

We begin with Theorem 6, (ii) in 6.1. By Lemma 2 of [17] for any $\varepsilon \in (0,1)$ we can find sets $\mathcal{E}_i^*(i = 1, \dots, N)$ of minimal size with

$$1 \geq W(\mathcal{E}_i^*|f_i) \geq 1 - \varepsilon, \tag{6.7}$$

$$|\mathcal{E}_i^*| \leq 2^{\left(\underline{H} + \frac{c(\varepsilon)}{\sqrt{n}}\right)n}. \tag{6.8}$$

How many can be equal to $\mathcal{F}$, say?

Now just repeat the previous proof in 6.1. Now (the sharper) (6.7) takes the role of (6.5). Instead of (6.6) we get now the stronger

$$M \leq 2^{2^{\left(\underline{H} + \frac{c(\varepsilon)}{\sqrt{n}}\right)n}} \tag{6.9}$$

and finally

$$N \leq 2^{(n \log |\mathcal{Y}|)2^{\left(\underline{H} + \frac{c(\varepsilon)}{\sqrt{n}}\right)n}} \cdot 2^{2^{\left(\underline{H} + \frac{c(\varepsilon)}{\sqrt{n}}\right)n}}$$

and thus

$$\frac{1}{n} \log\log N \leq \underline{H} + \frac{c(\varepsilon)}{\sqrt{n}} \quad \text{(Strong Converse)}. \tag{6.10}$$

Replacing f_i by F_i and $\underline{H}$ by $\overline{H}$ the same proof applies otherwise literally and gives a strong converse for randomized encoding.

Remark 16: The results obviously generalize to any constant K.

Problem 3: Are the optimal rates for 2–separable codes and ID–codes equal if they satisfy $\lambda_2 \leq e^{-\eta_2 n}$?

7 Analysis of a model with specific constraints: 2–separation and Renyi's entropy H_2

Let us assume that a set of persons $\mathcal{N} = \{1, 2, \dots, N\}$ are at a party. The persons move randomly between α rooms and the set of persons in room i at some time is A_i of cardinality

$$|A_i| = P_i N; \; i = 1, \dots, \alpha. \tag{7.1}$$

We say that the partition $\Pi = (A_1, \dots, A_\alpha)$ is of type $P = (P_1, P_2, \dots, P_\alpha) \in \mathcal{P}(\mathcal{N})$.

Let now $\Pi_1, \Pi_2, \dots, \Pi_m$ be a sequence of independent random partitions taking as values a partition of type $\mathcal{P}$ with equal probabilities. Equivalently we can say that a person $z \in \mathcal{N}$

belongs to the randomly choosen A_i with probability P_i independently of what happens to the other persons. (At discrete time points $1, 2, \dots$ the partition of the persons in several rooms is reported.)

Imagine now that somebody, the interrogator, has difficulties to distinguish any two persons in his interest at the party, but is reported the sequence of partitions described. So he knows at every time instance the set of persons in all rooms, but he cannot identify the persons in a set.

Let now $\lambda_{N,m}$ denote the probability that m such partitions separate any two persons in $\mathcal{N}$. Renyi [34] has shown that $m_2(N,\varepsilon)$, the smallest m with $\lambda_{N,m} \geq 1-\varepsilon$, satisfies

$$m_2(N,\varepsilon) \sim \frac{2\log_2 N + o(\varepsilon)}{H_2(P)}, \tag{7.2}$$

where H_2 is Renyi's entropy of order 2.

Now let us go a step further. The interrogator is at the receiver side of a noisy channel. For partition $\Pi_i = (A_{i1}, \dots, A_{i\alpha})$ let

$$F_i(z) = j, \quad \text{if} \quad z \in A_{ij}. \tag{7.3}$$

For every $z \in \mathcal{N}$ $\big(F_1(z), \dots, F_m(z)\big)$ is known to the encoder. How fast can the interrogator decide his question with high probability correctly?

Answer: Match $(F_1, \dots, F_m)$ with a 2–*separation code.*

It would be stupid to use a transmission code. There are several variations of this model.

In many situations of information transfer reduction to transmission would be of poor performance.

8 Bining via channels

In Section 5 we considered vertex colorings with different colors in each edge. They have been called strict colorings in [6] and [8]. Other colorings discussed there are

(α) colorings, where in every edge no color occurs more than ℓ times (leading to list–knowledge)

(β) colorings, where in every edge a high percentage of colors occurs only one time

(γ) colorings, which are good, in the senses of (α) and/or (β) in average under given probability distributions on vertices and/or edges.

The present investigations have born still another coloring (or bining) concept.

Indeed, let us look at K–separation. We know from Proposition 1 that we can choose N with second order rate R and K with rate κ, $R + 2\kappa \leq C_{sh}$, and achieve K–identification.

Further, by the Equivalence Theorem the hypergraph $\left(\mathcal{N}, \binom{\mathcal{N}}{K}\right)$ is in addition K–separable. *What does this mean*? Well, the "color" on vertex i is the randomized encoding P_i and within every edge $S \in \binom{\mathcal{N}}{K}$ containing i this i is decoded correctly with probability at least $1-\lambda$!

Notice that for the price of a small error probability λ now — in contrast to the situation in (β) (or also (γ)) — *every* vertex can be decoded correctly.

Furthermore, the theory in [6], [8] works, if the number of vertices, the number of edges, and the edge sizes are roughly of the same growth, namely exponential in n.

Here the edge sizes are at most exponential in n, but the number of vertices and edges can grow double exponentially in n!

9 K–identifiability, K–separability and related notions

We discuss here connections between code concepts.

To fix ideas let us first compare 1–identification (the classical identification) and 2–separation. In both cases we have a fixed *encoding structure* (set of codewords, set of probability distributions or set of randomized or non–randomized–feedback functions). In any case they specify via the channel a set of *output distributiuons*

$$\mathcal{Q} = \{Q_i : i \in \mathcal{N}\}. \tag{9.1}$$

The various code concepts associate with such a set a *decoding structure*.

In case of identification the decoding structure is

$$\mathcal{D} = \{\mathcal{D}_i : i \in \mathcal{N}\}. \tag{9.2}$$

It is of *precision* λ, if

$$Q_i(\mathcal{D}_i) \geq 1 - \lambda (i \in \mathcal{N}) \;\text{ and }\; Q_i(\mathcal{D}_j) \leq \lambda (i \neq j). \tag{9.3}$$

The precision relates to the whole encoding structure $\mathcal{Q}$, however, in a pairwise fashion (as specified in (9.3)).

The concept 2–separation allows more freedom in the decoding structure. We say $\mathcal{Q}$ is *2–separable with precision* λ, if for any $S = \{i,j\} \in \binom{\mathcal{N}}{2}$ there are two sets $\mathcal{D}_{Si}$ and $\mathcal{D}_{Sj}$ with

$$\mathcal{D}_{Si} \cap \mathcal{D}_{Sj} = \varnothing, Q_i(\mathcal{D}_{Si}), Q_j(\mathcal{D}_{Sj}) \geq 1 - \lambda. \tag{9.4}$$

These sets relate only to i and j.

Lemma 3. *1–identifiable with precision λ implies 2–separable with precision 2λ.*

Proof: Define $\mathcal{D}_{Si} = \mathcal{D}_i \smallsetminus \mathcal{D}_j$ and $\mathcal{D}_{Sj} = \mathcal{D}_j \smallsetminus \mathcal{D}_i$, then $\mathcal{Q}_\ell(\mathcal{D}_{S\ell}) \geq 1 - 2\lambda$ for $\ell = i, j$.

There is also a general connection.

Lemma 4. *K–identifiable with precision $\lambda(n)$ implies K–separable with precision $\lambda'(n) = \lceil n\kappa \rceil \lambda(n)$, where*

$$\kappa = rate\ (K) = \frac{1}{n} \log K.$$

Proof: See proof of Equivalence Theorem 1 in Section 3.

Problem 4: For $L \geq K$, how does K–identifiability relate to L–separability?

Finally we mention related concepts.

a. We say that a K–identification decoding is based on a 1–identification decoding $\{\mathcal{D}_i : i \in \mathcal{N}\}$ of precision λ, if

$$\mathcal{D}_S = \bigcup_{i \in S} \mathcal{D}_i, S \in \binom{\mathcal{N}}{K} \tag{9.5}$$

and

$$Q_i(\mathcal{D}_i) \geq 1 - \lambda \ \text{ for } \ i \in \mathcal{N}, Q_i(\mathcal{D}_S) \leq \lambda \ \text{ for } \ i \notin S. \tag{9.6}$$

For the disjoint sets

$$\mathcal{D}_{Si} = \mathcal{D}_i \smallsetminus \bigcup_{j \in S \smallsetminus \{i\}} \mathcal{D}_j \ \text{ for } \ i \in S$$

we have

$$Q_i(\mathcal{D}_{Si}) \geq 1 - 2\lambda \ \text{ for } \ i \in S, \tag{9.7}$$

a generalisation of Lemma 3.

b. As a weaker notion than K–separability we define for positive integers α, β with $\alpha + \beta = K$ that $\mathcal{Q}$ *is (α, β)–separable with precision* λ, if for every $S \in \binom{\mathcal{N}}{K}$ and every partition $\{S_0, S_1\}$ of S, where $|S_0| = \alpha$ and $|S_1| = \beta$, there are disjoint sets $\mathcal{D}_{S_0}$ and $\mathcal{D}_{S_1}$ with

$$Q_j(\mathcal{D}_{S_0}) \geq 1 - \lambda \ \text{ for } \ j \in S_0$$

and

$$Q_j(\mathcal{D}_{S_1}) \geq 1 - \lambda \ \text{ for } \ j \in S_1.$$

c. Analogously we say that $\mathcal{Q}$ is (α, β)–identifiable with precision λ, if there is a decoding structure $\left\{\mathcal{D}_{S'} : S' \in \binom{\mathcal{N}}{\alpha} \cup \binom{\mathcal{N}}{\beta}\right\}$ such that for $S = S_0 \dot\cup S_1$, $|S_0| = \alpha$, $|S_1| = \beta$

$$Q_i(\mathcal{D}_{S_\varepsilon}) \geq 1 - \lambda \ \text{ for } \ i \in S_\varepsilon$$

and

$$Q_i(\mathcal{D}_{S_\varepsilon}) \leq \lambda \ \text{ for } \ i \in S_{1-\varepsilon}$$

for $\varepsilon = 0, 1$.

K–identification concerns partitions $\{S, \mathcal{N} \smallsetminus S\}$, $S \in \binom{\mathcal{N}}{K}$. One can consider partitions π_ℓ, $\ell \in \mathcal{L}$, into more than 2 sets. Person ℓ wants to know the set in its partition, which contains the "message". There may be several channels. ("From which country is a sportsman?", "what is his age?" etc.)

This model includes compound channels, where the receiver knows the individual channel, broadcast channels (also with degraded message sets) etc.

Part III: Models with prior knowledge at the sender

10 Identification after group testing and a stronger form of the Rate–Distortion Theorem

Suppose that from the set $\mathcal{N}$ of persons any subset $\mathcal{S} \subset \mathcal{N}$ of persons may be the set of sick persons. Moreover it has been already tested that with probability q a person is sick and that the RV S has the distribution

$$\mathrm{Prob}(S = \mathcal{S}) = q^{|\mathcal{S}|}(1-q)^{N-|\mathcal{S}|}. \tag{10.1}$$

Decoder s wants to know whether he is sick. Any other information is of much less relevance to him. In terms of partitions

$$\pi_S = \big\{\{\mathcal{S} \subset \mathcal{N} : s \in \mathcal{S}\}, \{\mathcal{S} \subset \mathcal{N} : s \notin \mathcal{S}\}\big\} \tag{10.2}$$

he wants to know which member of π_S occured.

We can reformulate this problem by identifying $\mathcal{S} \subset \mathcal{N}$ with a word $x_{\mathcal{S}} = (x_1, \dots, x_N) \in \{0,1\}^N$, $x_s = 1$ iff $s \in \mathcal{S}$. Thus the distribution defined in (10.1) describes a DMS $\big(\{0,1\}^N, Q^N, X^N\big)$ with $Q^N(x^N) = \prod_{t=1}^N Q(x_t)$, where

$$Q(x_t) = \begin{cases} q & \text{for } \ x_t = 1 \\ 1-q & \text{for } \ x_t = 0, \end{cases} \tag{10.3}$$

and for $X^N = (X_1, \dots, X_N)$

$$\mathrm{Prob}(X^N = x^N) = Q^N(x^N). \tag{10.4}$$

For any encoding function $f_N : \{0,1\}^N \to \mathbb{N}$ and decoding function $g_t(1 \le t \le N) : \mathbb{N} \to \{0,1\}$ we can set

$$\hat{X}_t = g_t\big(f_N(X^N)\big) \tag{10.5}$$

and consider the error probability

$$\lambda_t = \mathbb{E}\ d(X_t, \hat{X}_t),$$

where d is the Hamming distance.

Now the Rate–distortion Theorem tells us how small a rate $R(\lambda)$ we can achieve with rate(f_N) under the constraint

$$\sum_{t=1}^{N} \mathbb{E}\ d(X_t, \hat{X}_t) \leq \lambda\ N. \tag{10.6}$$

However, we are interested in the stronger condition

$$\mathbb{E}\ d(X_t, \hat{X}_t) \leq \lambda \ \text{ for } \ 1 \leq t \leq N \tag{10.7}$$

and the corresponding minimal rate $R^*(\lambda)$. We know that

$$\lim_{\lambda \to 0} R(\lambda) = h(q)$$

and therefore as $\lambda \to 0$ by the Source Coding Theorem also $\lim\limits_{\lambda \to 0} R^*(\lambda) = h(q)$.

What can we say about $R^*(\lambda)$ if λ is kept at a prescribed level? In the standard random coding proof one shows that for the random partition F and every t

$$\mathbb{E}\ d\big(X_t, \hat{X}_t(F, X_t)\big) \leq \lambda \tag{10.8}$$

for rate$(F) \geq R(\lambda) + \delta$ (δ arbitrary small). Therefore $\sum_{t=1}^{N} \mathbb{E}\ d\big(X_t, \hat{X}_t(F, X_t)\big) \leq \lambda N$ and there exists a partition good for the sum of distortions.

Instead of (10.8) one can actually show by the standard approach (see [25]) that even for every t and $\eta > 0$

$$\text{Prob}\big\{d\big(X_t, \hat{X}_t(F, X_t)\big) > \lambda + \eta\big\} \leq \exp\big\{-\rho(\eta)N\big\}, \rho(\eta) > 0. \tag{10.9}$$

Since $N\exp\big\{-\rho(\eta)N\big\} < 1$ there exists a partition f such that $\mathbb{E}\ d\big(X_t, \hat{X}_t(f, X_t)\big) \leq \lambda + \eta$ for all t.

Theorem 7. *The identification after group testing in a group of N persons, everyone being independently sick with probability q, can be performed at error probability λ with $R(q, \lambda)N$ bits. Here $R(q, \lambda)$ is the rate–distortion function for the Bernoulli source with generic distribution $(q, 1-q)$ evaluated at distortion level λ.*

Remark 17: Instead of two properties (sick and not sick) there can be any finite number of properties k defining k classes and every person wants to know its class. This leads to Rate–distortion theory for a DMS.

In case the encoding of S is transmitted via a *noisy* channel an argument for the separation of source and channel coding is needed. To elaborate conditions under which the "separation principle" is valid is a major subject in Information Theory.

Part IV: Identification and transmission with multi–way channels

11 Simultaneous transfer: transmission and identification

The issue of simultaneouity comes up frequently in life and in science. In information theory we encounter situations where the same code is used for several channels, where several users are served by the same channel, where one code serves several users etc.

A. Let us discuss now a specific example. Suppose that one DMC is used *simultaneously* for transmission and identification. Since both, the transmission capacity and the (second order) identification capacity, equals C_{sh}, here is the best we can do: We use an (n, M) transmission code $\big\{(u_i, \mathcal{D}_i) : 1 \leq i \leq M\big\}$ with *average* error $\overline{\lambda}$. The randomness in the messages produces via this code a common random experiment for sender and receiver. Adding a few, say, $\sqrt{n}$ letters, we can get the desired identification code $(n + \sqrt{n}, N, \lambda)$ as in [17]. Hence, (C_{sh}, C_{sh}) is achievable.

Next suppose that there is a noiseless (passive) feedback channel and we use the same code as before. This generates an input process $X^n = (X_1, \dots, X_n)$ and an output process $Y^n = (Y_1, \dots, Y_n)$, which is known also to the sender by the feedback. So we get a common random experiment of rate $\frac{1}{n}H(Y^n)$. Again by the identification trick of [17] now

$$R_{\text{transm.}} \sim \frac{1}{n} I(X^n \wedge Y^n)$$
$$R^2_{\text{ident.}} \sim \frac{1}{n} H(Y^n).$$

It is now easy to show the direct part in

Theorem 8. *$\mathcal{R} = conv\big\{\big(I(X \wedge Y), H(Y)\big) : P_X \in \mathcal{P}(\mathcal{X})\big\}$ is the set of achievable pairs of rates for the simultaneous transmission and identification over the DMC with noiseless feedback.*

Proof of converse: Let the RV U take values in the set of codewords $\mathcal{U} = \{u_1, \dots, u_M\}$ for transmission with equal probabilities. Further let $F_i(u)$ be the randomized encoding for i and $u \in \mathcal{U}$, making use of the feedback. Then for the transmission

$$\frac{1}{M} \sum_{j=1}^{M} W\big(\mathcal{D}_j^c | F_i(u_j)\big) \leq \overline{\lambda} \quad \text{for all} \quad i \tag{11.1}$$

and for identification with decoding sets

$$\frac{1}{M} \sum_{j=1}^{M} W\big(\mathcal{D}_i^* | F_i(u_j)\big) \geq 1 - \lambda \quad \text{for} \quad i = 1, \dots, N \tag{11.2}$$

and

$$\frac{1}{M}\sum_{j=1}^{M} W\big(\mathcal{D}_j^*|F_i(u_i)\big) \le \lambda \ \text{ for } \ i \ne j. \tag{11.3}$$

For every i we get input variables $X_i^n = (X_{i1}, \dots, X_{in})$ and output variables $Y_i^n = (Y_{i1}, \dots, Y_{in})$.

By Shannon's weak converse proof for the DMC with feedback

$$\log M \le \frac{I(X_i^n, Y_i^n)}{1-\overline{\lambda}} \ \text{ for all } \ i \tag{11.4}$$

and by the weak converse proof for identification on the DMC with feedback ([27])

$$\log\log N \le \max_i H(Y_i^n) \tag{11.5}$$

Therefore by some i

$$\begin{aligned}&\left(\frac{1}{n}\log M, \frac{1}{n}\log\log N\right) \le \left(\frac{1}{n}I(X_{i_0}^n, Y_{i_0}^n), \frac{1}{n}H(Y_{i_0}^n)\right)\\ &\le \left(\frac{1}{n}\sum_{t=1}^{n} I(X_{i_0t}, Y_{i_0t}), \frac{1}{n}\sum_{t=1}^{n} H(Y_{i_0t})\right)\\ &\le \big(I(\overline{X}, \overline{Y}), H(\overline{Y})\big)\,,\end{aligned}$$

if we use the concavity of I and of H.

We draw attention of the fact that it is a lucky coincidence that these two proofs are available and can be combined. The known strong converses for the separate problems cannot be combined!

B. More generally there is a theory of multiple purpose information transfer. Different goal seeking achivities are optimized in combinations. The familiar compound and broadcast (also with degraded message sets) channels are included.

Not just transmission and identification, but any collection of the models in Section 2 can occur in various combinations. For example consider a MAC with three senders. For a given sportsman sender 1 says from which country he comes, sender 2 informs about the age groups, and sender 3 is concerned about the fields of activities.

C. Memory decreases the identification capacity of a discrete channel with alphabetic $\mathcal{X}$ and $\mathcal{Y}$ in case of noiseless feedback.

(α) For non–random strategies this immediately follows from the inequality

$$\max_{x^n} H\big(W^n(\cdot|x^n)\big) \le \sum_{t=1}^{n} \max_{x_t} H\big(W(\cdot|x_t)\big)$$

(β) For a randomized strategy F

$$H\big(W^n(\cdot|F)\big) = H(Y_1,\dots,Y_n) = H(Y_n|Y_1,\dots,Y_{n-1}) + H(Y_1,\dots,Y_{n-1})$$

and

$$\begin{aligned}
H(Y_n|Y_1,\dots,Y_{n-1}) &= \sum_{y^{n-1}} \mathrm{Prob}(Y^{n-1} = y^{n-1}) \cdot H(Y_n|y_1,\dots,y_{n-1}) \\
&= H\left(\sum_x W_n(\cdot|x) \sum_{y^{n-1}} \mathrm{Prob}\big(F_n(y_1\dots y_{n-1}) = x\big)\right) \\
&\le \max_{P_X} H(P_X W_n).
\end{aligned}$$

12 A proof of the weak converse to the identification coding theorem for the DMC

An (n,N,λ) identification code for the DMC with transmission probability matrix W is a system of pairs $\big\{(P_i,\mathcal{D}_i) : 1 \le i \le N\big\}$ with $P_i \in \mathcal{P}(\mathcal{X}^n)$ and $\mathcal{D}_i \subset \mathcal{X}^n$ with error probability of misacceptance and also misrejection less than λ, that is,

$$\sum_{x^n} P_i(x^n) W(\mathcal{D}_i|x^n) > 1-\lambda \text{ for all } i \text{ and } \sum_{x^n} P_i(x^n) W(\mathcal{D}_j|x^n) < \lambda \text{ for } i \ne j. \tag{12.1}$$

We know that any (second order) rate $R < C_{\mathrm{Shannon}} = C$ is achievable for any $\lambda > 0$ and all large n, that is, there are (n,N,λ) codes with $R \le \frac{1}{n}\log\log N$.

It is convenient to introduce the maximal code size

$$N(n,\lambda) = \max\big\{N : (n,N,\lambda) \text{ code exists } \big\}. \tag{12.2}$$

Already in [16] it was shown that for any exponentially small sequence of error probabilities $\lambda_n = e^{-\varepsilon n}$ $(\varepsilon > 0)$

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log\log N(n,\lambda_n) \le C. \tag{12.3}$$

This converse was named soft converse in [16]. We use here the more instructive name "exponential weak converse".

The (classical) weak converse states that

$$\inf_{\lambda>0} \overline{\lim_{n\to\infty}} \frac{1}{n} \log\log N(n,\lambda) \le C. \tag{12.4}$$

As a statement between these two we introduce now a *polynomial weak converse*:

For some $\alpha > 0$

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log\log N\left(n, \frac{1}{n^\alpha}\right) \leq C. \tag{12.5}$$

Such a statement was derived for $\alpha = 1$ in [35].

Again already in [16] a version of the strong converse was conjectured:

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log\log N(n,\lambda) \leq C \quad \text{for} \quad 0 \leq \lambda < 1/2.$$

In case of feedback this was proved in [17] and the conjecture of [16] was established by Han/Verdu [28] and with a simpler proof in [29].

We present here a new approach to converses for identification. The proof is simpler than its predecessors. (Except for those in case of feedback [17], [27].)

Moreover, the approach is applicable to multi–way channels.

Furthermore, in contrast to the proofs in [28], [29] the approach works also for channels without a strong converse for transmission.

We begin our analysis with any channel $W: \mathcal{X} \to \mathcal{Y}$, that is, a time free situation and its (N,λ) codes $\big\{(P_i, \mathcal{D}_i) : 1 \leq i \leq N\big\}$ with $P_i \in \mathcal{P}(\mathcal{X}), \mathcal{D}_i \subset \mathcal{Y}$,

$$\sum_x P_i(x) W(\mathcal{D}_i|x) > 1-\lambda \quad \text{for all} \quad i \quad \text{and} \quad \sum_x P_i(x) W(\mathcal{D}_j|x) < \lambda \quad (i \neq j).$$

For any distribution $P_X \in \mathcal{P}(\mathcal{X})$ we write P_{XY} for $P_X \times W$.

For any set $G \subset \mathcal{X} \times \mathcal{Y}$ we introduce

$$\rho(G) = \min_{(x,y)\in G} \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \tag{12.6}$$

and

$$\sigma(G) = \max_{(x,y)\in G} \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}. \tag{12.7}$$

The ratio $\rho(G)\sigma(G)^{-1}$ measures how "informationally balanced" the set G is under P_{XY}. Clearly $0 \leq \rho(G)\sigma(G)^{-1} \leq 1$ and the closer to 1 the ratio is the more balanced G is.

We state now our key results.

Lemma 5. (Codes in informationally balanced sets)

For any $G \subset \mathcal{X} \times \mathcal{Y}$, $P_{XY} = P_X W$, and any $\delta' < P_{XY}(G)$ there exists a transmission code $\big\{(u_i, \mathcal{E}_i) : 1 \leq i \leq M\big\}$ with

(i) $\mathcal{E}_i \subset G_{u_i} = \big\{y : (u_i, y) \in G\big\}$

(ii) $W(\mathcal{E}_i|u_i) > \delta'$ *for* $i = 1, 2, \ldots, M$

(iii) $M > \big(P_{XY}(G) - \delta'\big)\rho(G)$

(iv) $M < \frac{\sigma(G)}{\delta'}$ *(This holds for* any *code with (i) and (ii))*

(v) $P_Y\left(\bigcup\limits_{i=1}^{M} \mathcal{E}_i\right) \geq P_{XY}(G) - \delta'$.

(vi) *For* $Q(y) \triangleq \frac{1}{M}\sum\limits_{i=1}^{M} W(y|u_i) \quad Q(y) \geq \delta'\rho(G)\sigma(G)^{-1}P_Y(y)$, *if* $y \in \mathcal{E} = \bigcup\limits_{i=1}^{M} \mathcal{E}_i$.

Proof: Let $u_1 \in \mathcal{X}$ satisfy $W(G_{u_1}|u_1) > \delta'$. Its existence follows from $P_{XY}(G) > \delta'$. Set $\mathcal{E}_1 = G_{u_1}$, then define $(u_2, \mathcal{E}_2), \dots, (u_{j-1}, \mathcal{E}_{j-1})$ and add $u_j \in \mathcal{X}$ with $\mathcal{E}_j = G_{u_j} \smallsetminus \bigcup\limits_{i=1}^{j-1} \mathcal{E}_i$ and $W(\mathcal{E}_j|u_j) > \delta'$.

The procedure terminates at M, when no pair can be added subject to the constraints (i) and (ii). Consequently for all $x \in \mathcal{X}$

$$W\left(G_x \smallsetminus \bigcup_{i=1}^{M} \mathcal{E}_i | x\right) < \delta'. \tag{12.8}$$

Since obviously for all $(x, y) \in G$

$$W(y|x) \geq \rho(G)P_Y(y) \tag{12.9}$$

and since $1 \geq W(\mathcal{E}_i|u_i)$, we have

$$P_Y(\mathcal{E}_i) \leq \rho(G)^{-1}. \tag{12.10}$$

It follows from (12.8) that

$$P_{XY}\left(G \smallsetminus \mathcal{X} \times \bigcup_{i=1}^{M} \mathcal{E}_i\right) < \delta'$$

and therefore also with (12.10)

$$P_{XY}(G) < \delta' + \sum_{i=1}^{M} P_Y(\mathcal{E}_i) \leq \delta' + M\rho(G)^{-1}.$$

This is (iii).

From the definition of σ for $(x, y) \in G$ $P_Y(y)\sigma(G) \geq W(y|x)$ and thus

$$P_Y(\mathcal{E}_i)\sigma(G) \geq W(\mathcal{E}_i|u_i) \ \text{ for } \ i = 1, 2, \dots, M.$$

This gives (iv):

$$\sigma(G) \geq \sum_{i=1}^{M} W(\mathcal{E}_i|u_i) > M\delta'.$$

Further, (12.8) is equivalent to $W(G_x|x) - W\left(\bigcup_{i=1}^{M} \mathcal{E}_i \Big| x\right) < \delta'$, which implies

$\sum_x P_X(x)W(G_x|x) - \sum_x P_X(x)W\left(\bigcup_{i=1}^{M} \mathcal{E}_i \Big| x\right) = P_{XY}(G) - P_Y\left(\bigcup_{i=1}^{M} \mathcal{E}_i\right) < \delta'$ and hence (v).

Finally, by definition of ρ for $y \in \mathcal{E}_i \subset G_{u_i}$

$$W(y|u_i) \geq \rho(G)P_Y(y)$$

and by (iv)

$$\frac{1}{M}W(y|u_i) \geq \delta'\sigma(G)^{-1}\rho(G)P_Y(y).$$

Therefore

$$Q(y) \geq \delta'\sigma(G)^{-1}\rho(G)P_Y(y)$$

for all $y \in \bigcup_{i=1}^{M} \mathcal{E}_i$.

The freedom in the choice of G or even several G's makes the power of this approach. We explain this in Sections 13, 14 and 15.

Obviously, we get good bounds, if $\rho(G)$ and $\sigma(G)$ are close to each other. We achieve this with our next idea to partition

$$G_{XY} = \big\{(x,y) \in \mathcal{X} \times \mathcal{Y} : P_{XY}(x,y) > 0\big\}$$

into informationally balanced sets and a set with big value of ρ, which we exclude.

Introduce

$$\begin{aligned} G(I+\beta) &= G_{XY}(I+\beta) \\ &= \left\{(x,y) \in G_{XY} : \log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} < I(X \wedge Y) + \beta\right\} \end{aligned}$$

and for suitable θ, L, to be specified below, the partition

$$\begin{aligned} G_{XY}(I+\beta) &= \bigcup_{\ell=0}^{L-1} G_{XY}^{\ell}(I+\beta), \quad \text{where} \\ G_{XY}^{\ell}(I+\beta) &= G_{XY}(I+\beta-\ell\theta) - G_{XY}\big(I+\beta-(\ell+1)\theta\big). \end{aligned}$$

Its atoms are balanced, because

$$\frac{\sigma\big(G_{XY}^{\ell}(I+\beta)\big)}{\rho\big(G_{XY}^{\ell}(I+\beta)\big)} \leq e^{\theta}.$$

For the further analysis we need a simple fact about I–divergences.

Lemma 6. *For any PD's p, q on $\mathcal{Z}$ and any $\mathcal{Z}' \subset \mathcal{Z}$*

$$\sum_{z\in\mathcal{Z}'} p(z)\log\frac{p(z)}{q(z)} \geq -e^{-1}\log_2 e = -c, \quad \textit{say.}$$

Proof:

$$\begin{aligned}
\sum_{z\in\mathcal{Z}'} p(z)\log\frac{p(z)}{q(z)} &= p(\mathcal{Z}')\sum_{z\in\mathcal{Z}'}\frac{p(z)}{p(\mathcal{Z}')}\log\frac{p(z)/p(\mathcal{Z}')}{q(z)/q(\mathcal{Z}')} + p(\mathcal{Z}')\log\frac{p(\mathcal{Z}')}{q(\mathcal{Z}')}\\
&\geq p(\mathcal{Z}')\log\frac{p(\mathcal{Z}')}{q(\mathcal{Z}')} \quad \text{(by nonnegativity of } I\text{–divergence)}\\
&\geq p(\mathcal{Z}')\log p(\mathcal{Z}')\left(\text{since} \quad \log\frac{1}{q(\mathcal{Z}')} \geq 1\right)\\
&\geq \min_{0\leq t\leq 1} t\log t = -e^{-1}\log_2 e.
\end{aligned}$$

We apply this fact to the PD's P_{XY} and $P_X \times P_Y$ and

$$\mathcal{Z}' = \left\{(x,y) : \log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \leq 0\right\}.$$

Thus

$$0 \geq \sum_{(x,y)\in\mathcal{Z}'} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \geq -c$$

and therefore

$$\big(1 - P_{XY}\big(G(I+\beta)\big)\big)(I+\beta) - c \leq I$$

or

$$P_{XY}\big(G(I+\beta)\big) \geq \frac{\beta - c}{\beta + I}.$$

We can choose ℓ such that

$$P_{XY}\big(G^\ell_{XY}(I+\beta)\big) \geq \frac{\beta - c}{(\beta+I)L}. \tag{12.11}$$

The set $G^\ell_{XY}(I+\beta)$ serves as our representation of P_{XY}.

Lemma 7. *For any distribution P_{XY} and set $D \subset \mathcal{Y}$ with $P_Y(D) = \sum_{x\in\mathcal{X}} P_X(x)W(D|x) \geq 1-\lambda$ consider for any $\beta > 0$ and positive integer L the representative $G^\ell_{XY}(I+\beta)$. Then we have for $G = G^\ell_{XY}(I+\beta) \cap \mathcal{X} \times D$*

(i) $P_{XY}(G) \geq \frac{\beta-c}{(\beta+I)L} - \lambda = \delta$, *say.*

For any $\delta' < \delta$ there is a code

$$\big\{(u_i, \mathcal{E}_i) : 1 \le j \le M\big\} \quad \textit{with} \quad \mathcal{E}_j \subset G_{u_j} \subset D \quad \textit{for} \quad j = 1, \dots, M$$

and the properties

(ii) $M \le \frac{1}{\delta'} e^{I+\beta-\ell\cdot\theta}$

(iii) $P_Y\left(\bigcup_{i=1}^{M} \mathcal{E}_i\right) \ge \delta - \delta'$

(iv) $\frac{1}{M} \sum_{j=1}^{M} W(y|u_j) \ge \delta' e^{-\theta} P_Y(y)$ *for* $y \in E = \bigcup_{i=1}^{M} \mathcal{E}_i$

(v) $\frac{1}{M} \sum_{j=1}^{M} W(E|u_j) \ge \delta' e^{-\theta}(\delta - \delta') = \delta^*$*, say.*

Proof: (i) is a consequence of (12.11) and the assumption on D. Inequality (ii) follows from (iv) in Lemma 5 and inequality (iii) follows from (v) in Lemma 5 (and (i) above). Finally, this and (vi) in Lemma 5 give (iv) and (v).

Theorem 9. *Let the discrete (not necessarily memoryless) channel* $W^n : \prod_1^n \mathcal{X} \to \prod_1^n \mathcal{Y}$ *have an* (n, N, λ_n) *identification code* $\{(P_i, \mathcal{D}_i) : 1 \le i \le N\}$*, then*

$$\log\log N \le \max_i I(X_i^n \wedge Y_i^n) + o(n)$$

if $\lambda_n \le n^{-7}$*.*

Proof: Consider any pair $(P_i, \mathcal{D}_i)$ and apply Lemma 7 for $D = \mathcal{D}_i$, $P_X = P_i$. However, we write now P_{X^n} instead of P_X. Also, for $P_i \times W^n$ we write $P_{X^nY^n}$ (instead of P_{XY}) and thus we write the representation for $P_{X^nY^n}$ as $G = G^\ell_{X^nY^n}(I+\beta) \cap (\mathcal{X}^n \times D)$.

Our goal is to choose parameters so that M in (ii) of Lemma 7 becomes small and δ^* in (v) of Lemma 7 becomes large. The first property guarantees that $\binom{|\mathcal{X}^n|}{M}$ is so small that the number of representing encoding sets $\{u_j : 1 \le j \le M\}$ meets the desired double exponential bound.

The second property insures an appropriate bound on the multiplicity of representing encoding sets.

Accordingly the proof goes in two steps.

Step 1: We choose $\beta = \varepsilon n$ and for convenience we choose $\delta' = \delta/2$. Clearly, for n large by (12.11)

$$P_{XY}\big(G^\ell_{XY}(I+\beta)\big) \ge \frac{\beta}{(\beta+I)2L}. \tag{12.12}$$

We choose $\theta = \frac{\beta+I}{2L}$.

Using (12.12) and Lemma 7 (i) we get now

$$\delta_n^* \geq \frac{1}{4}\left(\frac{\beta}{(\beta+I)2L} - \lambda_n\right)^2 e^{-(I+\beta)/2L}.$$

Since $I = I(X^n \wedge Y^n) \leq n \log |\mathcal{X}|$, we get

$$\delta_n^* \geq \frac{1}{4}\left(\frac{\varepsilon}{(\log|\mathcal{X}|+\varepsilon)2L} - \lambda_n\right)^2 e^{-(\log|\mathcal{X}|+\varepsilon)(2L)^{-1}n}.$$

Notice that for any function $f(n) \to \infty (n \to \infty)$ the choice $L = L_n = n\ f(n)$ yields $\lim\limits_{n\to\infty} e^{-(\log|\mathcal{X}|+\varepsilon)L_n^{-1}n} = 1$ and the choices $f(n) = n^{1/2}$, $L_n = n^{3/2}$, $\lambda_n = n^{-7}$ yield $\delta_n^* \geq n^{-4}$ for n large.

These are not optimal calculations, but only polynomial growth and the fact $\delta_n^* \gg \lambda_n$ are relevant here!

By our choices and Lemma 7 (ii) – (v) $\delta \geq \lambda_n$ and

$$M \leq 2n^3\ e^{I(X_i^n \wedge Y_i^n)+\varepsilon n}. \tag{12.13}$$

This is the first desired property. The others are

$$P_Y\left(\bigcup_{i=1}^M \mathcal{E}_i\right) \geq \frac{\delta}{2} \geq \frac{1}{4}n^{-3/2}. \tag{12.14}$$

$$Q_{\mathcal{U}}(y) = \frac{1}{M}\sum_{j=1}^M W(y|u_i) \geq \frac{1}{2}n^{-3}P_Y(y) \tag{12.15}$$

and so

$$Q_{\mathcal{U}}\left(\bigcup_{i=1}^M \mathcal{E}_i\right) \geq \frac{1}{8}n^{-9/2}, \tag{12.16}$$

which is much bigger than $\lambda_n = n^{-7}$.

Step 2: If now $\mathcal{U}$ serves $K' \geq K$ other times as representative for $(P_{Y^j}, \mathcal{D}_{Y^j})$ with decoding sets $\{\mathcal{E}_i^j : 1 \leq i \leq M\}$, $j = 1, \ldots, K'$, then K' can be suitably bounded.

Indeed, set $\mathcal{E}^j = \bigcup\limits_{j=1}^M \mathcal{E}_i^j$ and define disjoint sets

$$\mathcal{E}'^j = \mathcal{E}^j - \bigcup_{j' \neq j} \mathcal{E}^{j'}; j = 1, 2, \ldots, K. \tag{12.17}$$

Since $\mathcal{E}^j \subset \mathcal{D}_{Y^j}$ and the identification code has error probabilities less than λ_n, we get from (12.14)

$$P_{Y^j}(\mathcal{E}'^j) \geq \frac{1}{4} n^{-3/2} - K\lambda_n \tag{12.18}$$

and thus by (12.15)

$$Q_{\mathcal{U}}\left(\bigcup_{j=1}^{K} \mathcal{E}'^j\right) = \sum_{j=1}^{K} Q_{\mathcal{U}}(\mathcal{E}'^j) \geq K\left(\frac{1}{4} n^{-3/2} - K\lambda_n\right) \cdot \frac{1}{2} n^{-3}.$$

Now for $K = 16\, n^{9/2}$ and $\lambda_n < \frac{1}{128} n^{-6}$ we have $\frac{1}{4}\, n^{-3/2} - K\lambda_n > \frac{1}{8}\, n^{-3/2}$ and thus

$$Q_{\mathcal{U}}\left(\bigcup_{j=1}^{K} \mathcal{E}'^j\right) > 1, \text{ a contradiction.}$$

So $\mathcal{U}$ serves at most $16\, n^{9/2}$ times as representative and the result follows with (12.13).

13 Two promissed results: characterisation of the capacity regions for the MAC and the BC for identification

We know from [3], [47] that the transmission capacity region $\mathcal{R}$ of a MAC $W : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ can be characterized as the convex hull of the set of pairs $(R_{\mathcal{X}}, R_{\mathcal{Y}})$ of non–negative numbers which satisfy for some input distribution $P_{XY} = P_X \times P_Y$

$$\begin{aligned} R_{\mathcal{X}} &\leq I(X \wedge Z|Y) \\ R_{\mathcal{Y}} &\leq I(Y \wedge Z|X) \\ R_{\mathcal{X}} + R_{\mathcal{Y}} &\leq I(XY \wedge Z). \end{aligned} \tag{13.1}$$

Also, in [3] there is a non–single letter characterisation.

$$\mathcal{R} = \left\{ \left(\frac{1}{n} I(X^n \wedge Z^n), \frac{1}{n} I(Y^n \wedge Z^n)\right) : n \in \mathbb{N}, P_{X^nY^n} \in \mathcal{P}(\mathcal{X}^n \times \mathcal{Y}^n), P_{X^nY^n} = P_{X^n} \times P_{Y^n} \right\}. \tag{13.2}$$

Quite surprisingly we can use this characterisation for the proof of the polynomial weak converse for identification via the MAC.

Theorem 10. *The second order identification capacity region for the MAC* equals *the first order transmission capacity region $\mathcal{R}$.*

The broadcast channel is a stochastic map

$$W : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}$$

with components $W_1 : \mathcal{X} \to \mathcal{Z}$ and $W_2 : \mathcal{X} \to \mathcal{Z}$ and the object space is

$$\mathcal{N} = \mathcal{N}_1 \times \mathcal{N}_2, |\mathcal{N}_i| = N_i \ (i = 1, 2,).$$

An identification code (n, N_1, N_2, λ) for the BC is a family

$$\big\{(P_{ij}, \mathcal{D}_i, \mathcal{F}_j) : 1 \leq i \leq N_1; 1 \leq j \leq N_2\big\},$$

where the $\mathcal{D}_i$'s are sets in $\mathcal{Y}^n$, the $\mathcal{F}_j$'s are sets in $\mathcal{Z}^n$ and $P_{ij} \in \mathcal{P}(\mathcal{X}^n)$, and

$$\sum_{x^n} W_1(\mathcal{D}_i|x^n) P_{ij}(x^n) \geq 1 - \lambda \ \text{ for all } \ i \ \text{ and } \ j \tag{13.3}$$

$$\sum_{x^n} W_1(\mathcal{D}_{i'}|x^n) P_{ij}(x^n) \leq \lambda \ \text{ for all } \ i \neq i' \ \text{ and all } \ j \tag{13.4}$$

$$\sum_{x^n} W_2(\mathcal{F}_j|x^n) P_{ij}(x^n) \geq 1 - \lambda \ \text{ for all } \ j \ \text{ and } \ i \tag{13.5}$$

$$\sum_{x^n} W_2(\mathcal{F}_{j'}|x^n) P_{ij}(x^n) \leq \lambda \ \text{ for all } \ j \neq j' \ \text{ and all } \ i. \tag{13.6}$$

Let $\mathcal{B}$ be the set of all achievable pairs of second order rates.

Our key observation is that for identification we can relate the capacity regions for identification of independent messages to the capacity region for identification for degraded message sets, $\mathcal{B}_\mathcal{Y}$ and $\mathcal{B}_\mathcal{Z}$.

Lemma 8 (Reduction). *For $R_\mathcal{Y} \leq R_\mathcal{Z}$ (resp. $R_\mathcal{Y} \geq R_\mathcal{Z}$) $(R_\mathcal{Y}, R_\mathcal{Z}) \in \mathcal{B}$ iff $(R_\mathcal{Y}, R_\mathcal{Z}) \in \mathcal{B}_\mathcal{Y}$ (resp. $\mathcal{B}_\mathcal{Z}$).*

Proof: Let $\big\{(P_{ij}, D_i, E_i) : 1 \leq i \leq N_\mathcal{Y}, 1 \leq j \leq N_\mathcal{Z}\big\}$ be an identification code for the BC with error probabilities $\leq \lambda$. Since $R_\mathcal{Y} \leq R_\mathcal{Z}$ we can define for

$$\ell = 1, \ldots, N_\mathcal{Y} \ \text{ and } \ m = 1, \ldots, \frac{N_\mathcal{Z}}{N_\mathcal{Y}}$$

$$Q_{\ell,m} = P_{\ell,(m-1)N_\mathcal{Y}+\ell}.$$

The $\mathcal{Y}$–decoder identifies ℓ and the $\mathcal{Z}$–decoder identifies $(m-1)N_\mathcal{Y} + \ell$ or equivalently ℓ **and** m, that is, the common part and a separate part.

If $\mathcal{R}_\mathcal{Y} < \mathcal{R}_\mathcal{Z}$, then $2^{2^{R_z n}} \cdot 2^{-2^{R_y n}} \sim 2^{2^{R_z n}}$.

If $R_\mathcal{Y} = R_\mathcal{Z}$, then we can make the same construction with rates $R_\mathcal{Y} - \varepsilon$ and $R_\mathcal{Z}$.

We need the direct part of the ABC Coding Theorem ([5], [31], [30]). Here, there are separate messages for decoder $\mathcal{Y}$ (resp. $\mathcal{Z}$) and common messages for both decoders.

Achievable are (with maximal errors)

$$\begin{aligned}\mathcal{A}_\mathcal{Y} = \big\{(R_0, R_1) : R_0 + R_1 \leq \min\big[I(X \wedge Y), I(X \wedge Y|U) + I(U \wedge Z)\big], R_0 \leq I(U \wedge Z),\\ U \ominus X \ominus YZ, \ \ \|\mathcal{U}\| \leq |\mathcal{X}| + 2\big\}.\end{aligned}$$

resp.

$$\mathcal{A}_{\mathcal{Z}} = \big\{(R_0, R_2) : R_0 + R_2 \leq \min\big[I(X \wedge Z), I(X \wedge Z|U) + I(U \wedge Y)\big], R_0 \leq I(U \wedge Y),$$
$$U \ominus X \ominus YZ, \quad \|\mathcal{U}\| \leq |\mathcal{X}| + 2\big\}.$$

Here is our surprising result.

Theorem 11. *For the (general) BC the set of achievable pairs of second order rates is given by*

$$\mathcal{B} = \mathcal{A}_{\mathcal{Y}} \cup \mathcal{A}_{\mathcal{Z}}.$$

14 The proof for MAC

The proof of achievability is straightforward by the second method of Ahlswede/Dueck [16] or the method of Section 18. Indeed, use an *average* error transmission code in blocklength n

$$\big\{(u_i, v_j, \mathcal{D}_{ij}) : 1 \leq i \leq M_{\mathcal{X}}, 1 \leq j \leq M_{\mathcal{Y}}\big\}$$

with

$$\frac{1}{M_{\mathcal{X}}} \frac{1}{M_{\mathcal{Y}}} \sum_{i,j} W(\mathcal{D}_{ij}^c | u_i, v_j) \leq \lambda. \tag{14.1}$$

Then of course also

$$\frac{1}{M_{\mathcal{X}}} \sum_i \left(\frac{1}{M_{\mathcal{Y}}} \sum_j W\left(\Big(\bigcup_{j'} \mathcal{D}_{ij'} \Big)^c | u_i, v_j \right) \right) \leq \lambda \tag{14.2}$$

and we have a random experiment U with $\mathrm{Prob}(U = u_i) = \frac{1}{M_{\mathcal{X}}}$, whose outcome is known to sender $S_{\mathcal{X}}$ and with probability at least $1 - \lambda$ also to the receiver.

Analogously, there is a random experiment V for the sender $S_{\mathcal{Y}}$ and the receiver. We have used blocklength n.

As in [16] with relatively few, say $\sqrt{n}$, letters (actually even $o(\log n)$) identification of *second* order rate $\sim \frac{1}{n} \log M_{\mathcal{X}}$ can be performed from $S_{\mathcal{X}}$ to the receiver. Finally, with other $\sqrt{n}$ letters the identification of *second* order rate $\sim \frac{1}{n} \log M_{\mathcal{Y}}$ can be done from $S_{\mathcal{Y}}$ to the receiver.

Remark 18: In our proof of the direct part the identification is done seperately for both encoders. The encoding strategy pair $(\mathcal{P}_i, Q_j)$ and the decodings $\mathcal{D}_i, \mathcal{F}_j$ identify i and j seperately. We can also choose $\mathcal{E}_{ij} = \mathcal{D}_i \cap \mathcal{F}_j$ and notice that

$$\sum_{x^n, y^n} W(\mathcal{E}_{ij} | x^n, y^n) P_i(x^n) Q_j(y^n) > 1 - 2\lambda$$

$$\sum_{x^n, y^n} W(\mathcal{E}_{i'j'}^c (x^n, y^n) P_i(x^n) Q_j(y^n) \leq 2\lambda \;\text{ for }\; (i', j') \neq (i, j).$$

On the other hand, starting with the $\mathcal{E}_{ij}$'s we can define $\mathcal{D}_i = \bigcup_j \mathcal{E}_{ij}$, $\mathcal{F}_j = \bigcup_i \mathcal{E}_{ij}$.

Remark 19: The decomposition principle does not hold for identification on the MAC. If both encoders have independent messages, but can cooperate, then

$$R_{\mathcal{X}\mathcal{Y}} = \max_{P_X \times P_Y} I(XY \wedge Z)$$

and $2^{2^{n\,R_{\mathcal{X}\mathcal{Y}}}}$ is **much** bigger than

$$2^{2^{n\,R_{\mathcal{X}}}} \cdot 2^{2^{n\,R_{\mathcal{Y}}}} \sim 2^{2^{n\,\max(R_{\mathcal{X}},R_{\mathcal{Y}})}}.$$

Problems:

5. Develop a theory for identification of correlated data (see "correlated codes" in [12]).
6. Develop approximation of output statistics for the MAC to obtain a strong converse. Use random coding instead of maximal coding with rates

$$\begin{aligned} I(X \wedge Z) &\leq R_{\mathcal{X}} \leq I(X \wedge Z|Y) \\ I(Y \wedge Z) &< R_{\mathcal{Y}} \leq I(Y \wedge Z|X) \\ I(XY \wedge Z) &\leq R_{\mathcal{X}} + R_{\mathcal{Y}} \end{aligned}$$

and code structure $\{u_1, \dots, u_{M_1}\}$ and $\{v_{i1}, \dots, v_{iM_2}\}$ for $i = 1, \dots, M_1$.

Converse proof: We follow closely the proof for a one–way channel. Here it is essential that our approach treats general channels with memory. Secondly we use the characterisation (13.2) of the rate–region $\mathcal{R}$ for the MAC.

In addition we partition our encoding pairs $(P_i \times Q_j)_{\substack{i=1,\dots,M_1 \\ j=1,\dots,M_2}}$ according to their corresponding pairs of mutual informations $\big(I(X_i^n \wedge Z_{ij}^n), I(Y_j^n \wedge Z_{ij}^n)\big)$ where $P_{X_i^n} = P_i$, $P_{Y_j^n} = Q_j$, $P_{Z_{ij}^n} = P_i \times Q_j W$, as follows.

Endow $\mathbb{R}^2$ and, particularly,

$$S = \big\{(r_1, r_2) : 0 \leq r_1 \leq \log|\mathcal{X}|, 0 \leq r_2 \leq \log|\mathcal{Y}|\big\}$$

with a rectangular lattice with side lengths η. So we get $g(\eta) = g_1(\eta) \cdot g_2(\eta)$ rectangles, if $g_1(\eta) = \frac{\log|\mathcal{X}|}{\eta}$, $g_2(\eta) = \frac{\log|\mathcal{Y}|}{\eta}$.

Label them as $S_{a,b}(1 \leq a \leq g_1(\eta), 1 \leq b \leq g_2(\eta)$ and associate with $P_i \times Q_j$ the rectangle $S_{a(i,j),b(i,j)}$, where

$$\frac{1}{n} I(X_i^n \wedge Z_{ij}^n), \frac{1}{n} I(Y_j^n \wedge Z_{ij}^n) \in S_{a(i,j),b(i,j)}. \tag{14.3}$$

There is a rectangle S^* with which at least $\frac{M_1 \cdot M_2}{g(\eta)}$ encodings $P_i \times Q_j$ are associated. Denote them by $(P_i \times Q_j)_{(i,j) \in \mathcal{M}(\eta)}$.

Their corresponding pairs of (normalized) mutual informations differ componentwise by at most η.

Furthermore, there is a row index i^* and a column index j^* so that

$$|\{(i^*,j):(i^*,j)\in\mathcal{M}(\eta)\}\geq\frac{|\mathcal{M}(\eta)|}{M_1}\geq\frac{M_2}{g(\eta)}, \tag{14.4}$$

$$|\{(i,j^*):(i,j^*)\in\mathcal{M}(\eta)\}\geq\frac{|\mathcal{M}(\eta)|}{M_2}\geq\frac{M_1}{g(\eta)}. \tag{14.5}$$

Now our previous converse proof comes in. To every triple $(P_i,Q_j,\mathcal{D}_{ij})$ we assign two codes $(\mathcal{U}_i^j,\mathcal{E}_i^j),(\mathcal{V}_j^i,\mathcal{F}_j^i)$, where $\mathcal{U}_i^j\subset\mathcal{X}^n$, $\mathcal{E}_i^j=\{E_{i1}^j,\dots,E_{iM_i^j}^j\}$, (pairwise disjoint), $\mathcal{V}_j^i\subset\mathcal{Y}^n$, $\mathcal{F}_j^i=\{F_{j1}^i,\dots,F_{jN_i^j}^i\}$ (pairwise disjoint), and all decoding sets are subsets from D_{ij}. Here

$$M_i^j\leq\exp\left\{\frac{1}{n}I(X_i^n\wedge Z_{ij}^n)+o(n)\right\}$$

$$N_j^i\leq\exp\left\{\frac{1}{n}I(Y_j^n\wedge Z_{ij}^n)+o(n)\right\}$$

and (14.3) holds.

Moreover, for all indices

$$\frac{1}{M_i^j}\sum_{u\in\mathcal{U}_i^j}\sum_{y^n}W(E_{in}^j\cap\mathcal{D}_{ij}|u,y^n)Q_j(y^n)\geq n^{-4} \tag{14.6}$$

and analogous relations hold for $\mathcal{V}_j^i$.

Now observe that for all $(i,j)\in\mathcal{M}(\eta)$

1.) $\frac{1}{n}\log M_i^j\leq R_1^*+\eta$
 $\frac{1}{n}\log N_j^i\leq R_2^*+\eta$

2.) By (14.4), (14.5) there are at most $\binom{|\mathcal{X}|^n}{2^{(R_1^*+\eta)n}}$ different codes $\mathcal{U}_{i^*}^j$ in row i^* and at most $\binom{|\mathcal{Y}|^n}{2^{(R_2^*+\eta)n}}$ codes $\mathcal{V}_{j^*}^i$ in column j^*.

Furthermore the multiplicity K_{i^*} of codes in row i^* (resp. K_{j^*} for column j^*) does not exceed n^6 (as previously).

Finally, therefore

$$\frac{1}{n}\log\log M_i\leq R_i^*+2\eta\ (i=1,2).$$

15 The proof for BC

The direct part: We use the Reduction Lemma and the ABC Coding Theorem mentioned in Section 13. Eventhough they hold also for maximal errors we use average errors so that the codes establish two common random experiments of the sender with both receivers, resp.

The converse part: Suppose w.l.o.g. that $R_{\mathcal{Y}} < R_{\mathcal{Z}} + \varepsilon$, ε arbitrarily small, and that the $\mathcal{Z}$–decoder has a separate part coded in two column number with the encodings $(P_{ij})_{\substack{i=1,\dots,M_1\\ j=1,\dots,M_2}}$.

We associate RV's and information quantities as follows:

X_{ij} has distribution $P_{X_{ij}} = P_{ij}$.

This generates output variables Z_{ij} with values in $\mathcal{Z}^n$. Let U_i be auxiliary RV's with

$$\text{Prob}(U_i = j) = \frac{1}{M_2}$$

and

$$\text{Prob}(\cdot|U_i = j) = P_{ij}.$$

Thus we get quantities

$$I(U_i \wedge Y_i^n), I(X_i^n \wedge Z_i^n|U_i), I(X_i^n \wedge Z_{ij}^n), \quad \text{and} \quad I(X_i^n \wedge Z_i^n).$$

As in the proof of Theorem 10 make η–approximations first for all $I(X_{ij}^n \wedge Z_{ij}^n)$ with biggest class of value I_{η_3}.

This gives as in the one–way channel coding theorem

$$\frac{1}{n} \log\log M_1 \cdot M_2 \sim \frac{1}{n} \log\log M_2 \leq I_{\eta_3} \tag{15.1}$$

In the remaining matrix keep I_{η_2} for $I(X_i^n \wedge Z_i^n|U_i)$ and then all $I(U_i \wedge Y_i^n)$ equal to I_{η_1}.

We upper bound the number of rows by upper bounding the number of codes (via Lemma 7) representing triples $(P_{U_i}, P_{Y_i^n|U_i}, \mathcal{D}_i)$. Thus for $\lambda_n = n^{-6}$ (as usual)

$$\frac{1}{n} \log\log M_1 \leq I_{\eta_1} + 2\eta. \tag{15.2}$$

Within row i^* a significant number of terms has

$$\frac{1}{n} I(X^n \wedge Z^n|U_{i^*} = j) \leq I_{\eta_2} + \beta^*.$$

This gives the desired column number estimate

$$\frac{1}{n} \log\log M_2 \leq \min(I_{\eta_1} + I_{\eta_2}, I_{\eta_3}).$$

Common randomness with two receivers

rates $\big(I(U \wedge Y), \min\big(I(X \wedge Z|U) + I(U \wedge Y), I(X \wedge Z)\big)\big)$.

$$\begin{aligned}\mathcal{B} &= \mathcal{B}_Y \cup \mathcal{B}_Z \\ &= \big\{(R_\mathcal{Y}, R_\mathcal{Z}) : \exists (R_0, R_1) \in \mathcal{A}_\mathcal{Y} : R_\mathcal{Y} \le R_0 + R_1, R_\mathcal{Z} \le R_1\big\} \\ &\cup \big\{(R_\mathcal{Y}, R_\mathcal{Z}) : \exists (R_0, R_1) \in \mathcal{A}_\mathcal{Z} : R_\mathcal{Y} \le R_0, R_\mathcal{Z} \le R_0 + R_1\big\}.\end{aligned}$$

Part V: Data compression

16 Noiseless coding for identification — identification entropies

Let $(\mathcal{N}, P)$ be a source, where $\mathcal{N} = \{1, 2, \dots, N\}$, $P = (P_1, \dots, P_N)$, and let $\mathcal{U} = (u_1, \dots, u_N)$ be a prefix code (PC) for this source with $\ell(u_i)$ as length of u_i. Introduce the RV U with $\text{Prob}(U = u_i) = P_i$ for $i = 1, 2, \dots, N$.

We use the PC for noiseless identification, that is user i wants to know whether the source output equals i, that is, whether U equals u_i or not. He iteratively checks whether U coincides with u_i in the first, second, etc. letter. What is the number $F_i(\mathcal{U}, P)$ of checkings? In order to calculate this quantity we introduce for the binary tree $\mathcal{T}_\mathcal{U}$ whose leaves are the codewords $u_1, \dots, u_N$ the branches $B_{ik}(1 \le i \le N; 0 \le k)$, where $B_{ik} = \{u \in \mathcal{U} : u$ coincides with u_i exactly until the k'th letter of $u_i\}$.

For $U \in B_{ik}$, $0 \le k \le \ell(u_i) - 1$, the answers are k times "Yes" and 1 time "No". For $U = u_i$ the answers are $\ell(u_i)$ times "Yes". Thus

$$F_i(\mathcal{U}, P) = \sum_{k=0}^{\ell(u_i)-1} P(U \in B_{ik})(k+1) + \ell(u_i) P$$

and

$$F(\mathcal{U}, P) = \max_{1 \le i \le N} F_i(\mathcal{U}, P),$$
$$F(P) = \min_{\mathcal{U}} F(\mathcal{U}, P).$$

Analogously, if $\tilde{\mathcal{U}}$ is a randomized coding, we introduce

$$F_i(\tilde{\mathcal{U}}, P), F(\tilde{\mathcal{U}}, P) \quad \text{and} \quad F_R(P).$$

What are the properties of $F(P)$ and $F_R(P)$? They are a kind of "identification entropies".

Moreover, as in [44] one can replace the lengths ℓ_i by $\varphi(\ell_i)$ where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and strictly monotone increasing.

Thus one gets functionals
$$F(P, \varphi) \text{ and } F_R(P, \varphi).$$

We shall analyse these quantities on another occasion and confine ourself here to deriving some simple facts.

Let us start with $P_N = \left(\frac{1}{N}, \dots, \frac{1}{N}\right)$ and set $f(N) = F(P_N)$. Clearly
$$f(2^n) \le 1 + \frac{1}{2} f(2^{n-1}), f(2) = 1$$

and therefore
$$f(2^n) \le 2 - 2^{-(n-1)}. \tag{16.1}$$

On the other hand it can be verified that
$$f(9) = 1 + \frac{10}{9} > 2 \text{ and more generally, } f(2^{n+1}) > 2.$$

1. What is $\sup_N \big(f(N)\big)$?
2. Is $F_R(P) \le 2$?
3. Suppose that encoder and decoder have access to a random experiment with unlimited capacity of common randomness (see [46]). Denote the best possible average codewords lengths by $F^*(P)$.

For $P = (P_1, \dots, P_N)$, $N \le 2^n$ write $P' = (P_1, \dots, P_N, 0, \dots, 0)$ with 2^n components. Use a binary regular tree of depth n with leaves $1, 2, \dots, 2^n$ represented in binary expansions.

The common random experiment with 2^n outcomes can be used to use 2^n cyclic permutions of $1, 2, \dots, 2^n$ for 2^n deterministic codes. For each i we get equally often 0 and 1 in its representation and an expected word length $\overline{\ell_i} \le 2 - \frac{1}{2^{n-1}}$. The error probability is 0. Therefore $F^*(P) \le 2 - 2^{-(n-1)} \le 2$ for all P.

17 Noiseless coding for multiple purposes

In classical theory of data compression the main concern is to achieve a short average length coding. Here we adress a problem of noiseless coding, where different persons are interested in different aspects of the data and their accessibility. We begin with a specified question.

17.1 Persons are interested in different components of a Bernoulli Source

Consider a discrete memoryless binary, symmetric source (BSC) producing the output $X^n = (X_1, \dots, X_n)$. Suppose that there are n persons and that person t is interested in the outcome of $X_t (1 \leq t \leq n)$. A multiple purpose encoding (or program) shall be a sequence $f = (f_i)_{i=1}^{\infty}$ of functions $f_i : \{0,1\}^n \to \{0,1\}$. Person t requests sequentially the values $f_1(X^n), f_2(X^n), \dots$ and stops as soon as he/she has identified the value of X_t. Let $\ell(f,t)$ denote the number of requests of person t for program f. We are interested in the quantity

$$L(n) = \min_f \max_{1 \leq t \leq n} \mathbb{E}\ell(f,t). \tag{17.1}$$

The choice $f_i(X^n) = X_i (1 \leq i \leq n)$ gives $\ell(f,t) = t$ and thus $\max\limits_{1 \leq t \leq n} \ell(f,t) = n$.

Since $\frac{1}{n} \sum\limits_{t=1}^{n} \ell(f,t) = \frac{n+1}{2}$, one should do better. In [15] we stated the problem to determine $L(n)$. Don Coppersmith [26] gave a rather precise bound.

Theorem 12. $\frac{n+1}{2} \leq L(n) \leq \frac{n+2}{2}$.

Proof: The lower bound is obvious, because

$$L(n) \geq \min_f \frac{1}{n} \sum_{t=1}^{n} \ell(f,t)$$

and

$$\mathbb{E}|\{t : 1 \leq t \leq n, \ell(f,t) \leq i\}| \leq i.$$

For the upper bound set $f_1(X^n) = X_1$ and for $2 \leq i \leq n$ set $f_i(X^n) = \begin{cases} X_i & \text{if } X_1 = 0 \\ X_{n+2-i} & \text{if } X_1 = 1. \end{cases}$

For $t > 1$ the stopping time is either t or $n+2-t$, each with probability $\frac{1}{2}$, so that the mean is $\ell(f,t) = \frac{n+2}{2}$, while obviously $\ell(f,1) = 1$. Thus $L(n) \leq \frac{n+2}{2}$.

Remark 20: A weaker upper bound, but more uniform distribution of the stopping times is obtained as follows: Let the first $\lceil \log_2 n \rceil$ bits be $\big(f_1(X^n), f_2(X^n), \dots, f_{\lceil \log n \rceil}(X^n)\big) = (X_1, X_2, \dots, X_{\lceil \log n \rceil})$ and let these $\log n$ bits index a cyclic shift of the remaining $n - \log n$ bits so that the distribution of stopping times is approximately uniform between $\log n$ and n for $t > \lceil \log n \rceil$. This leads to the weaker upper bound

$$L(n) \leq (n + \log_2 n + c)/2.$$

Remark 21: Notice that both procedures are probabilistic algorithms. They exploit the randomness of the source.

17.2 Noiseless source coding problems of infinite order: Ordering and Identification

We consider here a source coding version of the ordering problem and also of the identification problem.

To simplify technicalities we assume that $N = 2^n$. We also assume that any element of $\{0,1\}^n$ is source output with equal probabilities.

For any $u^n \in \{0,1\}^n$: Is the source output $x^n = (x_1, x_2, \dots, x_n)$ before u^n, that is, $x^n \le u^n$ (lexicographically), or not? There is a canonical encoding function $f = (f_1, \dots, f_n)$ with $f_t(X_1, \dots, X_n) = X_t$. The person interested in u^n stops, when his/her question is answered. He/she stops at the smallest t with $f_t(u_t) \neq f_t(X_t)$.

The distributions of the stopping times don't depend on u^n. Let T_n denote the expected stopping time.

Lemma 10. $T_n = 1 + \frac{1}{2}T_{n-1} = \frac{2^n-1}{2^{n-1}}$, $n \ge 1$.

This is a simple exercise. Notice that

$$\lim_{n\to\infty} T_n = 2. \tag{17.2}$$

So the compression rate exceeds any finite order.

Now let the question be "Does X^n equal u^n or not?" (Identification)

We use again a multi–purpose encoding function. Actually we can use the same function as before. There is also the same recursion for T_n. Notice that in case of identification for $X^n = u_n$ we have maximal running time, namely n.

17.3 Problems

7. It is interesting to study the previous problems for other distributions on $\{0,1\}^n$. In general the previous encoding function is not optimal (for instance if $\text{Prob}(X_1 = 0) = 1$).

   An instructive source is given by the distribution which assigns probability $\frac{1}{n}$ to the sequences starting with k 1's and continuing with 0's only. For $u^n = (1, 1, \dots, 1)$ the running time of the previous encoding function is always n. However, by choosing $f_1(X^n) = X_{\lceil \frac{n}{2} \rceil}$ etc. the worst case expected running time is still less than 2.

8. For any distribution P on $\{0,1\}^n$, is the worst case expected running time less than 2? In case the answer is negative, determine the best constant (independent of n) upper bound! An obvious algorithm: number probabilities in decreasing order; $P_1 \ge P_2 \ge \dots \ge P_N$ and divide as equaly as possible $P_1 + P_2 + \dots + P_{N_1}$, $P_{N_1+1} + \dots + P_N$. $f_1(X^n)$ says whether $i \in \{1, \dots, N_1\}$ or not, etc.

   We *conjecture* that the bound 2 is achievable, if randomisation in the encoding is permitted. Two simple examples illustrate the advantage of randomisation. Denote by $E_{P,i}(f)$ the expected running time for source distribution P, object i, and encoding function f.

For $P = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ and f based on division into two equal parts gives

$$T_{P,i}(f) = 1 + \frac{1}{2} \ (i = 1, 2, 3, 4).$$

For $Q = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ and f based on the division $\left\{\left\{\frac{1}{3}\right\}, \left\{\frac{1}{3}, \frac{1}{3}\right\}\right\}$ gives

$$T_{Q,1}(f) = 1, \ T_{Q,i}(f) = 1 + \frac{2}{3} \ (i = 2, 3).$$

Therefore $\max_i T_{P,i}(f) < \max_i T_{Q,i}(f)$, however, $\sum_i T_{P,i}(f) > \sum_i T_{Q,i}(f)$ and randomisation takes advantage of this fact, by smoothing out the differences between the individual running times.

Let F choose with probabilities $\frac{1}{3}$ the partitions $\big\{\{1\}, \{2, 3\}\big\}$, $\big\{\{1, 3\}, \{2\}\big\}$, $\big\{\{3\}, \{1, 2\}\big\}$ in the first step, the second step is canonical. Then

$$T_{Q,i}(F) = \frac{1}{3}\left(1 + \left(1 + \frac{2}{3}\right) + \left(1 + \frac{2}{3}\right)\right) = 1 + \frac{4}{9} < 1 + \frac{1}{2}(!).$$

9. It is also reasonable to study alphabetical source codes for identification. For example for different intervals of a pipeline different repairman are responsible. They want to know whether a defect occured in their interval or not.

10. Suppose that $N = 2^k$ numbers are stored in 0–1 bits in a machine. Upon request a further bit is revealed by the machine. What is the average number of requests so that person i knows whether i occured or not?

11. One can study multiple purpose coding problems with noise (see [22], which gives a common generalisation of Shannon's noiseless coding theorem and coding theorems for noisy channels). What are the generalisations (there is one in [22]) of Kraft's inequality?

12. These source coding problems open a whole area of research. Are there coding problems of an order between first order (as in the component problem) and infinite order (as in the ordering problem)?

13. It is remarkable in this context also that the ordering problem *via channels* is not easier than transmission, if maximal errors are used. However, if for the second kind error probability the average is taken, then the ordering problem becomes of infinite order (similar as the identification problem does). Indeed just map the numbers $1, \dots, N$ onto codewords of a transmission code $\left\{(u_i, \mathcal{D}_i) : 1 \le i \le\right\}$ as follows:

    For any $K < N$ write $j \in \{1, \dots, N\}$ as $j = rK + s$, $0 \le s < K$, and map j on u_r. Now just let N go to infinity and choose $K = \left\lceil \frac{N}{M} \right\rceil$.

14. It is also interesting that for maximal second kind error probabilities the identification problem via channels has second order behaviour whereas — as mentioned before — the ordering problem has first order behaviour.

    We therefore ask the following question:

    Is there a reasonable coding problem with average error of second kind as performance criterion which is neither of first order nor of infinite order behaviour? In the positive case, what is the hierarchy of all orders?

15. If $\kappa < \frac{1}{2}C_{sh}$, then first order capacity R_1 equals infinity. However, if $\kappa > \frac{1}{2}C_{sh}$, is then $R_1 > C_{sh}$ possible?

Part VI: Perspectives

Our models go considerably beyond Shannon's transmission model and the model of identification. They will greatly enlarge the body of Information Theory. We substantiate here this believe by a brief discussion of how already the identification model alone had a significant impact.

Right now the most visible influences are new approximation problems (like approximation of output statistics [29] or entropy approximations based on Schur–convexity [27] etc.), a new emphasis on random number generation [47] and, above all, an understanding of the concept of common randomness [17], in identification ([27], [35], [48]), cryptography [45], and classical transmission problems of arbitrarily varying channels ([45], [41], [43]), and the paper [42], with a novel capacity formula, which could not be derived before.

It is also fascinating to discover how transmission problems and identification problems in multi–user theory show often some kind of duality. Often identification problems are mathematically more complex and in other cases we encounter the opposite: there is a rather *complete* capacity theory for identification via multi–way channels in case of complete feedback ([27]), whereas for transmission with feedback we don't even understand the multiple access channel.

We conclude with three more recently encountered directions of research.

18 Comparison of identification rate and common randomness capacity: Identification rate can exceed common randomness capacity and vice versa

One of the observations of [17] was that random experiments, to whom the communicators have access, essentially influence the value of the identification capacity C_I. Actually, if sender and receiver have a common random capacity CR (for a precise definition see [45]) then by the so called $\sqrt{n}$–trick of [17] always

$$C_I \geq C_R \ \text{ if } \ C_I > 0. \tag{18.1}$$

For many channels (see [17], [46]), in particular for channels with feedback ([17], [27]), equality has been proved.

It seemed therefore plausible, that this is always the case, and that the theory of identification is basically understood, when common random capacities are known.

We report here a result of [43], which shows that this expected unification is not valid in general — **there remain two theories**.

Example 6: $C_I = 1, C_R = 0$. (Fundamental)

(Actually, in [43] one can find also an example with $0 < C_I < C_R$)

We use a Gilbert type construction of error correcting codes with constant weight words. This was done for certain parameters in [16]. The same arguments give for parameters needed here the following auxiliary result.

Proposition 5. Let $\mathcal{Z}$ be a finite set and let $\lambda \in (0, 1/2)$ be given. For $\varepsilon < (2^{2/\lambda}+1)^{-1}$ a family $A_1, \dots, A_N$ of subsets of $\mathcal{Z}$ exists with the properties

$$|A_i| = \varepsilon|\mathcal{Z}|, |A_i \cap A_j| < \lambda\varepsilon|\mathcal{Z}| \ (i \neq j)$$

and

$$N \geq |\mathcal{Z}|^{-1} 2^{\lfloor \varepsilon |\mathcal{Z}| \rfloor} - 1.$$

Notice that $\lambda \log\left(\frac{1}{3}-1\right) > 2$ and that for ℓ with $2^{-\ell} = \varepsilon$ necessarily $\ell > \frac{2}{\lambda}$.

Choose now $\mathcal{Z} = \{0,1\}^n$, $\varepsilon = 2^{-\ell}$ and A_i's as in the Proposition. Thus $|A_i| = 2^{n-\ell}$, $N(n,\lambda) = 2^{-n}\ 2^{2^{n-\ell}} - 1$ and $|A_i \cap A_j| < \lambda 2^{n-\ell}$.

Consider now a discrete channel $(W^n)_{n=1}^\infty$, where the input alphabets $\mathcal{X}_t = \{1, 2, \dots, N(t,\lambda)\}$ are increasing, $\mathcal{X}^n = \prod_{t=1}^n \mathcal{X}_t$ are the input words of length n, $\mathcal{Y}^n = \{0,1\}^n$ are the output words and $W^n : \mathcal{X}^n \rightsquigarrow \mathcal{Y}^n$ is defined by

$$W^n(\cdot|i_1 i_2 \dots i_n) = W^n(\cdot|i_n)$$

and $W^n(\cdot|i)$ is the uniform distribution on A_i for $1 \leq i \leq N(n,\lambda)$.

By Proposition 5 and $3/\lambda > \ell > 2/\lambda$

$$N(n,\lambda) \geq 2^{-n} 2^{2^{n-3/\lambda}}$$

and

$$C_I \geq \underline{\lim}_{n\to\infty} \frac{1}{n} \log\log N(n,\lambda) \geq 1.$$

However, for transmission every decoding set is contained in some A_i and for error probability λ must have cardinality $(1-\lambda)|A_i| = (1-\lambda)2^{n-\ell}$.

Therefore $M(n,\lambda) \leq \frac{2^n}{(1-\lambda)2^{n-\ell}} \leq 2^{\ell+1}$, if $\lambda < 1/2$, and $\frac{1}{n}\log M(n,\lambda) \leq \frac{\ell+1}{n} \leq \frac{3/\lambda+1}{n} \to 0$ $(n \to \infty)$. The transmission capacity is 0. Consequently also $C_R = 0$.

Remarks:

22. The case of bounded input alphabets remains to be analysed. What are "natural" candidates for equality of C_I and C_R?

23. For infinite alphabets one should work out conditions for finiteness of the identification capacity.

19 Robustness, Common Randomness and Identification

It is understood now ([45], [42]) how the theory of AV–channels is *intimately* related to the concept of robust common randomness. A key tool is the balanced hypergraph coloring ([8]). We sketch now another direction concerning robustness and identification.

For more robust channel models, for instance in jamming situations, where the jammer knows the word to be send (c.f. AV–channels with maximal error criterion), the communicators are forced to use the maximal error concept (see [1]). In case of identification this makes the randomisation in the encoding (see [16]) superfluous. Now, for a DMC W it was mentioned in [16] that in the absence of randomisation the identification capacity, say $C_I^*(W)$, equals the logarithm of the number of different row–vectors in W. This is easy to show, however, a formidable problem arises if the DMC W is replaced by the AVC $\mathcal{W}$. In fact, for 0–1–matrices only in $\mathcal{W}$ we are — exactly as for transmission — led to the equivalent Shannon–zero–capacity problem. But for general $\mathcal{W}$ the identification problem is quite different from the transmission problem.

In so far there is a lower bound on $C_I^*(\mathcal{W})$, which implies for $\mathcal{W} = \left\{ \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix} \right), \left(\begin{smallmatrix} 1 & 0 \\ \delta & 1-\delta \end{smallmatrix} \right) \right\}, \delta \in (0,1)$ that $C_I^*(\mathcal{W}) = 1$, which is obviously tight. It exceeds the known capacity for transmission. The capacity for $\mathcal{W} = \left\{ \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix} \right), \left(\begin{smallmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{smallmatrix} \right) \right\}$ is unknown.

20 Probabilistic Algorithms: Identification as Concept of Solution

Finally we mention as the perhaps most promising direction the study of probabilistic algorithms with identification as *concept of solution.* (For example: for any i, is there a root of a polynomial in interval i or not?)

The algorithm should be fast and have small error probabilities. Every algorithmic problem can be thus considered. This goes far beyond information theory. Of course, like in general information transfer also here a more general set of questions can be considered. As usual in complexity theory one may try to classify problems.

What rich treasures do we have in the much wider areas of information transfer?!

References

[1] C.E. Shannon, "A mathematical theory of communication", Bell Syst. Techn. J. 27, 339–425, 623–656, 1948.

[2] C.E. Shannon, "Two–way communication channels", Proc. 4th Berkeley Symp. Math. Statist. and Prob., Univ. of California Press, Berkeley, Vol. 1, 611–644, 1961.

[3] R. Ahlswede, "Multi–way communication channels", Proc. 2nd Int. Symposium on Inf. Theory, Tsahkadsor Armenian SSR, 1971. Akadémiai Kiadó, Budapest, 23–52, 1973.

[4] T.M. Cover, "Broadcast channels", IEEE Trans. Inf. Theory, Vol. IT 18, 2–14, 1972.

[5] T.M. Cover, "An achievable rate region for the broadcast channel", IEEE Trans. Inf. Theory, Vol. IT 21, 399–401, 1975.

[6] R. Ahlswede, "Channel capacities for list codes", J. Appl. Prob. 10, 824–836, 1973.

[7] D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources", IEEE Trans. Inf. Theory, Vol. IT 19, 471–480, 1973.

[8] R. Ahlswede, "Coloring hypergraphs: A new approach to multi–user source coding", Part I, J. Comb., Inf. and System Sciences, Vol. 1, 76–115, 1979, Part II, Vol. 5, No. 3, 220–268, 1980.

[9] A.C. Yao, "Some complexity questions related to distributive computing", Proc. 11th Annual Symposium on Theory of Computing, Atlanta, 209–213, 1979.

[10] R. Ahlswede and I. Csiszár, "Hypothesis testing under communication constraints", IEEE Trans. Inf. Theory, Vol. IT–32, No. 4, 533–543, 1986.

[11] R. Ahlswede and G. Dueck, "Good codes can be produced by a few permutations", IEEE Trans. Inf. Theory, Vol. IT 28, No. 3, 430–443, 1982.

[12] R. Ahlswede and T.S. Han, "On source coding with side information via a multiple–access channel and related problems", IEEE Trans. Inf. Theory, Vol. IT 29, No. 3, 396–412, 1983.

[13] J. Ja'Ja', "Identification is easier than decoding", Preprint.

[14] I. Csiszár and J. Körner, "Information Theory: Coding Theorems for Discrete Memoryless Systems", Academic Press, N.Y. 1981.

[15] R. Ahlswede, "Eight problems in information theory" in "Open Problems in Communication and Computation", T.M. Cover and B. Gopinath Editors, Springer Verlag, 1987.

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

[17] R. Ahlswede and G. Dueck, “Identification in the presence of feedback — a discovery of new capacity formulas”, IEEE Trans. Inf. Theory, Vol. 35, 30–39, 1989.

[18] M. Lothaire, “Combinatorics on words”, Encycl. of Math. and its Applications, Vol. 17, 1982.

[19] E. Mittenecker and E. Raab, “Informationstheorie für Psychologen”, Verlag für Psychologie, Dr. C. Hofgrefe, Göttingen 1973.

[20] P. Stucki, “Advances in digital image processing”, Theory, Application, Implentation, The IBM Res. Symp. Series.

[21] J. Von Neumann, “The Computer and the Brain”, Yale University Press, 1958.

[22] R. Ahlswede and P. Gács, “Two contributions to information theory”, Colloquia Mathematica Societatis János Bolyai, 16. Topics in Information Theory, I. Csiszár and P. Elias Edit., Keszthely, Hungaria, 1975, 17–40.

[23] C. Cherry, “On human communication, A Review, a survey and a criticism”, MIT–Press 1957, 1966.

[24] R.G. Gallager, “A perspective on multi–access channels”, IEEE Trans. Inf. Theory IT–31, No. 2, March 1985.

[25] T. Berger, “Rate–distortion Theory”, Prentice–Hall, Inc. Englewood Cliffs, N.J., 1971.

[26] D. Coppersmith, Private Communication in 1987.

[27] R. Ahlswede and B. Verboven, “On identification via multi–way channels with feedback”, IEEE Trans. Inf. Theory, Vol. 37, No. 5, 1519–1526, 1991.

[28] T.S. Han and S. Verdú, “New results in the theory and application of identification via channels”, IEEE Trans. on Inform. Theory, Vol. IT–38, 14–25, 1992.

[29] T.S. Han and S. Verdú, “Approximation theory of output statistics”, IEEE Trans. Inform. Theory, Vol. 39, No. 3, 1993.

[30] J. Körner and K. Marton, “General broadcast channels with degraded message sets”, IEEE Trans. Inf. Theory, IT 23, 60–64, 1977.

[31] E.C. van der Meulen, “Random coding theorems for the general discrete memoryless broadcast channel”, IEEE Trans. Inf. Theory, IT 21, 180–190, 1975.

[32] M.S. Pinsker, “Capacity region of noiseless broadcast channels”, (in Russian), Problemi Peredachii Informatsii 14, No 2, 28–32, 1978.

[33] B. Verboven and E.C. van der Meulen, “Capacity bounds for identification via broadcast channels that are optimal for the deterministic broadcast channel”, IEEE Trans. Inf. Theory, Vol. 36, No 6, 1197–1205, 1990. IEEE Workshop on Inf. Theory, Salvador, Brazil, June 1992.

[34] A. Renýi, “On the foundations of information theory”, Rev. Inst. Internat. Stat. 33, 1–14, 1965.

[35] R. Ahlswede and Z. Zhang, “New directions in the theory of identification via channels”, SFB 343 “Diskrete Strukturen in der Mathematik”, Bielefeld, Preprint 94–010, 1994, IEEE Trans. Inform. Theory, Vol. 41, No. 4, 1040–1050, 1995.

[36] A.D. Wyner, “The wire–tape channel”, Bell Syst. Tech. J., Vol. 54, 1355–1387, 1975.

[37] R. Ahlswede and N. Cai, “Information and Control: The matching channel”, Preprint 95–035, SFB 343 “Diskrete Strukturen in der Mathematik”, Universität Bielefeld, IEEE Trans. Inf. Theory ...

[38] R. Ahlswede, Jian–ping Ye, and Z. Zhang, “Creating order in sequence spaces with simple machines”, Information and Computation, Vol. 89, No. 1, 47–94, 1990.

[39] R. Ahlswede, E. Yang, and Z. Zhang, “Identification via compressed data”, IEEE Trans. Inform. Theory, Vol. 43, No. 1, 48–70, 1997.

[40] J. Singh, “Great Ideas in Information Theory, Language and Cybernetics”, Dover Publication, Inc. New York, 1966.

[41] R. Ahlswede and N. Cai, “Arbitrarily varying multiple–access channels”, Part I: Ericson’s symmetrizability is adequate, Gubner’s conjecture is true, Preprint 96–068, SFB “Diskrete Strukturen in der Mathematik”, Universität Bielefeld; Part II: Correlated sender’s side information, correlated messages and ambiguous transmission, Preprint 97–006, SFB 343 “Diskrete Strukturen in der Mathematik”, Universität Bielefeld, to appear in IEEE Trans. on Inf. Theory.

[42] R. Ahlswede and N. Cai, “The AVC with noiseless feedback and maximal error probability: a capacity formula with a trichotomy”, Preprint 96–064, SFB 343 “Diskrete Strukturen in der Mathematik”, Universität Bielefeld, submitted to IEEE Trans. Inf. Theory.

[43] R. Ahlswede and C. Kleinewächter, “Identification versus common randomness”, in preparation.

[43] C.C. Campbell, “Definition of entropy by means of a coding problem”, Z. Wahrscheinlichkeitstheorie u. verw. Geb., 113–119, 1966.

[44] R. Ahlswede and B. Balkenhol, “Data compression for identification”, in preparation.

[45] R. Ahlswede and I. Csiszár, “Common randomness in information theory and cryptography, Part I: Secret sharing”, IEEE Trans. Information Theory, Vol. 39, No. 4, 1121–1132; Part II: CR capacity, Preprint 95–101, SFB 343 “Diskrete Strukturen in der Mathematik”, Universität Bielefeld, to appear in IEEE Trans. Inf. Theory.

[46] R. Ahlswede, “The capacity region of a channel with two senders and two receivers”, Ann. Probability, Vol. 2, No. 5, 805–814, 1974.

[47] T.S. Han, Forthcoming book.

[48] R. Ahlswede and V. Balakirsky, "Identification under random processes", Preprint 95–098, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, Problemy peredachii informatsii (special issue devoted to M.S. Pinsker), vol. 32, no. 1, 144–160, Jan.–March 1996.

[49] W.H. Kautz and R.C. Singleton, "Nonrandom binary superimposed codes", IEEE Trans. Inform. Theory 10, 363–377, 1964.

[50] A.G. Dyachkov and V.V. Rykov, "Bounds on the length of disjunctive codes", Problemy Peredachi Informatsii 18, No. 3, 7–13 (in Russian), 1982.

[51] P. Erdös, P. Frankl, and Z. Füredi, "Families of finite sets in which no set is covered by the union of r others", Israel J. Math. 51, 79–89, 1985.

[52] M. Ruszinko, "On the upper bound of the size of the r–cover–free families", J. Combinatorial Theory, Ser. A 66, 302–310, 1994.

[53] Z. Füredi, "A note on r–vover–free families", to appear in J. Combinatorial Theory, Ser. A.

[54] N.N. Kuzjurin, "On the difference between asymptotically good packings and coverings", European J. Combin. 16, 35–40, 1995.