# On Concepts of Performance Parameters for Channels

R. Ahlswede

**Abstract.** Among the mostly investigated parameters for noisy channels are code size, error probability in decoding, block length; rate, capacity, reliability function; delay, complexity of coding. There are several statements about connections between these quantities. They carry names like "coding theorem", "converse theorem" (weak, strong, ...), "direct theorem", "capacity theorem", "lower bound", "upper bound", etc. There are analogous notions for source coding.

This note has become necessary after the author noticed that Information Theory suffers from a lack of precision in terminology. Its purpose is to open a discussion about this situation with the goal to gain more clarity.

There is also some confusion concerning the scopes of analytical and combinatorial methods in probabilistic coding theory, particularly in the theory of identification. We present a covering (or approximation) lemma for hypergraphs, which especially makes strong converse proofs in this area transparent and dramatically simplifies them.

## 1 Channels

It is beyond our intention to consider questions of modelling, like what is a channel in reality, which parts of a communication situation constitute a channel etc. Shannon's mathematical description in terms of transmission probabilities is the basis for our discussion.

Also, in most parts of this note we speak about one–way channels, but there will be also comments on multi–way channels and compound channels.

Abstractly, let $\mathcal{I}$ be any set, whose elements are called input symbols and let $\emptyset$ be any set, whose elements are called output symbols.

An (abstract) channel $W : \mathcal{I} \to (\emptyset, \mathcal{E})$ is a set of probability distributions

$$W = \{W(\cdot|i) : i \in \mathcal{I}\} \tag{1.1}$$

on $(\emptyset, \mathcal{E})$.

So for every input symbol $i$ and every (measurable) $E \in \mathcal{E}$ of output symbols $W(E|i)$ specifies the probability that a symbol in $E$ will be received, if symbol $i$ has been sent.

The set $\mathcal{I}$ does not have to carry additional structure.

Of particular interest are channels with "time–structure", that means, symbols are words over an alphabet, say $\mathcal{X}$ for the inputs and $\mathcal{Y}$ for the outputs. Here $\mathcal{X}^n = \prod_{t=1}^{n} \mathcal{X}_t$ with $\mathcal{X}_t = \mathcal{X}$ for $t \in \mathbb{N}$ (the natural numbers) are the input

R. Ahlswede et al. (Eds.): Information Transfer and Combinatorics, LNCS 4123, pp. 639–663, 2006.

words of (block)–length $n$ and $\mathcal{Y}^n = \prod_{t=1}^{n} \mathcal{Y}_t$ with $\mathcal{Y}_t = \mathcal{Y}$ for $t \in \mathbb{N}$ are the output words of length $n$.

Moreover, again for the purpose of this discussion we can assume that a transmitted word of length $n$ leads to a received word of length $n$. So we can define a (constant block length) channel by the set of stochastic matrices

$$\mathcal{K} = \{W^n : \mathcal{X}^n \to \mathcal{Y}^n : n \in \mathbb{N}\}. \tag{1.2}$$

In most channels with time–structure there are (compatibility) relations between these matrices.

We don't have to enter these delicate issues. Instead, we present now three channel concepts, which serve as key examples in this note.

**DMC:** The most familiar channel is the discrete memoryless channel, defined by the transmission probabilities

$$W^n(y^n|x^n) = \prod_{t=1}^{n} W(y_t|x_t) \tag{1.3}$$

for $W : \mathcal{X} \to \mathcal{Y}$, $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, $y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$, and $n \in \mathbb{N}$.

**NDMC:** The *nonstationary* discrete memoryless channel is given by a sequence $(W_t)_{t=1}^{\infty}$ of stochastic matrices $W_t : \mathcal{X} \to \mathcal{Y}$ and the rule for the transmission of words

$$W^n(y^n|x^n) = \prod_{t=1}^{n} W_t(y_t, x_t). \tag{1.4}$$

Other names are "inhomogeneous channel", "non–constant" channel.

Especially, if $W_t = \begin{cases} W & \text{for } t \text{ even} \\ V & \text{for } t \text{ odd} \end{cases}$

one gets a "periodic" channel of period 2 or a "parallel" channel. (c.f. [32], [2])

**ADMC:** Suppose now that we have two channels $\mathcal{K}_1$ and $\mathcal{K}_2$ as defined in (1.2). Then following [3] we can associate with them an *averaged* channel

$$\mathcal{A} = \left\{ \left( \frac{1}{2} W_1^n + \frac{1}{2} W_2^n : \mathcal{X}^n \to \mathcal{Y}^n \right) : n \in \mathbb{N} \right\} \tag{1.5}$$

and when both constituents, $\mathcal{K}_1$ and $\mathcal{K}_2$ are DMC's (resp. NDMC's) we term it ADMC (resp. ANDMC).

It is a very simple channel with "strong memory", suitable for theoretical investigations. They are considered in [3] in much greater generality (any number of constituents, infinite alphabets) and have been renamed by Han and Verdu "mixed channels" in several papers (see [29]).

*We shall see below that channel parameters, which have been introduced for the DMC, where their meaning is without ambiguities, have been used for general time–structured channels for which sometimes their formal or operational meaning is not clear.*

NONSTATIONARITY and MEMORY, incorporated in our examples of channels, are tests for concepts measuring channel performance.

## 2 Three Unquestioned Concepts: The Two Most Basic, Code Size and Error Probability, Then Further Block Length

Starting with the abstract channel $W : \mathcal{I} \to (\emptyset, \mathcal{E})$ we define a *code*

$$\mathcal{C} = \big\{(u_i, D_i) : i \in I\big\} \text{ with } u_i \in \mathcal{I}, D_i \in \mathcal{E}$$

for $i \in I$ and pairwise disjoint $D_i$'s.

$$M = |\mathcal{C}| \text{ is the code size} \tag{2.1}$$

$$e(\mathcal{C}) = \max_{i \in I} W(D_i^c | u_i) \tag{2.2}$$

is the (maximal) probability of error and

$$\overline{e}(\mathcal{C}) = \frac{1}{M} \sum_{i=1}^{M} W(D_i^c | u_i) \tag{2.3}$$

is the average probability of error.

One can study now the functions

$$M(\lambda) = \max_{\mathcal{C}} \big\{|\mathcal{C}| : e(\mathcal{C}) \leq \lambda\big\} \text{ (resp. } \overline{M}(\lambda)) \tag{2.4}$$

and

$$\lambda(M) = \min_{\mathcal{C}} \big\{e(\mathcal{C}) : |\mathcal{C}| = M\big\} \text{ (resp. } \overline{\lambda}(M)), \tag{2.5}$$

that is, finiteness, growth, convexity properties etc.

It is convenient to say that $\mathcal{C}$ is an $(M, \lambda)$–code, if

$$|\mathcal{C}| \geq M \text{ and } e(\mathcal{C}) \leq \lambda. \tag{2.6}$$

Now we add time–structure, that means here, we go to the channel defined in (1.2). The parameter $n$ is called the *block length* or word length.

It is to be indicated in the previous definitions. So, if $u_i \in \mathcal{X}^n$ and $D_i \subset \mathcal{Y}^n$ then we speak about a code $\mathcal{C}(n)$ and definitions (2.4), (2.5), and (2.6) are to be modified accordingly:

$$M(n, \lambda) = \max_{\mathcal{C}(n)} \big\{|\mathcal{C}(n)| : e\big(\mathcal{C}(n)\big) \leq \lambda\big\} \tag{2.7}$$

$$\lambda(n, M) = \min_{\mathcal{C}(n)} \big\{e\big(\mathcal{C}(n)\big) : |\mathcal{C}(n)| = M\big\} \tag{2.8}$$

$$\mathcal{C}(n) \text{ is an } (M, n, \lambda)\text{–code, if } |\mathcal{C}(n)| \geq M, e\big(\mathcal{C}(n)\big) \leq \lambda. \tag{2.9}$$

**Remark 1:** One could study blocklength as function of $M$ and $\lambda$ in smooth cases, but this would be tedious for the general model $\mathcal{K}$, because monotonicity properties are lacking for $M(n, \lambda)$ and $\lambda(M, n)$.

We recall next Shannon's fundamental statement about the two most basic parameters.

## 3 Stochastic Inequalities: The Role of the Information Function

We consider a channel $W : \mathcal{X} \to \mathcal{Y}$ with finite alphabets. To an input distribution $P$, that is a PD on $\mathcal{X}$, we assign the output distribution $Q = PW$, that is a PD on $\mathcal{Y}$, and the joint distribution $\tilde{P}$ on $\mathcal{X} \times \mathcal{Y}$, where $\tilde{P}(x, y) = P(x)W(y|x)$.

Following Shannon [38] we associate with $(P, W)$ or $\tilde{P}$ the *information function (per letter)* $I : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where

$$I(x, y) = \begin{cases} \log \frac{\tilde{P}(x,y)}{P(x)Q(y)} \\ 0 \end{cases} \quad , \text{ if } \tilde{P}(x, y) = 0. \tag{3.1}$$

If $X$ is an (input) RV with values in $\mathcal{X}$ and distribution $P_X = P$ and if $Y$ is an (output) RV with values in $\mathcal{Y}$ and distribution $P_Y = Q$ such that the joint distribution $P_{XY}$ equals $\tilde{P}$, *then* $I(X, Y)$ *is a RV*. Its distribution function will be denoted by $F$, so

$$F(\alpha) = \Pr\big\{I(X, Y) \le \alpha\big\} = \tilde{P}\big(\big\{(x, y) : I(x, y) \le \alpha\big\}\big). \tag{3.2}$$

We call an $(M, \overline{\lambda})$–code $\big\{(u_i, D_i) : 1 \le i \le M\big\}$ *canonical*, if $P(u_i) = \frac{1}{M}$ for $i = 1, \ldots, M$ and the decoding sets are defined by maximum likelihood decoding, which results in a (minimal) average error probability $\overline{\lambda}$.

**Theorem.** Shannon [38]
*For a canonical* $(M, \overline{\lambda})$*–code and the corresponding information function there are the relations*

$$\frac{1}{2}F\left(\log \frac{M}{2}\right) \le \overline{\lambda} \le F\left(\log \frac{M}{2}\right). \tag{3.3}$$

**Remarks**

2. Shannon carries in his formulas a blocklength $n$, but this is nowhere used in the arguments. The bounds hold for abstract channels (without time structure). The same comment applies to his presentation of his random coding inequality: there exists a code of length $M$ and average probability of error

$$\overline{\lambda} \le F(\log M + \theta) + e^{-\theta}, \theta > 0.$$

3. Let us emphasize that all of Shannon's bounds involve the information function (per letter), which is highlighted also in Fano [24], where it is called mutual information. (One may argue which terminology should be used, but certainly we don't need the third "information spectrum" introduced more recently by Han!) In contrast, Fano's inequality is *not a stochastic inequality*. It works with the *average* (or expected) mutual information $I(X \wedge Y)$ (also written as $I(X; Y)$), which is a constant. Something has been given away.

## 4 Derived Parameters of Performance: Rates for Code Sizes, Rates for Error Probabilities, Capacity, Reliability

The concept of rate involves a renormalisation in order to put quantities into a more convenient scale, some times per unit. Exponentially growing functions are renormalized by using the logarithmic function. In Information Theory the prime example is $M(n, \lambda)$ (see 2.7). Generally speaking, with any function $f : \mathbb{N} \to \mathbb{R}_+$ (or, equivalently, any sequence $\big(f(1), f(2), f(3), \dots\big)$ of non–negative numbers) we can associate a rate function $\mathrm{rate}(f)$, where

$$\mathrm{rate}\big(f(n)\big) = \frac{1}{n} \log f(n). \tag{4.1}$$

We also speak of the *rate at* $n$, when we mean

$$\mathrm{rate}_n(f) \triangleq \mathrm{rate}\big(f(n)\big) = \frac{1}{n} \log f(n). \tag{4.2}$$

This catches statements like "an increase of rate" or "rate changes".

In Information Theory $f$ is related to the channel $\mathcal{K}$ or more specifically $f(n)$ depends on $W^n$. For example choose $f(n) = M(n, \lambda)$ for $n \in \mathbb{N}$, $\lambda$ constant. Then $\mathrm{rate}(f)$ is a *rate function* for certain code sizes.

Now comes a *second step*: for many *stationary* systems like stationary channels (c.f. DMC) $f$ behaves very regular and instead of dealing with a whole rate function one just wants to associate a *number* with it.

We state for the three channels introduced in Section 1 the results – not necessarily the strongest known – relevant for our discussion.

**DMC:** There is a constant $C = C(W)$ (actually known to equal $\max\limits_P I(W|P)$) such that

(a) for every $\lambda \in (0, 1)$ and $\delta > 0$ there exists an $n_0 = n_0(\lambda, \delta)$ such that *for all* $n \geq n_0$ there exist
$$(n, e^{(C-\delta)n}, \lambda)\text{–codes},$$

(b) for every $\lambda \in (0, 1)$ and $\delta > 0$ there exists an $n_0 = n_0(\lambda, \delta)$ such that *for all* $n \geq n_0$ there does *not* exist an
$$(n, e^{(C+\delta)n}, \lambda)\text{–code}.$$

**ADMC:** There is a constant $C$ (actually known to equal $\max\limits_P \min\limits_{i=1,2} I(W_i|P)$ [3]) such that

(a) holds

(c) for every $\delta > 0$ there *exists* a $\lambda \in (0, 1)$ and an $n_0 = n_0(\lambda, \delta)$ such that *for all* $n \geq n_0$ there does *not* exist an
$$(n, e^{(C+\delta)n}, \lambda)\text{–code}.$$

**NDMC:** There is a sequence of numbers $\big(C(n)\big)_{n=1}^{\infty}$ (which actually can be chosen as $C(n) = \frac{1}{n}\sum_{t=1}^{n} \max_{P} I(W_t|P)$ [2]) such that

(a′) for every $\lambda \in (0,1)$ and $\delta > 0$ there exists an $n_0 = n_0(\lambda, \delta)$ such that for all $n \geq n_0$ there exist

$$(n, e^{(C(n)-\delta)n}, \lambda)\text{–codes.}$$

(b′) for every $\lambda \in (0,1)$ and $\delta > 0$ there exists an $n_0 = n_0(\lambda, \delta)$ such that for all $n \geq n_0$ there does *not* exist an

$$(n, e^{(C(n)+\delta)n}, \lambda)\text{–code.}$$

(This is still true for infinite output alphabets, for infinite input alphabets in general not. There the analogue of (c), say (c′) is often still true, but also not always.)

Notice that with every sequence $\big(C(n)\big)_{n=1}^{\infty}$ satisfying (a′) and (b′) or (a′) and (c′) also every sequence $\big(C(n)+o(1)\big)_{n=1}^{\infty}$ does. In this sense the sequence is not unique, whereas earlier the constant $C$ is.

The *pair* of statements ((a), (b)) has been called by Wolfowitz *Coding theorem with strong converse* and the number $C$ has been called the *strong capacity* in [2]. For the ADMC there is no $C$ satisfying (a) *and* (b), so this channel *does not have* a strong capacity.

The pair of statements ((a), (c)) have been called by Wolfowitz coding theorem with *weak converse* and the number $C$ has been called in [2] the *weak capacity*. So the ADMC does have a weak capacity.

(For completeness we refer to two standard textbooks. On page 9 of Gallager [27] one reads "The converse to the coding theorem is stated and proved in varying degrees of generality in chapter 4, 7, and 8. In imprecise terms, it states that if the entropy of a discrete source, in bits per second, is greater than $C$, then independent of the encoding and decoding used in transmitting the source output at the destination cannot be less than some positive number which depends on the source and on $C$. Also, as shown in chapter 9, if $R$ is the minimum number of binary digits per second required to reproduce a source within a given level of average distortion, and if $R > C$, then, independent of the encoding and decoding, the source output cannot be transmitted over the channel and reproduced within that given average level of distortion."

In spite of its pleasant preciseness in most cases, there seems to be no definition of the weak converse in the book by Csiszár and Körner [22].)

**Now the NDMC has in general no strong and no weak capacity (see our example in Section 7)**

However, if we replace the concept of capacity by that of a capacity function $\big(C(n)\big)_{n=1}^{\infty}$ then the pair ((a′), (b′)) (resp. ((a′), (c′)) may be called coding theorem with strong (resp. weak) converse and accordingly one can speak about *strong* (*resp. weak*) *capacity functions*, defined modulo $o(1)$.

These concepts have been used or at least accepted – except for the author – also by Wolfowitz, Kemperman, Augustin and also Dobrushin [23], Pinsker [35]. The concept of information stability (Gelfand/Yaglom; Pinsker) defined for *sequences of numbers* and *not* – like some authors do nowadays – for a *constant only*, is in full agreement at least with the ((a), (c)) or ((a′), (c′)) concepts. Equivalent formulations are

(a′) $\inf\limits_{\lambda>0} \underline{\lim\limits_{n\to\infty}} \left(\frac{1}{n}\log M(n,\lambda) - C(n)\right) \geq 0$

(b′) for all $\lambda \in (0,1)$ $\overline{\lim\limits_{n\to\infty}} \left(\frac{1}{n}\log M(n,\lambda) - C(n)\right) \leq 0$

(c′) $\inf\limits_{\lambda>0} \overline{\lim\limits_{n\to\infty}} \left(\frac{1}{n}\log M(n,\lambda) - C(n)\right) \leq 0.$

(For a constant $C$ this gives (a), (b), (c).)

**Remarks**

4. A standard way of expressing (c) is: for rates above capacity the error probability is bounded away from 0 for *all large n*. ([25], called "*partial converse*" on page 44.)
5. There are cases (c.f. [3]), where the uniformity in $\lambda$ valid in (b) or (b′) holds only for $\lambda \in (0, \lambda_1)$ with an absolute constant $\lambda_1$ – a "medium" strong converse. It also occurs in "second order" estimates of [31] with $\lambda_1 = \frac{1}{2}$.
6. There are cases where (c) (or (c′)) don't hold for constant $\lambda$'s but for $\lambda = \lambda(n)$ going to 0 sufficiently fast, in one case [17] like $\frac{1}{n}$ and in another like $\frac{1}{n^4}$ [19]. In both cases $\lambda(n)$ decreases reciprocal to a polynomial and it makes sense to speak of polynomial–weak converses. The soft–converse of [12] is for $\lambda(n) = e^{o(n)}$. Any decline condition on $\lambda_n$ could be considered.
7. For our key example in Section 7 ((a′), (c′) holds, but not ((a), (c)). It can be shown that for the constant $C = 0$ and any $\delta > 0$ there is a $\lambda(\delta) > 0$ such that $(n, e^{(C+\delta)n})$–codes have error probability exceeding $\lambda(\delta)$ for *infinitely many n*.

   **By Remark 1 this is weaker than (c) and equivalent to**

$$\inf_{\lambda>0} \underline{\lim_{n\to\infty}} \frac{1}{n}\log M(n,\lambda) = C.$$

Now comes a seemingly small twist. Why bother about "weak capacity", "strong capacity" etc. and their existence – every channel should have a capacity.

**Definition:** $\underline{C}$ is called the (pessimistic) capacity of a channel $\mathcal{K}$, if it is the supremum over all numbers $C$ for which (a) holds. Since $C = 0$ satisfies (a), the number $\underline{C} = \underline{C}(\mathcal{K})$ exists. Notice that there are no requirements concerning (b) or (c) here.

*To every general $\mathcal{K}$ a constant performance parameter has been assigned* ! What does it do for us?

First of all the name "pessimistic" refers to the fact that another number $\overline{C} = \overline{C}(\mathcal{K})$ can be introduced, which is at least a large as $\underline{C}$.

**Definition:** $\overline{C}$ is called the (optimistic) capacity of a channel $\mathcal{K}$, if it is the supremum over all numbers $C$ for which in (a) the condition "for all $n \geq n_0(\lambda, \delta)$" is replaced by "for infinitely many $n$" or equivalently

$$\overline{C} = \inf_{\lambda > 0} \overline{\lim_{n \to \infty}} \frac{1}{n} \log M(n, \lambda).$$

Here it is measured whether for every $\lambda \;\; R < \overline{C}$ this "rate" is occasionally, but infinitely often achievable.

(Let us briefly mention that "the reliability function" $E(R)$ is commonly defined through the values

$$\underline{E}(R) = - \underline{\lim_{n \to \infty}} \frac{1}{n} \log \lambda(e^{Rn}, n)$$

$$\overline{E}(R) = - \overline{\lim_{n \to \infty}} \frac{1}{n} \log \lambda(e^{Rn}, n)$$

if they coincide. Again further differentiation could be gained by considering the sequence

$$E_n(R_n) = -\frac{1}{n} \log \lambda(e^{R_n n}, n), \;\; n \in \mathbb{N},$$

for sequences of rates $(R_n)_{n=1}^{\infty}$. But that shall not be pursuit here.)

In the light of old work [2] we were shocked when we learnt that these two definitions were given in [22] and that the pessimistic capacity was used throughout that book. Since the restriction there is to the DMC–situation it makes actually no difference. However, several of our Theorems had just been defined away. Recently we were even more surprised when we learned that these definitions were not new at all and have indeed been standard and deeply rooted in the community of information theorists (the pessimistic capacity $\underline{C}$ is used in [24], [42], [21] and the optimistic capacity $\overline{C}$ is used in [22] on page 223 and in [33]).

Fano [24] uses $\underline{C}$, but he at the same time emphasizes throughout the book that he deals with "constant channels".

After quick comments about the optimistic capacity concept in the next section we report on another surprise concerning $\underline{C}$.

## 5 A Misleading Orientation at the DMC: The Optimistic Rate Concept Seems Absurd

Apparently for the DMC the optimistic as well as the pessimistic capacities, $\overline{C}$ and $\underline{C}$, equal $C(W)$. For multi–way channels and compound channels $\{W(\cdot|\cdot, s) : s \in \mathcal{S}\}$ the optimistic view suggests a dream world.

A. Recently Cover explained that under this view for the broadcast channel $(W : \mathcal{X} \to \mathcal{Y}, V : \mathcal{X} \to \mathcal{Z})$ the rate pair $(R_{\mathcal{Y}}, R_{\mathcal{Z}}) = \big(C(W), C(V)\big)$ is in the capacity region, which in fact equals $\big\{(R_{\mathcal{Y}}, R_{\mathcal{Z}}) : 0 \leq R_{\mathcal{Y}} \leq C(W), 0 \leq R_{\mathcal{Z}} \leq C(W)\big\}$.

Just assign periodically time intervals of lengths $m_1, n_1, m_2, n_2, m_3, n_3, \ldots$ to the DMC's $W$ and $V$ for transmission. Just choose every interval very long in comparison to the sum of the lengths of its predecessors. Thus again and again every channel comes in its rate close, and finally arbitrary close, to its capacity. The same argument applies to the MAC, TWC etc. – so in any situation where the communicators have a choice of the channels for different time intervals.

B. The reader may quickly convince himself that $\overline{C} = \min\limits_{s \in \mathcal{S}} C\big(W(\cdot|\cdot, s)\big) \geq \max\limits_{P} \min\limits_{s} I\big(W(\cdot|\cdot, s)|P\big)$ for the compound channel. For the sake of the argument choose $\mathcal{S} = \{1, 2\}$. The sender not knowing the individual channel transmits for channel $W(\cdot|\cdot, 1)$ on the $m$–intervals and for channel $W(\cdot|\cdot, 2)$ on the $n$–intervals. The receiver *can test* the channel and knows in which intervals to decode!

C. As a curious Gedankenexperiment: Is there anything one can do in this context for the AVC?

For the semicontinuous compound channel, $|\mathcal{S}| = \infty$, the ordinary weak capacity (((a),(c)) hold) is unknown. We guess that optimism does not help here, because it does seem to help if there are infinitely many proper cases.

The big issue in all problems here is of course delay. It ought to be incorporated (Space–time coding).

## 6 A "Paradox" for Product of Channels

Let us be given $s$ channels $(W_j^n)_{n=1}^{\infty}, 1 \leq j \leq s$. Here $W_j^n : \mathcal{X}_j^n \to \mathcal{Y}_j^n$, $1 \leq j \leq s$. The product of these channels $(W^{*n})_{n=1}^{\infty}$ is defined by

$$W^{*n} = \prod_{j=1}^{s} W_j^n : \prod_{j=1}^{s} \mathcal{X}_j^n \to \prod_{j=1}^{s} \mathcal{Y}_j^n.$$

A paper by Wyner [42] is very instructive for our discussion. We quote therefore literally the beginning of the paper (page 423) and also its Theorem with a sketch of the proof (page 425), because it is perhaps instructive for the reader to see how delicate things are even for leading experts in the field.

"In this paper we shall consider the product or parallel combination of channels, and show that (1) the *capacity of the product channel is the sum of the capacities of the component channels*, and (2) the "strong converse" holds for the product channel if it holds for each of the component channels. The result is valid for any class of channels (with or without memory, continuous or discrete) provided that the capacities exist. "Capacity" is defined here *as the supremum of those rates for which arbitrarily high reliability is achievable with block coding for sufficiently long delay.*

Let us remark here that there are two ways in which "channel capacity" is commonly defined. The first definition takes the channel capacity to be the supremum of the "information" processed by the channel, where "information" is the difference of the input "uncertainty" and the "equivocation" at the output. *The second definition, which is the one we use here, takes the channel capacity to be the maximum "error free rate"*. For certain classes of channels (e.g., memoryless channels, and finite state indecomposable channels) it has been established that these two definitions are equivalent. In fact, this equivalence is the essence of the Fundamental Theorem of Information Theory. For such channels, (1) above follows directly. The second definition, however, is applicable to a broader class of channels than the first. One very important such class are time–continuous channels."

**Theorem**

(1) *Let $C^*$ be the capacity of the product of $s$ channels with capacities $C_1, C_2, \ldots, C_s$ respectively. Then*

$$C^* = \sum_{j=1}^{s} C_j. \qquad ((6.1))$$

(2) *If the strong converse holds for each of these $s$ channels, then it holds for the product channel.*

The proof of (1) is divided into two parts. In the first (the "direct half") we will show that any $R < \sum_{j=1}^{s} C_j$ is a permissible rate. This will establish that $C^* \geqq \sum_{j=1}^{s} C_j$. In the second ("weak converse") we will show that no $R > \sum_{j=1}^{s} C_j$ is a permissible rate, establishing that $C^* \leqq \sum_{j=1}^{s} C_j$. The proof of (2) parallels that of the weak converse.

It will suffice to prove the theorem for the product of two channels ($s = 2$), the result for arbitrary $s$ following immediately by induction."

Let's first remark that $C^* \geq \sum_{j=1}^{s} C_j$ for the pessimistic capacities (apparently used here) follows immediately from the fact that by taking products of codes the errors at most behave additive. By proving the reverse inequality the weak converse, statement (c) in Section 4 is *tacitly assumed* for the component channels and from there on everything is okay. The point is that this assumption does not appear as a hypothesis in the Theorem! Indeed our key example of Section 7 shows that (6.1) is in general not true. The two factor channels used in the example don't have a weak converse (or weak capacity for that matter).

The reader is reminded that having proved a weak converse for the number $\underline{C}$, the pessimistic capacity, is equivalent to having shown that the weak capacity exists.

## 7 The Pessimistic Capacity Definition: An Information Theoretic Perpetuum Mobile

Consider the two matrices $V^1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $V^0 = \left(\begin{smallmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{smallmatrix}\right)$. We know that $C(V^1) = 1$ and $C(V^0) = 0$.

Consider a NDMC $\mathcal{K}$ with $W_t \in \{V^0, V^1\}$ for $t \in \mathbb{N}$ and a NDMC $\mathcal{K}^*$ with $t$–th matrix $W_t^*$ also from $\{V^0, V^1\}$ but *different* from $W_t$. Further consider the product channel $(\mathcal{K}, \mathcal{K}^*)$ specified by $W_1 W_1^* W_2 W_2^*$ – again a NDMC.

With the choice $(m_1, n_1, m_2, n_2, \dots)$, where for instance $n_i \geq 2^{m_i}$, $m_{i+1} \geq 2^{n_i}$ we define channel $\mathcal{K}$ completely by requiring that $W_t = V^1$ in the $m_i$–length intervals and $W_t = V^0$ in the $n_i$–length intervals. By their growth properties we have for the pessimistic capacities $\underline{C}(\mathcal{K}) = \underline{C}(\mathcal{K}^*) = 0$. However, apparently $\underline{C}(\mathcal{K}, \mathcal{K}^*) = 1$.

## 8 A Way Out of the Dilemma: Capacity Functions

If $M(n, \lambda)$ fluctuates very strongly in $n$ and therefore also $\mathrm{rate}_n(M)$, then it does not make much sense to describe its growth by one number $\underline{C}$. At least one has to be aware of the very limited value of theorems involving that number.

For the key example in Section 7 $\underline{C}(\mathcal{K}) = \underline{C}(\mathcal{K}^*) = 0$ and on the other hand $\overline{C}(\mathcal{K}) = \overline{C}(\mathcal{K}^*) = 1$. In contrast we can choose the sequence $(c_n)_{n=1}^{\infty} = \left(\frac{1}{n} \sum_{t=1}^{n} C(W_t)\right)_{n=1}^{\infty}$ for channel $\mathcal{K}$ and $(c_n^*)_{n=1}^{\infty} = \left(\frac{1}{n} \sum_{t=1}^{n} C(W_t^*)\right)_{n=1}^{\infty}$ for channel $\mathcal{K}^*$, who are always *between* 0 and 1.

They are (even strong) capacity functions and for the product channel $\mathcal{K} \times \mathcal{K}^*$ we have the capacity function $(c_n + c_n^*)_{n=1}^{\infty}$, which equals identically 1, what it should be. Moreover thus also in general the "perpetuum mobile of information" disappears. We have been able to prove the

**Theorem.** *For two channels $\mathcal{K}_1$ and $\mathcal{K}_2$*

(i) *with weak capacity functions their product has the sum of those functions as weak capacity function*
(ii) *with strong capacity functions their product has the sum of those functions as strong capacity function.*

We hope that we have made clear that capacity functions in conjunction with converse proofs carry in general more information – perhaps not over, but *about channels* – than optimistic or pessimistic capacities. This applies even for channels without a weak capacity function because they can be made this way at least as large $\underline{C}$ and still satisfy (a).

Our conclusion is, that

1. when speaking about capacity formulas in non standard situations one must clearly state which definition is being used.
2. there is no "true" definition nor can definitions be justified by authority.

3. presently weak capacity functions have most arguments in their favour, also in comparison to strong capacity functions, because of their wide validity and the primary interest in direct theorems. To call channels without a strong capacity "channels without capacity" ([41]) is no more reasonable than to name an optimistic or a pessimistic capacity "the capacity".
4. we must try to help enlightening the structure of channels. For that purpose for instance $\underline{C}$ can be a useful bound on the weak capacity function, because it may be computable whereas the function isn't.
5. Similar comments are in order for other quantities in Information Theory, rates for data compression, reliability functions, complexity measures.

## 9 Some Concepts of Performance from Channels with Phases

In this Section we explore other capacity concepts involving the phase of the channel, which for stationary systems is not relevant, but becomes an issue otherwise. Again the NDMC $(W_t)_{t=1}^{\infty}$ serves as a genuine example. In a phase change by $m$ we are dealing with $(W_{t+m})_{t=1}^{\infty}$. "Capacity" results for the class of channels $\left\{(W_{t+m})_{t=1}^{\infty} : 0 \leq m < \infty\right\}$ in the spirit of a compound channel, that is, for codes which are good simultaneously for all $m$ are generally unknown. The AVC can be produced as a special case and even more so the zero–error capacity problem.

An exception is for instance the case where $(W_t)_{t=1}^{\infty}$ is almost periodic in the sense of Harald Bohr. Because these functions have a mean also $\big(C(W_t)\big)_{t=1}^{\infty}$ has a mean and it has been shown that there is a strong capacity [2].

Now we greatly simplify the situation and look only at $(W_t)_{t=1}^{\infty}$ where

$$W_t \in \left\{ \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{smallmatrix}\right) \right\}$$

and thus $C(W_t) \in \{0, 1\}$. Moreover, we leave error probabilities aside and look only at $0-1$–sequences $(C_1, C_2, C_3, \dots)$ and the associated $C(n) = \frac{1}{n} \sum\limits_{t=1}^{n} C_t \in [0, 1]$.

So we just play with $0-1$–sequences $(a_n)_{n=1}^{\infty}$ and associated Cesaro–means $A_n = \frac{1}{n} \sum\limits_{t=1}^{n} a_t$ and $A_{m+1,m+n} = \frac{1}{n} \sum\limits_{t=m+1}^{m+n} a_t$.

First of all there are the familiar

$$\underline{A} = \underline{\lim_{n\to\infty}} A_n \text{ (the pessimistic mean)} \tag{9.1}$$

$$\overline{A} = \overline{\lim_{n\to\infty}} A_n \text{ (the optimistic mean).} \tag{9.2}$$

We introduce now a new concept

$$\underset{=}{A} = \underline{\lim_{n\to\infty}} \inf_{m\geq 0} A_{m+1,m+n} \text{ (the pessimistic phase independent mean).} \tag{9.3}$$

The "inf" reflects that the system could be in any phase (*known to but not controlled by the communicators*). Next we assume that the communicators can *choose* the phase $m$ for an intended $n$ and define

$$\overline{\underline{A}} = \overline{\lim_{n\to\infty}} \sup_{m\geq 0} A_{m+1,m+n} \text{ (super optimistic mean).} \tag{9.4}$$

We shall show first

**Lemma**

$$\overline{\underline{\lim_{n\to\infty}}} \inf_{m\geq 0} A_{m+1,m+n} = \underline{\underline{A}} \tag{9.5}$$

$$\underline{\overline{\lim_{n\to\infty}}} \sup_{m\geq 0} A_{m+1,m+n} = \overline{\underline{A}} \tag{9.6}$$

**Proof:** We prove only (9.5), the proof for (9.6) being "symmetrically" the same. We have to show that

$$\underline{\underline{A}} = \underline{\overline{\lim_{n\to\infty}}} \inf_{m\geq 0} A_{m+1,m+n} \geq \overline{\lim_{n\to\infty}} \inf_{m\geq 0} A_{m+1,m+n}. \tag{9.7}$$

For every $n$ let $m(n)$ give minimal $A_{m+1,m+n}$. The number exists because these means take at most $n+1$ different values. Let $n^*$ be such that $A_{m(n^*)+1,m(n^*)+n^*}$ is within $\varepsilon$ of $\underline{\underline{A}}$ and choose a much bigger $N^*$ for which $A_{m(N^*)+1,m(N^*)+N^*}$ is within $\varepsilon$ of the expression at the right side of (9.7) and $N^* \geq \frac{1}{\varepsilon} n^*$ holds.

Choose $r$ such that $rn^* + 1 \leq N^* \leq (r+1)n^*$ and write

$$\begin{aligned} N^* A_{m(N^*)+1,m(N^*)+N^*} &= \sum_{s=0}^{r-1} \sum_{t=m(N^*)+sn^*+1}^{m(N^*)+(s+1)n^*} a_t + \sum_{t=m(N^*)+rn^*+1}^{m(N^*)+N^*} a_t \\ &\geq r \cdot n^* A_{m(n^*)+n^*} \geq r \cdot n^* (\underline{\underline{A}} - \varepsilon) \\ &\geq (N^* - n^*)(\underline{\underline{A}} - \varepsilon) \geq N^*(1-\varepsilon)(\underline{\underline{A}} - \varepsilon). \end{aligned}$$

Finally, by changing the order of operations we get four more definitions, however, they give nothing new. In fact,

$$\inf_m \underline{\lim_{n\to\infty}} A_{m+1,m+n} = \sup_m \underline{\lim_{n\to\infty}} A_{m+1,m+n} = \underline{A} \tag{9.8}$$

$$\inf_m \overline{\lim_{n\to\infty}} A_{m+1,m+n} = \sup_m \overline{\lim_{n\to\infty}} A_{m+1,m+n} = \overline{A}, \tag{9.9}$$

because for an $m_0$ close to an optimal phase the first $m_0$ positions don't affect the asymptotic behaviour.

The list of quantities considered is not intended to be complete in any sense, but serves our illustration.

We look now at $\underline{\underline{A}} \leq \underline{A} \leq \overline{A} \leq \overline{\overline{A}}$ in four examples to see what constellations of values can occur.

We describe a $0-1$–sequence $(a_n)_{n=1}^{\infty}$ by the lengths of its alternating strings of 1's and 0's: $(k_1, \ell_1, k_2, \ell_2, k_3, \dots)$

**Example 1:** $k_t = k$, $\ell_t = \ell$ for $t = 1, 2, \ldots$; a periodic case:

$$\underline{\underline{A}} = \underline{A} = \overline{A} = \overline{\overline{A}} = \frac{k}{k+\ell}.$$

**Example 2:** $k_t = \ell_t = t$ for $t = 1, 2, \ldots$. Use $\sum_{t=1}^{n} k_t = \sum_{t=1}^{n} \ell_t = \frac{n(n+1)}{2}$ and verify

$$0 = \underline{\underline{A}} < \frac{1}{2} = \underline{A} = \overline{A} < 1 = \overline{\overline{A}}.$$

**Example 3:** $k_t = \sum_{s=1}^{t-1} k_s$, $\ell_t = \sum_{s=1}^{t-1} \ell_s$ for $t = 1, 2, \ldots$

$$0 = \underline{\underline{A}} < \frac{1}{2} = \underline{A} < \frac{2}{3} = \overline{A} < 1 = \overline{\overline{A}}.$$

Here all four values are different.

**Example 4:** $k_t = \sum_{s=1}^{t-1} k_s$, $\ell_t = t$ for $t = 2, 3, \ldots, k_1 = 1$

$$0 = \underline{\underline{A}} < 1 = \underline{A} = \overline{A} = \overline{\overline{A}}.$$

All four quantities say something about $(A_n)_{n=1}^{\infty}$, they all say less than the *full record*, the sequence itself (corresponding to our capacity function).

## 10 Some Comments on a Formula for the Pessimistic Capacity

A noticeable observation of Verdu and Han [39] is that $\underline{C}$ can be expressed for every channel $\mathcal{K}$ in terms of a stochastic limit (per letter) mutual information.

The renewed interest in such questions originated with the Theory of Identification, where converse proofs for the DMC required that output distributions of a channel, generated by an arbitrary input distribution (randomized encoding for a message), be "approximately" generated by input distributions of controllable sizes of the carriers. Already in [12] it was shown that essentially sizes of $\sim e^{Cn}$ would do and then in [30], [31] the bound was improved (strong converse) by a natural random selection approach. They termed the name "resolvability" of a channel for this size problem.

The approximation problem (like the rate distortion problem) is a "covering problem" as opposed to a "packing problem" of channel coding, but often these problems are very close to each other, actually ratewise identical for standard channels like the DMC. To establish the strong second order identification capacity for more general channels required in the approach of [30] that resolvability must equal capacity and for that the strong converse for $\mathcal{K}$ was needed.

This led them to study the ADMC [3], which according to Han [28] plaid a key role in the further development. Jacobs has first shown that there are channels with a weak converse, but without a strong converse. In his example the abstract reasoning did not give a channel capacity formula. This is reported in [32] and mentioned in [3], from where the following facts should be kept in mind.

1. The ADMC has no strong converse but a weak converse (see Section 4 for precise terminology).
2. The term weak capacity was introduced.
3. The weak capacity (and also the $\lambda$–capacity were determined for the ADMC by linking it to the familiar max min –formula for the compound channel in terms of (per letter)–mutual information.
4. It was shown that $\lim\limits_{n\to\infty} \frac{1}{n} \max_{X^n} I(X^n \wedge Y^n)$ does not describe the weak capacity in general. Compare this with Wyner's first capacity definition in Section 6.
5. It was shown that Fano's inequality, involving only the *average* mutual information $I(X^n \wedge Y^n)$, fails to give the weak converse for the ADMC.

The observation of [39] is again natural, one should use the information function of the ADMC directly rather than the max min –formula. They defined for general $\mathcal{K}$ the *sequence* of pairs

$$(\mathbf{X}, \mathbf{Y}) = (X^n, Y^n)_{n=1}^{\infty} \tag{10.1}$$

and

$$\underline{I}(\mathbf{X} \wedge \mathbf{Y}) = \sup \left\{ \alpha : \underline{\lim_{n\to\infty}} \Pr \left\{ (x^n, y^n) : \frac{1}{n} I(x^n, y^n) \leq \alpha \right\} = 0 \right\}. \tag{10.2}$$

Their general formula asserts

$$\underline{C} = \sup_{\mathbf{X}} \underline{I}(\mathbf{X} \wedge \mathbf{Y}). \tag{10.3}$$

The reader should be aware that

$\alpha$.) The stochastic inequalities used for the derivation (10.3) are both (in particular also Theorem 4 of [39]) not new.

$\beta$.) Finally, there is a very important point. In order to show that a certain quantity $K$ (for instance $\sup\limits_{\mathbf{X}} \underline{I}(\mathbf{X} \wedge \mathbf{Y})$) equals $\underline{C}$ one has to show $K \geq \underline{C}$ and then (by definition of $\underline{C}$) that $K + \delta$, any $\delta > 0$, is not a rate achievable for arbitrary small error probabilities or equivalently, that $\inf\limits_{\lambda} \underline{\lim\limits_{n\to\infty}} \log M(n, \lambda) < K + \delta$. For this one does *not need* the *weak* converse

(b) $\inf\limits_{\lambda} \overline{\lim\limits_{n\to\infty}} \log M(n, \lambda) \leq K$, but only

$$\inf_{\lambda} \underline{\lim_{n\to\infty}} \log M(n, \lambda) \leq K \tag{10.4}$$

(see also Section 4) The statement may be termed the "weak–weak converse" or the "<u>weak</u>–converse" or "occasional–converse" or whatever. Keep

in mind that the fact that the weak converse does not hold for the factors led to the "information theoretic perpetuum mobile". The remark on page 1153 "Wolfowitz ... referred to the conventional capacity of Definition 1 (which is always defined) as *weak capacity*" is not only wrong, because Wolfowitz never used the term "weak capacity", it is – as we have explained – very misleading. After we have commented on the drawbacks of the pessimistic capacity, especially also for channel NDMC, we want to say that on the other hand the formula $\sup_{\mathbf{X}} \underline{I}(\mathbf{X} \wedge \mathbf{Y})$ and also its dual $\sup_{\mathbf{X}} \overline{I}(\mathbf{X} \wedge \mathbf{Y})$ are helpful in characterizing or bounding quantities of interest not only in their original context, Theory of Identification. Han has written a book [29] in which he introduces these quantities and their analogues into all major areas of Information Theory.

## 11 Pessimistic Capacity Functions

We think that the following concept suggests itself as one result of the discussion.

**Definition:** A sequence $(C_n)_{n=1}^{\infty}$ of non–negative numbers is *a* capacity sequence of $\mathcal{K}$, if

$$\inf_{\lambda>0} \underline{\lim}_{n\to\infty} \left( \frac{1}{n} \log M(n, \lambda) - C_n \right) = 0.$$

The sequence $(\underline{C}, \underline{C}, \underline{C}, \dots)$ is a capacity sequence, so by definition there are always capacity sequences.

Replacing $\alpha$ by $\alpha_n$ in (10.2) one can characterize capacity sequences in term of sequences defined in terms of (per letter) information functions. Every channel $\mathcal{K}$ has a class of capacity sequences $\mathcal{C}(\mathcal{K})$.

It can be studied. In addition to the constant function one may look for instance at the class of functions of period $m$, say $\mathcal{C}(\mathcal{K}, m) \subset \mathcal{C}(\mathcal{K})$. More generally complexity measures $\mu$ for the sequences may be used and accordingly one gets say $\mathcal{C}(\mathcal{K}, \mu \leq \rho)$, a space of capacity functions of $\mu$–complexity less than $\rho$.

This seems to be a big machinery, but channels $\mathcal{K}$ with no connections between $W^n$ and $W^{n\prime}$ required in general constitute a *wild* class of channels. The capacity sequence space $\mathcal{C}(\mathcal{K})$ characterizes a channel in time like a capacity region for multi–way channels characterizes the possibilities for the communicators.

Its now not hard to show that for the product channel $\mathcal{K}_1 \times \mathcal{K}_2$ for any $f \in \mathcal{C}(\mathcal{K}_1 \times \mathcal{K}_2)$ there exist $f_i \in \mathcal{C}(\mathcal{K}_i); i = 1, 2,;$ such that $f_1 + f_2 \geq f$. The component channels together can do what the product channel can do. This way, both, the non–stationarity and perpetuum mobile problem are taken care of.

We wonder how all this looks in the light of "quantum parallelism".

We finally quote statements by Shannon. In [37] he writes "Theorem 4, of course, is analogous to known results for the ordinary capacity $C$, *where the product channel has the sum of the ordinary capacities* and the sum channel has an equivalent number of letters equal to the sum of the equivalent numbers of letters for the individual channels. We conjecture, but have not been able to

prove, that the equalities in Theorem 4 hold in general – not just under the conditions given". Both conjectures have been disproved (Haemers and Alon).

## 12 Identification

Ahlswede and Dueck, considering not the problem that the receiver wants to recover a message (*transmission problem*), but wants to decide whether or not the sent message is identical to an arbitrarily chosen one (*identification problem*), defined an $(n, N, \lambda_1, \lambda_2)$ identification (ID) code to be a collection of pairs

$$\{(P_i, \mathcal{D}_i) : i = 1, \ldots, N\},$$

with probability distributions $P_i$ on $\mathcal{X}^n$ and $\mathcal{D}_i \subset \mathcal{Y}^n$, such that the error probabilities of first resp. second kind satisfy

$$P_i W^n(\mathcal{D}_i^c) = \sum_{x^n \in \mathcal{X}^n} P_i(x^n) W^n(\mathcal{D}_i^c | x^n) \leq \lambda_1,$$

$$P_j W^n(\mathcal{D}_i) = \sum_{x^n \in \mathcal{X}^n} P_j(x^n) W^n(\mathcal{D}_i | x^n) \leq \lambda_2,$$

for all $i, j = 1, \ldots, N$, $i \neq j$. Define $N(n, \lambda_1, \lambda_2)$ to be the maximal $N$ such that a $(n, N, \lambda_1, \lambda_2)$ ID code exists.

With these definitions one has for a DMC

**Theorem.** (Ahlswede, Dueck [12]) *For every $\lambda_1, \lambda_2 > 0$ and $\delta > 0$, and for every sufficiently large $n$*

$$N(n, \lambda_1, \lambda_2) \geq \exp(\exp(n(C(W) - \delta))).$$

The next two sections are devoted to a (comparably short) proof of the following strong converse

**Theorem.** *Let $\lambda_1, \lambda_2 > 0$ such that $\lambda_1 + \lambda_2 < 1$. Then for every $\delta > 0$ and every sufficiently large $n$*

$$N(n, \lambda_1, \lambda_2) \leq \exp(\exp(n(C(W) + \delta))).$$

The strong converse to the coding theorem for identification via a DMC was conjectured in [12] (In case of complete feedback the strong converse was established already in [13]) and proved by Han and Verdu [31] and in a simpler way in [30]. However, even the second proof is rather complicated. The authors emphasize that they used and developed analytical methods and take the position that combinatorial techniques for instance of [6], [7] find their limitations on this kind of problem (see also Newsletter on Moscow workshop in 1994). We demonstrate now that this is not the case (see also the remarks on page XIX of [C1]).

Here we come back to the very first idea from [12], essentially to replace the distributions $P_i$ by uniform distributions on "small" subsets of $\mathcal{X}^n$, namely with cardinality slightly above $\exp(nC(W))$.

## 13 A Novel Hypergraph Covering Lemma

The core of the proof is the following result about hypergraphs. Recall that a *hypergraph* is a pair $\Gamma = (\mathcal{V}, \mathcal{E})$ with a finite set $\mathcal{V}$ of vertices, and a finite set $\mathcal{E}$ of (hyper–) edges $E \subset \mathcal{V}$. We call $\Gamma$ $e$–uniform, if all its edges have cardinality $e$. For an edge $E \in \mathcal{E}$ denote the characteristic function of $E \subset \mathcal{V}$ by $1_E$.

A result from large deviation theory will be used in the sequel:

**Lemma 1.** *For an i.i.d. sequence $Z_1, \ldots, Z_L$ of random variables with values in $[0,1]$ with expectation $\mathbb{E}Z_i = \mu$, and $0 < \varepsilon < 1$*

$$\Pr\left\{\frac{1}{L}\sum_{i=1}^{L} Z_i > (1+\varepsilon)\mu\right\} \leq \exp(-LD((1+\varepsilon)\mu\|\mu)),$$

$$\Pr\left\{\frac{1}{L}\sum_{i=1}^{L} Z_i < (1-\varepsilon)\mu\right\} \leq \exp(-LD((1-\varepsilon)\mu\|\mu)),$$

*where $D(\alpha\|\beta)$ is the information divergence of the binary distributions $(\alpha, 1-\alpha)$ and $(\beta, 1-\beta)$. Since*

$$D((1+\varepsilon)\mu\|\mu) \geq \frac{\varepsilon^2\mu}{2\ln 2} \text{ for } |\varepsilon| \leq \frac{1}{2},$$

*it follows that*

$$\Pr\left\{\frac{1}{L}\sum_{i=1}^{L} Z_i \notin [(1-\varepsilon)\mu, (1+\varepsilon)\mu]\right\} \leq 2\exp\left(-L\cdot\frac{\varepsilon^2\mu}{2\ln 2}\right).$$

**Proof:** The first two inequalities are for instance a consequence of Sanov's Theorem (c.f. [21], also Lemma LD in [12]). The lower bound on $D$ is elementary calculus.

**Lemma 2.** (Novel hypergraph covering, presented also in "Winter School on Coding and Information Theory, Ebeltoft, Dänemark, Dezember 1998" and in "Twin Conferences: 1. Search and Complexity and 2. Information Theory in Mathematics, Balatonelle, Ungarn, July 2000".)

*Let $\Gamma = (\mathcal{V}, \mathcal{E})$ be an $e$–uniform hypergraph, and $P$ a probability distribution on $\mathcal{E}$. Define the probability distribution $Q$ on $\mathcal{V}$ by*

$$Q(v) = \sum_{E\in\mathcal{E}} P(E)\frac{1}{e}1_E(v),$$

*and fix $\varepsilon, \tau > 0$. Then there exist vertices $\mathcal{V}_0 \subset \mathcal{V}$ and edges $E_1, \ldots, E_L \in \mathcal{E}$ such that with*

$$\bar{Q}(v) = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{e}1_{E_i}(v)$$

*the following holds:*

$$Q(\mathcal{V}_0) \leq \tau,$$
$$\forall v \in \mathcal{V} \setminus \mathcal{V}_0 \quad (1-\varepsilon)Q(v) \leq \bar{Q}(v) \leq (1+\varepsilon)Q(v),$$
$$L \leq 1 + \frac{|\mathcal{V}|}{e}\frac{2\ln 2\log(2|\mathcal{V}|)}{\varepsilon^2\tau}.$$

For ease of application we formulate and prove a slightly more general version of this:

**Lemma 3.** *Let $\Gamma = (\mathcal{V}, \mathcal{E})$ be a hypergraph, with a measure $Q_E$ on each edge $E$, such that $Q_E(v) \leq \eta$ for all $E$, $v \in E$. For a probability distribution $P$ on $\mathcal{E}$ define*

$$Q = \sum_{E\in\mathcal{E}} P(E)Q_E,$$

*and fix $\varepsilon, \tau > 0$. Then there exist vertices $\mathcal{V}_0 \subset \mathcal{V}$ and edges $E_1, \ldots, E_L \in \mathcal{E}$ such that with*

$$\bar{Q} = \frac{1}{L}\sum_{i=1}^{L} Q_{E_i}$$

*the following holds:*

$$Q(\mathcal{V}_0) \leq \tau,$$
$$\forall v \in \mathcal{V} \setminus \mathcal{V}_0 \quad (1-\varepsilon)Q(v) \leq \bar{Q}(v) \leq (1+\varepsilon)Q(v),$$
$$L \leq 1 + \eta|\mathcal{V}|\frac{2\ln 2\log(2|\mathcal{V}|)}{\varepsilon^2\tau}.$$

**Proof:** Define i.i.d. random variables $Y_1, \ldots, Y_L$ with

$$\Pr\{Y_i = E\} = P(E) \text{ for } E \in \mathcal{E}.$$

For $v \in \mathcal{V}$ define $X_i = Q_{Y_i}(v)$. Clearly $\mathbb{E}X_i = Q(v)$, hence it is natural to use a large deviation estimate to prove the bounds on $\bar{Q}$. Applying Lemma 1 to the random variables $\eta^{-1}X_i$ we find

$$\Pr\left\{\frac{1}{L}\sum_{i=1}^{L} X_i \notin [(1-\varepsilon)Q(v), (1+\varepsilon)Q(v)]\right\} \leq 2\exp\left(-L\cdot\frac{\varepsilon^2 Q(v)}{2\eta ln2}\right).$$

Now we define

$$\mathcal{V}_0 = \left\{v \in \mathcal{V} : Q(v) < \frac{1}{|\mathcal{V}|}\tau\right\},$$

and observe that $Q(\mathcal{V}_0) \leq \tau$. Hence,

$$\Pr\left\{\exists v \in \mathcal{V}\setminus\mathcal{V}_0 : \frac{1}{L}\sum_{i=1}^{L} Q_{Y_i}(v) \notin [(1-\varepsilon)Q(v), (1+\varepsilon)Q(v)]\right\}$$
$$\leq 2|\mathcal{V}|\exp\left(-L\cdot\frac{\varepsilon^2\tau}{2\eta|\mathcal{V}|ln2}\right).$$

The right hand side becomes less than 1, if

$$L > \eta|\mathcal{V}|\frac{2ln2\log(2|\mathcal{V}|)}{\varepsilon^2\tau},$$

hence there exist instances $E_i$ of the $Y_i$ with the desired properties.

The interpretation of this result is as follows: $Q$ is the expectation measure of the measures $Q_E$, which are sampled by the $Q_{E_i}$. The lemma says how close the sampling average $\bar{Q}$ can be to $Q$. In fact, assuming $Q_E(E) = q \leq 1$ for all $E \in \mathcal{E}$, one easily sees that

$$\|Q - \bar{Q}\|_1 \leq 2\varepsilon + 2\tau.$$

## 14 Proof of Converse

Let $\{(P_i, D_i) : i = 1, \ldots, N\}$ be a $(n, N, \lambda_1, \lambda_2)$ ID code, $\lambda_1 + \lambda_2 = 1 - \lambda < 1$. Our goal is to construct a $(n, N, \lambda_1 + \lambda/3, \lambda_2 + \lambda/3)$ ID code $\{(\bar{P}_i, D_i) : i = 1, \ldots, N\}$ with $KL$–distributions $\bar{P}_i$ on $\mathcal{X}^n$, i.e. all the probabilities are rational with common denominator $KL$ to be specified below.

Fix $i$ for the moment. For a distribution $T$ on $\mathcal{X}$ we introduce

$$\mathcal{T}_T^n = \{x^n \in \mathcal{X}^n : \forall x\ N(x|x^n) = nT(x)\},$$

and call $T$ *empirical distribution* if this is nonempty. There are less than $(n+1)^{|\mathcal{X}|}$ many empirical distributions.

For an empirical distribution $T$ define

$$P_i^T(x^n) = \frac{P_i(x^n)}{P_i(\mathcal{T}_T^n)} \text{ for } x^n \in \mathcal{T}_T^n,$$

which is a probability distribution on $\mathcal{T}_T^n$ (which we extend by 0 to all of $\mathcal{X}^n$). Note:

$$P_i = \sum_{T \text{ emp. distr.}} P_i(\mathcal{T}_T^n)P_i^T.$$

For $x^n \in \mathcal{T}_T^n$ and

$$\alpha = \sqrt{\frac{9|\mathcal{X}||\mathcal{Y}|}{\lambda}}$$

we consider the set of *conditional typical sequences*

$$\mathcal{T}_{W,\alpha}^n(x^n) = \{y^n \in \mathcal{Y}^n : \ ...\}.$$

It is well known that these sets are contained in the set of *$TW$–typical sequences* on $\mathcal{Y}^n$,

$$\mathcal{T}_{TW,..\alpha}^n = \{y^n \in \mathcal{Y}^n : \ ...\}.$$

Define now the measures $Q_{x^n}$ by

$$Q_{x^n}(y^n) = W^n(y^n|x^n) \cdot 1_{\mathcal{T}_{W,\alpha}^n(x^n)}(y^n).$$

By the properties of typical sequences and choice of $\alpha$ we have

$$\|Q_{x^n} - W(\cdot|x^n)\|_1 \leq \frac{\lambda}{9}.$$

Now with $\varepsilon = \tau = \lambda/36$ apply Lemma 3 to the hypergraph with vertex set $\mathcal{T}^n_{TW,..\alpha}$ and edges $\mathcal{T}^n_{W,\alpha}(x^n)$, $x^n \in \mathcal{T}^n_T$, carrying measure $W(\cdot|x^n)$, and the probability distribution $P_i^T$ on the edge set: we get a $L$–distribution $\bar{P}_i^T$ with

$$\|P_i^T Q - \bar{P}_i^T Q\|_1 \leq \frac{\lambda}{9},$$

$$L \leq \exp(nI(T;W) + O(\sqrt{n})) \leq \exp(nC(W) + O(\sqrt{n})),$$

where the constants depend explicitly on $\alpha, \delta, \tau$. By construction we get

$$\|P_i^T W^n - \bar{P}_i^T W^n\|_1 \leq \frac{\lambda}{3}.$$

In fact by the proof of the lemma we can choose $L = \exp(nC(W) + O(\sqrt{n}))$, independent of $i$ and $T$.

Now chose a $K$–distribution $R$ on the set of all empirical distributions such that

$$\sum_{T \text{ emp.distr.}} |P_i(\mathcal{T}^n_T) - R(T)| \leq \frac{\lambda}{3},$$

which is possible for

$$K = \lceil 3(n+1)^{|\mathcal{X}|}/\lambda \rceil.$$

Defining

$$\bar{P}_i = \sum_{T \text{ emp.distr.}} R(T)\bar{P}_i^T$$

we can summarize

$$\frac{1}{2}\|P_i W^n - \bar{P}_i W^n\|_1 \leq \frac{\lambda}{3},$$

where $\bar{P}_i$ is a $KL$–distribution. Since for all $\mathcal{D} \subset \mathcal{Y}^n$

$$|P_i W^n(\mathcal{D}) - \bar{P}_i W^n(\mathcal{D})| \leq \frac{1}{2}\|P_i W^n - \bar{P}_i W^n\|_1$$

the collection $\{(\bar{P}_i, \mathcal{D}_i) : i = 1, \ldots, N\}$ is indeed a $(n, N, \lambda_1 + \lambda/3, \lambda_2 + \lambda/3)$ ID code.

The proof is concluded by two observations: because of $\lambda_1 + \lambda_2 + 2\lambda/3 < 1$ we have $\bar{P}_i \neq \bar{P}_j$ for $i \neq j$. Since the $\bar{P}_i$ however are $KL$–distributions, we find

$$N \leq |\mathcal{X}^n|^{KL} = \exp(n \log |\mathcal{X}| \cdot KL) \leq \exp(\exp(n(C(W) + \delta))),$$

the last if only $n$ is large enough.

## References


1. R. Ahlswede, Certain results in coding theory for compound channels, Proc. Colloquium Inf. Th. Debrecen (Hungary), 35–60, 1967.
2. R. Ahlswede, Beiträge zur Shannonschen Informationstheorie im Fall nichtstationärer Kanäle, Z. Wahrscheinlichkeitstheorie und verw. Geb. 10, 1–42, 1968. (Dipl. Thesis Nichtstationäre Kanäle, Göttingen 1963.)
3. R. Ahlswede, The weak capacity of averaged channels, Z. Wahrscheinlichkeitstheorie und verw. Geb. 11, 61–73, 1968.
4. R. Ahlswede, On two–way communication channels and a problem by Zarankiewicz, Sixth Prague Conf. on Inf. Th., Stat. Dec. Fct's and Rand. Proc., Sept. 1971, Publ. House Chechosl. Academy of Sc., 23–37, 1973.
5. R. Ahlswede, An elementary proof of the strong converse theorem for the multiple–access channel, J. Combinatorics, Information and System Sciences, Vol. 7, No. 3, 216–230, 1982.
6. R. Ahlswede, Coloring hypergraphs: A new approach to multi–user source coding I, Journ. of Combinatorics, Information and System Sciences, Vol. 4, No. 1, 76–115, 1979.
7. R. Ahlswede, Coloring hypergraphs: A new approach to multi–user source coding II, Journ. of Combinatorics, Information and System Sciences, Vol. 5, No. 3, 220–268, 1980.
8. 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; Problems of Information Transmission, Vol. 32, No. 1, 123–138, 1996.
9. 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, 1993.
10. R. Ahlswede and I. Csiszár, Common randomness in information theory and cryptography, part II: CR capacity, Preprint 95–101, SFB 343 Diskrete Strukturen in der Mathematik, Universität Bielefeld, IEEE Trans. Inf. Theory, Vol. 44, No. 1, 55–62, 1998.
11. R. Ahlswede and G. Dueck, Every bad code has a good subcode: a local converse to the coding theorem, Z. Wahrscheinlichkeitstheorie und verw. Geb. 34, 179–182, 1976.
12. R. Ahlswede and G. Dueck, Identification via channels, IEEE Trans. Inf. Theory, Vol. 35, No. 1, 15–29, 1989.
13. R. Ahlswede and G. Dueck, Identification in the presence of feedback — a discovery of new capacity formulas, IEEE Trans. on Inf. Theory, Vol. 35, No. 1, 30–39, 1989.
14. R. Ahlswede and B. Verboven, On identification via multi–way channels with feedback, IEEE Trans. Information Theory, Vol. 37, No. 5, 1519–1526, 1991.
15. R. Ahlswede and J. Wolfowitz, The structure of capacity functions for compound channels, Proc. of the Internat. Symposium on Probability and Information Theory at McMaster University, Canada, April 1968, 12–54, 1969.
16. R. Ahlswede and Z. Zhang, New directions in the theory of identification via channels, Preprint 94–010, SFB 343 Diskrete Strukturen in der Mathematik, Universität Bielefeld, IEEE Trans. Information Theory, Vol. 41, No. 4, 1040–1050, 1995.
17. R. Ahlswede, N. Cai, and Z. Zhang, Erasure, list, and detection zero–error capacities for low noise and a relation to identification, Preprint 93–068, SFB 343 Diskrete Strukturen in der Mathematik, Universität Bielefeld, IEEE Trans. Information Theory, Vol. 42, No. 1, 55–62, 1996.

18. R. Ahlswede, P. Gács, and J. Körner, Bounds on conditional probabilities with applications in multiuser communication, Z. Wahrscheinlichkeitstheorie und verw. Geb. 34, 157–177, 1976.
19. R. Ahlswede, General theory of information transfer, Preprint 97–118, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, 1997; General theory of information transfer:updated, General Theory of Information Transfer and Combinatorics, a Special Issue of Discrete Applied Mathematics, to appear.
20. R. Ash, Information Theory, Interscience Tracts in Pure and Applied Mathematics, No. 19, Wiley & Sons, New York, 1965.
21. T.M. Cover and J.A. Thomas, Elements of Information Theory, Wiley, Series in Telecommunications, J. Wiley & Sons, 1991.
22. I. Csiszár and J. Körner, Information Theory — Coding Theorem for Discrete Memoryless Systems, Academic, New York, 1981.
23. R.L. Dobrushin, General formulation of Shannon's main theorem of information theory, Usp. Math. Nauk., 14, 3–104, 1959. Translated in Am. Math. Soc. Trans., 33, 323–438, 1962.
24. R.M. Fano, Transmission of Information: A Statistical Theory of Communication, Wiley, New York, 1961.
25. A. Feinstein, Foundations of Information Theory, McGraw–Hill, New York, 1958.
26. R.G. Gallager, A simple derivation of the coding theorem and some applications, IEEE Trans. Inf. Theory, 3–18, 1965.
27. R.G. Gallager, Information Theory and Reliable Communication, J. Wiley and Sons, Inc., New York, 1968.
28. T.S. Han, Oral communication in 1998.
29. T.S. Han, Information – Spectrum Methods in Information Theory, April 1998 (in Japanese).
30. T.S. Han and S. Verdú, Approximation theory of output statistics, IEEE Trans. Inf. Theory, IT–39(3), 752–772, 1993.
31. T.S. Han and S. Verdú, New results in the theory of identification via channels, IEEE Trans. Inf. Theory, Vol. 39, No. 3, 752–772, 1993.
32. K. Jacobs, Almost periodic channels, Colloquium on Combinatorial Methods in Probability Theory, 118–126, Matematisk Institute, Aarhus University, August 1–10, 1962.
33. F. Jelinek, Probabilistic Information Theory, 1968.
34. H. Kesten, Some remarks on the capacity of compound channels in the semicontinuous case, Inform. and Control 4, 169–184, 1961.
35. M.S. Pinsker, Information and Stability of Random Variables and Processes, Izd. Akad. Nauk, 1960.
36. C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal, Vol. 27, 379–423, 623–656, 1948.
37. C.E. Shannon, The zero error capacity of a noisy channel, IRE, Trans. Inf. Theory, Vol. 2, 8–19, 1956.
38. C.E. Shannon, Certain results in coding theory for noisy channels, Inform. and Control 1, 6–25, 1957.
39. S. Verdú and T.S. Han, A general formula for channel capacity, IEEE Trans. Inf. Theory, Vol. 40, No. 4, 1147–1157, 1994.
40. J. Wolfowitz, The coding of messages subject to chance errors, Illinois Journal of Mathematics, 1, 591–606, 1957.
41. J. Wolfowitz, Coding theorems of information theory, 3rd. edition, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 31, Springer-Verlag, Berlin-New York, 1978.
42. A.D. Wyner, The capacity of the product channel, Information and Control 9, 423–430, 1966.

**Appendix: Concepts of Performance from Number Theory**

We can identify the $0-1$–sequence $(a_t)_{t=1}^{\infty}$ with the set of numbers $\mathcal{A} \subset \mathbb{N}$, where

$$\mathcal{A} = \{t \in \mathbb{N} : a_t = 1\}. \tag{A.1}$$

Then the lower asymptotic density equals the pessimistic mean, so

$$\underline{d}(\mathcal{A}) = \underline{A} \tag{A.2}$$

and the upper asymptotic density equals the optimistic mean, so

$$\overline{d}(\mathcal{A}) = \overline{A}. \tag{A.3}$$

If both coincide they agree with the asymptotic density $d(\mathcal{A})$. Another well–known and frequently used concept is logarithmic density $\delta$ again with lower and upper branches

$$\underline{\delta}(\mathcal{A}) = \underline{\lim_{n\to\infty}} \frac{1}{\log n} \sum_{\substack{a\in\mathcal{A}\\ a\le n}} \frac{1}{a} \tag{A.4}$$

$$\overline{\delta}(a) = \overline{\lim_{n\to\infty}} \frac{1}{\log n} \sum_{\substack{a\in\mathcal{A}\\ a\le n}} \frac{1}{a}. \tag{A.5}$$

If they are equal, then the logarithmic density $\delta(\mathcal{A}) = \underline{\delta}(\mathcal{A}) = \overline{\delta}(\mathcal{A})$ exists.

Equivalently, they can be written in the form of (lower, upper, ...) Dirichlet densities

$$\underline{\delta}(\mathcal{A}) = \underline{\lim_{s\to 1^+}} \sum_{a\in\mathcal{A}} \frac{1}{a^s} \tag{A.6}$$

$$\overline{\delta}(\mathcal{A}) = \overline{\lim_{s\to 1^+}} \sum_{a\in\mathcal{A}} \frac{1}{a^s} \tag{A.7}$$

which often can be handled analytically more easily.

It is well–known that for every $\mathcal{A} \subset \mathbb{N}$

$$\underline{d}(\mathcal{A}) \le \underline{\delta}(\mathcal{A}) \le \overline{\delta}(\mathcal{A}) \le \overline{d}(\mathcal{A}). \tag{A.8}$$

Whereas the measures of the previous Section $\underline{\underline{A}}$ and $\overline{\overline{A}}$ are *outside* the interval $\big(\underline{d}(\mathcal{A}), \overline{d}(\mathcal{A})\big)$ these measures are *inside*.

Operationally their meaning is not so clear except that they put more weight on the beginning of the sequence – a realistic property where time is limited.

Even though they don't seem to have an immediate information theoretical interpretation, they get one as bounds on the limit points of $(A_n)_{n=1}^{\infty}$ and also on $\underline{A}$, $\overline{A}$. For instance in a widely developed calculus on pessimistic capacities $\underline{\delta}$ helps in evaluations.

The other famous concept of density in Number Theory is

$$\sigma(\mathcal{A}) = \inf_{n\geq 1} \frac{1}{n}|\{a \in \mathcal{A} : a \leq n\}|, \tag{A.9}$$

the Schnirelmann density. It is in so far peculiar as $1 \notin \mathcal{A}$ implies already $\sigma(\mathcal{A}) = 0$.

As first application we consider a situation where the communicators have *restrictions* on transmission lengths $n$ and on phases $m$, say to be members of $\mathbb{N}$ and $\mathcal{M}$. Following these rules, what are the time points at which there can be activity? One answer is the

**Lemma** (Schnirelmann)**.** *Let* $0 \in \mathcal{M} \subset \mathbb{N} \cup \{0\}$ *and* $0 \in \mathbb{N} \subset \mathbb{N} \cup \{0\}$*, if* $\sigma(\mathcal{M}) + \sigma(B) \geq 1$*, then* $n \in \mathcal{M} + \mathbb{N}$ *for every* $n \in \mathbb{N}$.

But now we come closer to home.

**Definition:** For channel $\mathcal{K}$ we define for every $\lambda \in (0,1)$ the *Schnirelmann* $\lambda$*–capacity*

$$S(\lambda) = \sigma\left(\left\{\frac{1}{n}\log M(n,\lambda) : n \in \mathbb{N}\right\}\right).$$

A pleasant property of $\sigma$ is that $\sigma(\mathcal{A}) = \gamma$ implies

$$\frac{1}{n}|\{a \in \mathcal{A} : a \leq n\}| \geq \gamma \text{ for all } n \in \mathbb{N}. \tag{A.10}$$

Therefore $\frac{1}{n}\log M(n,\lambda) \geq S(\lambda)$ for all $n$. For a DMC we have for the quantity $\min\limits_{\lambda>0} S(\lambda) = \log M(1,0) \leq C_{\text{zero}}(W)$.

$S(\lambda)$ lower bounds the pessimistic $\lambda$–capacity (see [15])

$$\underline{C}(\lambda) = \underline{\lim}_{n\to\infty} \frac{1}{n}\log M(n,\lambda).$$

**Remark 8:** This quantity in conjunction with a weak converse has been determined (except for finitely many discontinuities in [15]) for compound channels with the average error criterion, after it was noticed in [3] that for this error concept – as opposed to the maximal error concept – there is no strong converse.

The behaviour of $\underline{C}(\lambda)$ is the same as for average errors for the case of maximal errors *and* randomisation in the encoding. Conjunction of average error criterion and randomisation lead to no improvement.

**Problem:** For which DMC's and for which $\lambda$ do we have

$$S(\lambda) = \underline{C}(\lambda)?$$

For instance consider a BSC $\left(\begin{smallmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{smallmatrix}\right)$ and $\lambda > \varepsilon$, then $\log M(1,\lambda) = 1$. On the other hand we know that $\underline{C}(\lambda) = 1 - h(\varepsilon)$. For $\lambda$ large enough it is conceivable that $\frac{1}{n}\log M(n,\lambda) \geq 1 - h(\varepsilon)$ for all $n \in \mathbb{N}$. For general channels $\mathcal{K}$ many things can happen.

Theoretically and practically it is still meaningful to investigate $S(\lambda)$ where it is smaller than $\underline{C}(\lambda)$.