

Chapter 6

Towards Capacity Functions

6.1 Lecture on Concepts of Performance Parameters for Channels

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.

6.1.1 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 \mathcal{O} be any set, whose elements are called output symbols.

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

$$W = \{W(\cdot|i) : i \in \mathcal{I}\} \quad (6.1)$$

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

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 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

$$\mathcal{K} = \{W^n : \mathcal{X}^n \rightarrow \mathcal{Y}^n : n \in \mathbb{N}\}. \quad (6.2)$$

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

$$W^n(y^n|x^n) = \prod_{t=1}^n W(y_t|x_t) \quad (6.3)$$

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

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

$$W^n(y^n|x^n) = \prod_{t=1}^n W_t(y_t, x_t). \quad (6.4)$$

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. [2, 37]).

ADMC: Suppose now that we have two channels \mathcal{K}_1 and \mathcal{K}_2 as defined in (6.2). Then following [3] we can associate with them an *averaged* channel

$$\mathcal{A} = \left\{ \left(\frac{1}{2} W_1^n + \frac{1}{2} W_2^n : \mathcal{X}^n \rightarrow \mathcal{Y}^n \right) : n \in \mathbb{N} \right\} \quad (6.5)$$

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 chapters (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.

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

6.1.2 Three Unquestioned Concepts: The Two Most Basic, Code Size and Error Probability, then Further Block Length

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

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

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

$$M = |\mathcal{C}| \text{ is the code size} \quad (6.6)$$

$$e(\mathcal{C}) = \max_{i \in I} W(D_i^c | u_i) \quad (6.7)$$

is the (maximal) probability of error and

$$\bar{e}(\mathcal{C}) = \frac{1}{M} \sum_{i=1}^M W(D_i^c | u_i) \quad (6.8)$$

is the average probability of error.

One can study now the functions

$$M(\lambda) = \max_{\mathcal{C}} \{|\mathcal{C}| : e(\mathcal{C}) \leq \lambda\} \text{ (resp. } \bar{M}(\lambda)) \quad (6.9)$$

and

$$\lambda(M) = \min_{\mathcal{C}} \{e(\mathcal{C}) : |\mathcal{C}| = M\} \text{ (resp. } \bar{\lambda}(M)), \quad (6.10)$$

that is, finiteness, growth, convexity properties etc.

It is convenient to say that \mathcal{C} is an (M, λ) -code, if

$$|\mathcal{C}| \geq M \text{ and } e(\mathcal{C}) \leq \lambda. \quad (6.11)$$

Now we add time-structure, that means here, we go to the channel defined in (6.2). 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 (6.9), (6.10), and (6.11) are to be modified accordingly:

$$M(n, \lambda) = \max_{\mathcal{C}(n)} \{ |\mathcal{C}(n)| : e(\mathcal{C}(n)) \leq \lambda \} \quad (6.12)$$

$$\lambda(n, M) = \min_{\mathcal{C}(n)} \{ e(\mathcal{C}(n)) : |\mathcal{C}(n)| = M \} \quad (6.13)$$

$$\mathcal{C}(n) \text{ is an } (M, n, \lambda)\text{-code, if } |\mathcal{C}(n)| \geq M, e(\mathcal{C}(n)) \leq \lambda. \quad (6.14)$$

Remark One could study blocklength as function of M and λ 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.

6.1.3 Stochastic Inequalities: The Role of the Information Function

We consider a channel $W : \mathcal{X} \rightarrow \mathcal{Y}$ 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 [47] we associate with (P, W) or \tilde{P} the *information function (per letter)* $I : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, where

$$I(x, y) = \begin{cases} \log \frac{\tilde{P}(x, y)}{P(x)Q(y)} \\ 0 \end{cases}, \text{ if } \tilde{P}(x, y) = 0. \quad (6.15)$$

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

$$F(\alpha) = \Pr\{I(X, Y) \leq \alpha\} = \tilde{P}(\{(x, y) : I(x, y) \leq \alpha\}). \quad (6.16)$$

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

Theorem 36 (Shannon 1957, [47]) *For a canonical $(M, \bar{\lambda})$ -code and the corresponding information function there are the relations*

$$\frac{1}{2}F\left(\log \frac{M}{2}\right) \leq \bar{\lambda} \leq F\left(\log \frac{M}{2}\right). \quad (6.17)$$

Remarks

1. 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

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

2. Let us emphasize that all of Shannon's bounds involve the information function (per letter), which is highlighted also in Fano [22], where it is called mutual information. 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.

6.1.4 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 Chap. 2). Generally speaking, with any function $f : \mathbb{N} \rightarrow \mathbb{R}_+$ (or, equivalently, any sequence $(f(1), f(2), f(3), \dots)$ of non-negative numbers) we can associate a rate function $\text{rate}(f)$, where

$$\text{rate}(f(n)) = \frac{1}{n} \log f(n). \quad (6.18)$$

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

$$\text{rate}_n(f) \triangleq \text{rate}(f(n)) = \frac{1}{n} \log f(n). \quad (6.19)$$

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 $n \in \mathbb{N}$, λ constant. Then $\text{rate}(f)$ 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 Sect. 6.1 the results—not necessarily the strongest known—relevant for our discussion.

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

- (i) 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

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

- (ii) 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 $\max_P \min_{i=1,2} I(W_i|P)$ [3]) such that

- (i) holds
- (ii) 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 $(C(n))_{n=1}^\infty$ (which actually can be chosen as $C(n) = \frac{1}{n} \sum_{t=1}^n \max_P I(W_t|P)$ [2]) such that

- (i') 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

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

- (ii') 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 (i), say (iii') is often still true, but also not always.)

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

The pair of statements [(i), (ii)] 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 (i) and (ii), so this channel *does not have* a strong capacity.

The pair of statements [(i), (ii)] 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 [24] one reads “The converse to the coding theorem is stated and proved in varying degrees of generality in Chaps. 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 Chap. 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 [17].)

Now the NDMC has in general no strong and no weak capacity (see our example in Sect. 6.1.7).

However, if we replace the concept of capacity by that of a capacity function $(C(n))_{n=1}^{\infty}$ then the pair [(a'), (ii')] (resp. [(i'), (iii')]) 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 [18, 19], Pinsker [44]. 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 [(i), (iii)] or [(i'), (iii')] concepts. Equivalent formulations are

- (i') $\inf_{\lambda > 0} \lim_{n \rightarrow \infty} \left(\frac{1}{n} \log M(n, \lambda) - C(n) \right) \geq 0$
- (ii') for all $\lambda \in (0, 1) \quad \overline{\lim}_{n \rightarrow \infty} \left(\frac{1}{n} \log M(n, \lambda) - C(n) \right) \leq 0$
- (iii') $\inf_{\lambda > 0} \overline{\lim}_{n \rightarrow \infty} \left(\frac{1}{n} \log M(n, \lambda) - C(n) \right) \leq 0.$

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

Remarks

1. A standard way of expressing (iii) is: for rates above capacity the error probability is bounded away from 0 for *all large* n . ([23], called “*partial converse*” on page 44.)
2. There are cases (c.f. [3]), where the uniformity in λ valid in (ii) or (ii') holds only for $\lambda \in (0, \lambda_1)$ with an absolute constant λ_1 —a “medium” strong converse. It also occurs in “second order” estimates of [32] with $\lambda_1 = \frac{1}{2}$.
3. There are cases where (iii) [or (iii')] don't hold for constant λ 's but for $\lambda = \lambda(n)$ going to 0 sufficiently fast, in one case [9] like $\frac{1}{n}$ and in another like $\frac{1}{n^4}$ [7]. 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 λ_n could be considered.
4. For our key example in Sect. 6.1.7 [(i'), (iii')] holds, but not [(i), (iii)]. 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 the first remark of this lecture, this is weaker than (c) and equivalent to

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

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

Definition 29 \underline{C} is called the (pessimistic) capacity of a channel \mathcal{K} , if it is the supremum over all numbers C for which (i) holds. Since $C = 0$ satisfies (a), the number $\underline{C} = \underline{C}(\mathcal{K})$ exists. Notice that there are no requirements concerning (ii) or (iii) 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 as large as \underline{C} .

Definition 30 \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

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

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

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

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

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

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

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

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

In the light of old work [2] we were shocked when we learnt that these two definitions were given in [17] 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 [16, 22, 50], and the optimistic capacity \overline{C} is used in [17] on page 223 and in [38]).

Fano [22] 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} .

6.1.5 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.

1. Recently Cover explained that under this view for the broadcast channel ($W : \mathcal{X} \rightarrow \mathcal{Y}, V : \mathcal{X} \rightarrow \mathcal{Z}$) the rate pair $(R_{\mathcal{Y}}, R_{\mathcal{Z}}) = (C(W), C(V))$ is in the capacity region, which in fact equals $\{(R_{\mathcal{Y}}, R_{\mathcal{Z}}) : 0 \leq R_{\mathcal{Y}} \leq C(W), 0 \leq R_{\mathcal{Z}} \leq C(W)\}$. Just assign periodically time intervals of lengths $m_1, n_1, m_2, n_2, m_3, n_3, \dots$ 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.
2. The reader may quickly convince himself that $\overline{C} = \min_{s \in \mathcal{S}} C(W(\cdot|\cdot, s)) \geq \max_P \min_s I(W(\cdot|\cdot, s)|P)$ 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!

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

For the semicontinuous compound channel, $|S| = \infty$, the ordinary weak capacity [(i),(iii)] 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).

6.1.6 A “Paradox” for Products of Channels

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

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

A chapter by Wyner [50] is very instructive for our discussion. We quote therefore literally the beginning of the chapter (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 chapter 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 37 (i) Let C^* be the capacity of the product of s channels with capacities C_1, C_2, \dots, C_s respectively. Then

$$C^* = \sum_{j=1}^s C_j. \quad (6.20)$$

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

The proof of (i) 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 $C^* \geq \sum_{j=1}^s C_j$. In the second (“weak converse”) we will show that no $R > \sum_{j=1}^s C_j$ is a permissible rate, establishing that $C^* \leq \sum_{j=1}^s C_j$. The proof of (ii) 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 (iii) in Sect. 6.1.4 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 Sect. 6.1.7 shows that (6.20) is in general not true. The two factor channels used in the example do not 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.

6.1.7 The Pessimistic Capacity Definition: An Information-Theoretic Perpetuum Mobile

Consider the two matrices $V^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $V^0 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. 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 $\underline{C}(\mathcal{K}, \mathcal{K}^*) = 1$.

6.1.8 A Way Out of the Dilemma: Capacity Functions

If $M(n, \lambda)$ fluctuates very strongly in n and therefore also $\text{rate}_n(M)$, 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 Sect. 6.1.7 $\underline{C}(\mathcal{K}) = \underline{C}(\mathcal{K}^*) = 0$ and on the other hand $\overline{C}(\mathcal{K}) = \overline{C}(\mathcal{K}^*) = 1$. In contrast we can choose the sequence $(c_n)_{n=1}^\infty = (\frac{1}{n} \sum_{t=1}^n C(W_t))_{n=1}^\infty$ for channel \mathcal{K} and $(c_n^*)_{n=1}^\infty = (\frac{1}{n} \sum_{t=1}^n C(W_t^*))_{n=1}^\infty$ 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 38 *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 (i).

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” ([49]) 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.

6.1.9 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 $\{(W_{t+m})_{t=1}^{\infty} : 0 \leq m < \infty\}$ 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 $(C(W_t))_{t=1}^{\infty}$ 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

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

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 $C(n) = \frac{1}{n} \sum_{t=1}^n C_t \in [0, 1]$.

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

First of all there are the familiar

$$\underline{A} = \lim_{n \rightarrow \infty} A_n \text{ (the pessimistic mean)} \quad (6.21)$$

$$\overline{A} = \overline{\lim}_{n \rightarrow \infty} A_n \text{ (the optimistic mean).} \quad (6.22)$$

We introduce now a new concept

$$A = \lim_{n \rightarrow \infty} \inf_{m \geq 0} A_{m+1, m+n} \text{ (the pessimistic phase independent mean).} \quad (6.23)$$

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

$$\overline{\overline{A}} = \overline{\lim}_{n \rightarrow \infty} \sup_{m \geq 0} A_{m+1, m+n} \text{ (super optimistic mean).} \quad (6.24)$$

We shall show first

Lemma 55

$$\overline{\lim}_{n \rightarrow \infty} \inf_{m \geq 0} A_{m+1, m+n} = \underline{A} \quad (6.25)$$

$$\underline{\lim}_{n \rightarrow \infty} \sup_{m \geq 0} A_{m+1, m+n} = \overline{A} \quad (6.26)$$

Proof We prove only (6.25), the proof for (6.26) being “symmetrically” the same. We have to show that

$$\underline{A} = \underline{\lim}_{n \rightarrow \infty} \inf_{m \geq 0} A_{m+1, m+n} \geq \overline{\lim}_{n \rightarrow \infty} \inf_{m \geq 0} A_{m+1, m+n}. \quad (6.27)$$

□

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 ε of \underline{A} and choose a much bigger N^* for which $A_{m(N^*)+1, m(N^*)+N^*}$ is within ε of the expression at the right side of (6.27) and $N^* \geq \frac{1}{\varepsilon} n^*$ holds.

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

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

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

$$\inf_m \underline{\lim}_{n \rightarrow \infty} A_{m+1, m+n} = \sup_m \underline{\lim}_{n \rightarrow \infty} A_{m+1, m+n} = \underline{A} \quad (6.28)$$

$$\inf_m \overline{\lim}_{n \rightarrow \infty} A_{m+1, m+n} = \sup_m \overline{\lim}_{n \rightarrow \infty} A_{m+1, m+n} = \overline{A}, \quad (6.29)$$

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 $\underline{A} \leq \underline{A} \leq \overline{A} \leq \overline{A}$ 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: $(k_1, \ell_1, k_2, \ell_2, k_3, \dots)$

Examples

1. $k_t = k, \ell_t = \ell$ for $t = 1, 2, \dots$; a periodic case:

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

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

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

3. $k_t = \sum_{s=1}^{t-1} k_s, \ell_t = \sum_{s=1}^{t-1} \ell_s$ for $t = 1, 2, \dots$

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

Here all four values are different.

4. $k_t = \sum_{s=1}^{t-1} k_s, \ell_t = t$ for $t = 2, 3, \dots, k_1 = 1$

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

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).

6.1.10 Some Comments on a Formula for the Pessimistic Capacity

A noticeable observation of Verdú and Han [48] 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 [31, 32], 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 [31] 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] played 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 [37] 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 Sect. 6.1.4 for precise terminology).
2. The term weak capacity was introduced.
3. The weak capacity (and also the λ -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 $\lim_{n \rightarrow \infty} \frac{1}{n} \max_{X^n} I(X^n \wedge Y^n)$ does not describe the weak capacity in general. Compare this with Wyner's first capacity definition in Sect. 6.1.6.
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 [48] 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

$$(\mathbf{X}, \mathbf{Y}) = (X^n, Y^n)_{n=1}^{\infty} \quad (6.30)$$

and

$$\underline{I}(\mathbf{X} \wedge \mathbf{Y}) = \sup \left\{ \alpha : \lim_{n \rightarrow \infty} \Pr \left\{ (x^n, y^n) : \frac{1}{n} I(x^n, y^n) \leq \alpha \right\} = 0 \right\}. \quad (6.31)$$

Their general formula asserts

$$\underline{C} = \sup_{\mathbf{X}} \underline{I}(\mathbf{X} \wedge \mathbf{Y}). \quad (6.32)$$

The reader should be aware that

- α . The stochastic inequalities used for the derivation (10.3) are both (in particular also Theorem 4 of [48]) not new.
- β . Finally, there is a very important point. In order to show that a certain quantity K (for instance $\sup_{\mathbf{X}} \underline{I}(\mathbf{X} \wedge \mathbf{Y})$) 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 $\inf_{\lambda} \lim_{n \rightarrow \infty} \log M(n, \lambda) < K + \delta$. For this one does *not need* the *weak* converse (ii) $\inf_{\lambda} \lim_{n \rightarrow \infty} \log M(n, \lambda) \leq K$, but only

$$\inf_{\lambda} \lim_{n \rightarrow \infty} \log M(n, \lambda) \leq K \quad (6.33)$$

(see also Sect. 6.1.4) The statement may be termed the “weak-weak converse” or the “weak-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 perpetual 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 $\sup_{\mathbf{X}} \bar{I}(\mathbf{X} \wedge \mathbf{Y})$ 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.

6.1.11 Pessimistic Capacity Functions

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

Definition 31 A sequence $(C_n)_{n=1}^{\infty}$ of non-negative numbers is a capacity sequence of \mathcal{K} , if

$$\inf_{\lambda > 0} \lim_{n \rightarrow \infty} \left(\frac{1}{n} \log M(n, \lambda) - C_n \right) = 0.$$

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

Replacing α by α_n in (6.31) 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 μ for the sequences may be used and accordingly one gets say $\mathcal{C}(\mathcal{K}, \mu \leq \rho)$, a space of capacity functions of μ -complexity less than ρ .

This seems to be a big machinery, but channels \mathcal{K} with no connections between W^n and W^m 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 perpetual 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 [46] 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).

6.2 Lecture on Comments to “On Concepts of Performance Parameters for Channels”

This contribution has appeared in the preprint series of the SFB 343 at Bielefeld University in the year 2000 [6].

Its declared purpose was to open a discussion about basic concepts in Information Theory in order to gain more clarity. This had become necessary after the work [12] for identification gave also new interpretation to classical theory of transmission leading in particular to “A general formula for channel capacity” by S. Verdú and T.S. Han, which however, does not cover the results of [2].

The final answer is not found until now.

We give now a brief reaction to some comments and criticism we received.

1. Strong objections were made against statement α) after (6.32), which implies the claim that the inequality

$$\bar{\lambda} \geq F(\log M - \Theta) - e^{-\Theta}, \quad \Theta > 0 \quad (6.34)$$

is essentially not new. Particularly it has been asserted that this inequality is not comparable to Shannon’s

$$\bar{\lambda} \geq \frac{1}{2} F\left(\log \frac{M}{2}\right) \quad (6.35)$$

in (6.17) of Shannon’s Theorem, because it is stronger.

Therefore we have to justify our statement by a proof. Indeed, just notice that Shannon worked with the constant $\frac{1}{2}$ for simplicity in the same way as one usually extracts from a code with average error probability $\bar{\lambda}$ a subcode of size $\frac{M}{2}$ with maximal probability of error not exceeding $2\bar{\lambda}$. However, again by the pigeon-hole principle for any $\beta \in (0, \frac{1}{2})$ there are $M\beta$ codewords with individual error probabilities $\leq \frac{\bar{\lambda}}{1-\beta}$. (This argument was used in [1, 2, 13]).

Now just replace in Shannon’s proof $\frac{1}{2}$ by β to get the inequality

$$\bar{\lambda} \geq (1 - \beta)F(\log M\beta). \quad (6.36)$$

Equating now $F(\log M - \Theta)$ with $F(\log M\beta)$ we get $\beta = e^{-\Theta}$ and it suffices to show that

$$(1 - e^{-\Theta})F(\log M - \Theta) \geq F(\log M - \Theta) - e^{-\Theta}.$$

Indeed $e^{-\Theta} \geq e^{-\Theta}F(\log M - \Theta)$, because F is a distribution function. So (6.36) is even slightly stronger than (6.34). \square

For beginners we carry out the details.

Introduce $W^*(u|y) = \frac{\tilde{P}(u,y)}{Q(y)}$ and notice that $I(u, y) \leq \log M\beta$ is equivalent with $\frac{W^*(u|y)}{P(u)} \leq M\beta$ or, since $P(u) = M^{-1}$,

$$W^*(u|y) \leq \beta. \quad (6.37)$$

Now concentrate attention on those pairs (u, y) for which (6.37) holds.

Consider the bipartite graph $G = (\mathcal{U}, \mathcal{Y}, \mathcal{E})$ with vertex sets $\mathcal{U} = \{u_1, \dots, u_M\}$, \mathcal{Y} , and edge set $\mathcal{E} = \{(u, y) : W^*(u|y) \leq \beta\}$

Clearly,

$$\tilde{P}(\mathcal{E}) = F(\log M\beta) \quad (6.38)$$

We partition now \mathcal{Y} into

$$\mathcal{Y}_+ = \{y \in \mathcal{Y} : \text{exists } u \text{ with } W^*(u|y) > \beta\}, \quad \mathcal{Y}_- = \mathcal{Y} \setminus \mathcal{Y}_+ \quad (6.39)$$

and correspondingly we partition \mathcal{E} into

$$\mathcal{E}_+ = \{(u, y) \in \mathcal{E} : y \in \mathcal{Y}_+\}, \quad \mathcal{E}_- = \mathcal{E} \setminus \mathcal{E}_+. \quad (6.40)$$

Clearly $\tilde{P}(\mathcal{E}) = \tilde{P}(\mathcal{E}_+) + \tilde{P}(\mathcal{E}_-)$. For $y \in \mathcal{Y}_+$ ML-decoding chooses a u with $W^*(u|y) > \beta$, but (u, y) is not in \mathcal{E} and not in \mathcal{E}_+ . Therefore all $(u', y) \in \mathcal{E}_+$ contribute to the error probability. The total contribution is $\tilde{P}(\mathcal{E}_+)$.

The contribution of the edges in \mathcal{E}_- to the error probability is, if f is the ML-decoder,

$$\sum_{y \in \mathcal{Y}_-} Q(y)(1 - W^*(f(y)|y)) \geq \tilde{P}(\mathcal{E}_-)(1 - \beta)$$

and hence

$$\bar{\lambda} \geq \tilde{P}(\mathcal{E}_+) + \tilde{P}(\mathcal{E}_-)(1 - \beta)$$

(even slightly stronger than 6.36), written explicitly

$$\bar{\lambda} \geq \tilde{P}(\{(u, y) : \log \frac{W(y|u)}{Q(y)} \leq \log M - \Theta\}) - e^{\Theta} \tilde{P}(\{(u, y) :$$

$$\text{for all } u' \in \mathcal{U} \log \frac{W(y|u')}{Q(y)} \leq \log M - \Theta\}).$$

For those who do not accept it as Shannon's result it would be only consequential to name it then the Shannon/Ahlsweide inequality.

2. We also have some good news. In Sect. 6.1.5 we argued that the optimistic capacity concept seems absurd and we provided convincing examples.

Investigations [10] in Cryptography made us aware that this concept, a dual to the pessimistic capacity, finds a natural place here, because *one wants to protect also against enemies having fortunate time for themselves in using their wire-tapping channel!*

3. Finally, being concerned about performance criteria, we should not forget that in Information Theory, similarly as in Statistics, asymptotic theory gives a first coarse understanding, but never should be the last word.

In particular with the availability of a lot of computing power not only small, but even medium size samples call for algorithmic procedures with suitable parameters. This was the message from Ziv in his contribution [52].

4. S. Dodunekov in [20] explained that for linear codes with parameters block-length n , dimension k and minimal distance d , fixing two parameters and optimizing the third of the quantities (a) $N_q(k, d)$, (b) $K_q(n, d)$, (c) $D_q(n, k)$ *the first one gives the most accurate description.*

5. It has been pointed out that there are different concepts of capacity, but that usually in a chapter it is clearly explained which is used and a definition is given. It can be felt from those reactions that Doob's famous (and infamous with respect to Shannon and his work) criticism [21] about the way in which Information Theory proceeds from Coding Theorem n to Coding Theorem $n + 1$ keeps people alert.

That mostly researchers know which concepts they are using and that they sometimes even give definitions, that is still not enough.

For instance we have been reminded that our statement "there seems to be no definition of the weak converse in the book [17]" is wrong and that a look at the index leads to the problem session, where a definition is given. This is correct.

However, it is not stated there on page 112 that this definition of a converse, referred to by us as "weak weak converse, ... or ...", is not the definition, which was established in Information Theory at least since the book [49] by Wolfowitz came out in 1961 and was used at least by most of the Mathematicians and Statisticians working in the field.

Unfortunately this happened even though since 1974 we had many discussions and joint work with Körner and Csiszár, a great part of which entered the book and influenced the shape of the later parts of the book. It was also given to us for reading prior to publication and we have to apologize for being such a poor reader. Otherwise we would have noticed what we only noticed in 1998.

It is clear now why [2] and [3] are not cited in the book, because they don’t fit into the frame.

This frame became the orientation for people starting to learn Information Theory via typical sequences. Shannon’s stochastic inequalities ([47]) perhaps were not taught anymore.

6. We know that the weak capacity has the additivity property for parallel channels. We draw attention to the fact that *Shannon (and also later Lovász) conjectured* this property to hold also for his zero-error capacity (which was disproved by Haemers [27]). Apparently, Shannon liked to have this property!

We all do, often naively, *time-sharing*, which is justified, if there is an additivity property!

We like to add that without thinking in terms of time-sharing we never would have discovered and proved our characterization of the (weak) capacity region for the MAC in [4] (with a different proof in [5]).

So our message is “Shannon also seems to think that additivity is an important property” and not “Shannon made a wrong conjecture”.

The additivity property for quantum channels is of great interest to the community of Quantum Informationtheorists. This led M. Horodecki to his question quoted in [36]. The answer is positive for degraded channels, but not in general!

7. Once the situation is understood it is time to improve it. We suggest below a unified description of capacity concepts with conventions for their notations.

In every science it is occasionally necessary to agree on some standards—a permanent fight against the second law of thermodynamics. We all know how important the settings of such standards have been in Physics, Chemistry, Biology etc. Every advice to the standards proposed here is welcome.

We start here with the case corresponding to B) under 4. above and restrict ourself to one-way channels.

We consider a general channel \mathcal{K} with time structure, which is defined in (6.2) of [6]. Recall that for a positive integer n and a non-negative real number λ . $N(n, \lambda)$ denotes for \mathcal{K} the maximal cardinality of a code of block-length n with an error probability not exceeding λ .

Often λ depends on n and we write $\lambda(n)$. The sequence $\{\lambda(n)\}$ is typically of one of the following forms

- (i) $\lambda(n) = 0$ for all $n \in \mathbb{N}$
- (ii) $\lambda(n) = \lambda$, $0 < \lambda \leq 1$, for all $n \in \mathbb{N}$
- (iii) $\lambda(n) = n^{-\alpha}$, $0 < \alpha$, for all $n \in \mathbb{N}$
- (iv) $\lambda(n) = e^{-En}$, $0 < E$, for all $n \in \mathbb{N}$

We speak about zero, constant, polynomial, and exponential error probabilities.

With the sequence $\{N(n, \lambda(n))\}$ we associate a very basic and convenient performance parameter for the channel \mathcal{K} the

$$\text{rate-error function } R : \Lambda \rightarrow \mathbb{R}_+,$$

where Λ is the space of all non-negative real-valued sequences and for every $\{\lambda(n)\} \in \Lambda$.

$R(\{\lambda(n)\})$ is the largest real number with $\frac{1}{n} \log N(n, \lambda(n)) \geq R(\{\lambda(n)\}) - \delta$ for every $\delta > 0$ and all large n .

How does it relate to capacities? In the four cases described we get for $R(\{\lambda(n)\})$ the values

- (i') $C(0)$, that is, Shannon's zero error capacity
- (ii') $C(\lambda)$, that is, the λ -capacity introduced in [13]
- (iii') $C(\alpha)$, that is, a novel α -capacity
- (iv') $C(E)$, that is, the E -capacity introduced by Evgueni Haroutunian in [33].

A word about notation is necessary. The functions $C(0)$, $C(\lambda)$, $C(\alpha)$, and $C(E)$ are distinguished only by their arguments, these will always appear explicitly. All our results have to be interpreted with this understanding.

This convention was made already in [13] where not only the maximal error probability λ but also the average error probability $\bar{\lambda}$, the maximal error probability λ_R for randomized encoding, and the average error probability $\bar{\lambda}_R$ for randomized encoding were considered. For example, one of our theorems in [13] says that

$$C(\lambda_R) = C(\bar{\lambda}) = C(\bar{\lambda}_R).$$

under certain conditions where $\lambda_R = \bar{\lambda} = \bar{\lambda}_R$. Taken literally this is a trivial statement. In the light of our notation it means that these functions coincide for certain values of the argument. This notation result is no confusion or ambiguity, and has the advantage of suggestiveness new and typographical simplicity.

An important point about the introduced rate-error function and the capacities is their existence for every channel \mathcal{K} .

The same is the case for the (ordinary) capacity

$$C = \inf_{0 < \lambda \leq 1} C(\lambda).$$

Our rate-error function may be called pessimistic and it has an optimistic twin $\bar{R}(\{\lambda(n)\})$, the largest real number with

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log N(n, \lambda(n)) \geq \bar{R}(\{\lambda(n)\}).$$

Correspondingly we get the optimistic capacities $\bar{C}(0)$, $\bar{C}(\lambda)$, $\bar{C}(\alpha)$, $\bar{C}(E)$, and \bar{C} . Of course for a DMC $\bar{C}(0) = C(0)$ etc.

They are relevant, for example, if a wire-tapper chooses the times of an attack.

Again all these quantities exist. Moreover, the error criteria $\bar{\lambda}$, λ_R , $\bar{\lambda}_R$ lead to analogs of the capacities in (iii)–(iv'), namely, $C(\bar{\alpha})$, $C(\alpha_R)$, $C(\bar{\alpha}_R)$, $C(\bar{E})$, $C(E_R)$, and $C(\bar{E}_R)$ and similarly for \bar{C} . In [13] compound channels, a certain class of channels, were considered. The concepts are even more relevant for the more sophisticated AVC, for example.

8. Well-known is Shannon’s *rate-distortion function* in source coding. It is amazing that our preceding analogue for channel coding was not introduced. However, it must be said that Shannon introduced his function “informationally” (and so does Toby Berger in [15]) and not “operational” as we did. In channel coding he gave two definitions for the channel capacity, an informational and an operational one. This is very well discussed in the explanation of Aaron Wyner, which we cite in [6]. Unfortunately, some well-known text books like [16] or [51] give the informational one. But $\{\max_{p^n} I(W^n|P^n)\}$ describes the operational capacity C only in special cases like the DMC and is too large for instance for averaged channels (memory!). Here lies one of the main roots for conceptional confusions about channel capacity!

9. As we have explained earlier a nice property to have is additivity of a capacity for parallel channels. This is the case for the ordinary capacity C , if $C = \bar{C}$ and this is exactly in the case where the *weak converse* holds. We also say in this case that the *weak capacity* exists (see [3]). *So this quantity does not exist automatically.* This is sometimes overlooked and even more true for the strong capacity introduced in [3], when the strong converse holds.

This must be kept in mind when we compare results. Generality is of course easier to obtain for capacities with weaker properties than for those with stronger properties.

In proving upper bounds like the weak converse it is helpful to prove first bounds, which can be obtained easier, like polynomial or exponential (see also soft converse in [12]) weak converses discussed in [7].

10. We just indicate that in the spirit of [6] one should introduce also

rate sequence – error functions

- saying in particular much more about non-stationary channels than rate-error functions,

classes of rate sequence – error functions

- catching tighter descriptions suggested in Sect. 11 of [6].

Of course associated with these performance criteria are capacity concepts.

11. The discussion should be continued and should some time include other performance criteria like the analoga to (i) and (iii) above. Also analoga of (i), (ii), (iii) for combinatorial channel models and also criteria for sources are to be classified.

It should be specific about distinctions stemming from multi-way channels—feedback situations—non-block codes—delay—synchronization.

Beyond capacities for Shannon’s transmission there are those for identification (second order)—common randomness—general information transfer (first and second order).

In combinatorial channel models to be distinguished are the various error concepts:

failure of error detection or wrong detection—defects—insertions—deletions—localized errors—unidirectional errors—etc.

The case of feedback just brings us to search (see Chap. 6) with the recently studied lies with cost constraints etc.

Also J. Körner's models of a combinatorial universe with information aspects, starting with ambiguous alphabets and going along trifferent paths, definitely must be included (Sperner Capacity).

Recent work by G. Katona and K. Tichler in search deserves immediate attention. There a test is a partition of the search space \mathcal{X} into 3 sets ($\mathcal{Y}, \mathcal{N}, \mathcal{A}$). If the object x searched for satisfies $x \in \mathcal{Y}$ the answer is Yes, if it satisfies $x \in \mathcal{N}$ the answer is No, it is for $x \in \mathcal{A}$ arbitrary Yes or No.

Also to be classified are the performance criteria in the very important work on codes introduced by Kautz/Singleton [39] and studied by Lindström, Dyachkov, Erdős and many others (see survey [26]).

The results found an application in the probabilistic model of K -identification in [7].

Finally, analogous performance criteria are to be defined in Statistics in particular in the interplay between Multi-user Source Coding Theory and Hypothesis Testing or Estimation starting with work [8, 11], with Csiszar and Burnashev and continued by many others (see survey [30] by Han and Amari).

6.3 Lecture on the “Capacity” of the Product Channel

6.3.1 The Strong Converse of the Coding Theorem

Central in the proof of the strong converse for the DMC is the following lemma. As Fano's Lemma it does not make use of the time structure, i.e., the block length n is not involved and can hence be chosen w.l.o.g. as $n = 1$.

Let \mathcal{X} and \mathcal{Y} denote finite sets, $W = (w(y|x))_{x \in \mathcal{X}, y \in \mathcal{Y}}$ a stochastic matrix and let Q be an arbitrary probability distribution on \mathcal{Y} . Further for all $x \in \mathcal{X}$ we have some $\Theta_x > 0$ and define

$$B_x(\Theta_x, Q) \triangleq \left\{ y \in \mathcal{Y} : \frac{w(y|x)}{Q(y)} \geq 2^{\Theta_x} \right\}.$$

Lemma 56 *If for an (N, λ) -code $\{(u_i, D_i) : i = 1, \dots, N\}$ ($u_i \in \mathcal{X}, D_i \subset \mathcal{Y}$ for all i , $D_i \cap D_j = \emptyset$ for all $i \neq j$) there exists a probability distribution Q on \mathcal{Y} with $\max_{u_i} \sum_{y \in B_{u_i}(\Theta_{u_i}, Q)} w(y|u_i) < \gamma$ and $\lambda + \gamma < 1$, then*

$$N < (1 - \lambda - \gamma)^{-1} 2^{\frac{1}{N} \sum_{i=1}^N \Theta_{u_i}}.$$

Proof We shortly write Θ_i instead of Θ_{u_i} and set for $i = 1, \dots, N$

$$A_i \triangleq \left\{ y \in D_i : \frac{w(y|u_i)}{Q(y)} < 2^{\Theta_i} \right\} = D_i \cap (B_{u_i}(\Theta_i, Q))^c. \quad \square$$

For $y \in A_i$ then by definition $2^{\Theta_i} Q(y) > w(y|u_i)$ and hence for all i

$$\begin{aligned} 2^{\Theta_i} Q(D_i) &\geq 2^{\Theta_i} Q(A_i) > w(A_i|u_i) = w(D_i|u_i) - w(D_i \setminus A_i|u_i) \\ &\geq w(D_i|u_i) - B_{u_i}(\Theta_i, Q) \geq 1 - \lambda - \gamma \end{aligned}$$

by the assumptions. Division by $Q(D_i)$ and taking logarithm on both sides yields for all $i = 1, \dots, n$ $\Theta_i \geq \log \left(\frac{1-\lambda-\gamma}{Q(D_i)} \right)$ and hence

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \Theta_i &\geq \frac{1}{N} \sum_{i=1}^N \log \left(\frac{1-\lambda-\gamma}{Q(D_i)} \right) = - \sum_{i=1}^N \log Q(D_i) + \log(1-\lambda-\gamma) \\ &> - \sum_{i=1}^N \frac{1}{N} \log \frac{1}{N} + \log(1-\lambda-\gamma) = \log N + \log(1-\lambda-\gamma), \end{aligned}$$

since the relative entropy $\sum_{i=1}^N \frac{1}{N} \log \left(\frac{\frac{1}{N}}{Q(D_i)} \right) > 0$.

Exponentiation on both sides yields the statement of the lemma. \square

Theorem 39 (Strong Converse for the DMC) *For all $\lambda \in (0, 1)$*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log N(n, \lambda) \leq C.$$

The strong converse for the DMC was first proved by Wolfowitz. The following proof is due to Kemperman.

Proof With the notation of the preceding lemma we now choose $\mathcal{X} = \mathcal{X}^n$, $\mathcal{Y} = \mathcal{Y}^n$ and $W = W^n$ and let $\{(u_i, D_i) : i = 1, \dots, N\}$ be an (n, N, λ) -code. We partition $\mathcal{U} \triangleq \{u_1, \dots, u_N\}$ into subcodes $\mathcal{U}(P)$, where each code word in $\mathcal{U}(P)$ is of type $P \in \mathcal{P}$. \square

Let the type P^* be such that $\mathcal{U}(P^*)$ is a subcode of maximum length $M \triangleq |\mathcal{U}(P^*)|$. Then $N \leq (n+1)^{|\mathcal{X}|} \cdot M$.

Further let $Q \triangleq P^* \cdot W$. In each code word $u = (u_{(1)}, \dots, u_{(n)}) \in \mathcal{U}(P^*)$ by definition every $x \in \mathcal{X}$ has frequency exactly $\langle u|x \rangle = P^*(x) \cdot n$. We denote by \mathbb{E}_P and Var_P the expected value and the variance of a random variable on a set \mathcal{X} with probability distribution P .

$$\begin{aligned}
\mathbb{E}_{w(\cdot|u)} \log \frac{w(\cdot|u)}{Q^n(\cdot)} &= \sum_{t=1}^n \mathbb{E}_{w(\cdot|u_{(t)})} \log \frac{w(\cdot|u_{(t)})}{Q(\cdot)} \\
&= \sum_{t=1}^n \sum_{y \in \mathcal{Y}} w(y|u_{(t)}) \log \frac{w(y|u_{(t)})}{Q(y)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} n P^*(x) w(y|x) \cdot \log \frac{w(y|x)}{Q(y)} \\
&= n \cdot I(P^*, W).
\end{aligned}$$

Next we shall show that $\text{Var}_{w(\cdot|u)} \cdot \log \frac{w(\cdot|u)}{Q^n(\cdot)}$ is bounded.

$$\begin{aligned}
\text{Var}_{w(\cdot|u)} \log \frac{w(\cdot|u)}{Q^n(\cdot)} &= \sum_{t=1}^n \text{Var}_{w(\cdot|u_{(t)})} \log \frac{w(\cdot|u_{(t)})}{Q(\cdot)} \\
&\leq \sum_{t=1}^n \mathbb{E}_{w(\cdot|u_{(t)})} \left(\log \frac{w(\cdot|u_{(t)})}{Q(\cdot)} + \log P^*(u_{(t)}) \right)^2 \\
&\quad (\text{since } \text{Var}X \leq \mathbb{E}(X + a)^2 \text{ for all } a \in \mathbb{R}) \\
&= \sum_{t=1}^n \sum_{y \in \mathcal{Y}} w(y|u_{(t)}) \cdot \left(\log \frac{P^*(u_{(t)})w(y|u_{(t)})}{Q(y)} \right)^2 \\
&= \sum_{y \in \mathcal{Y}} n P^*(u_{(t)}) w(y|u_{(t)}) \cdot \left(\log \frac{P^*(u_{(t)})w(y|u_{(t)})}{Q(y)} \right)^2 \\
&= n \sum_{y \in \mathcal{Y}} Q(y) \frac{P^*(u_{(t)}) \cdot w(y|u_{(t)})}{Q(y)} \left(\log \frac{P^*(u_{(t)})w(y|u_{(t)})}{Q(y)} \right)^2.
\end{aligned}$$

Now observe that the function $x \cdot \log^2 x$ is uniformly continuous and therefore bounded by some constant. Hence the function $\sum_{i=1}^b x_i \log^2 x_i$ is also bounded for every fixed b . It can be shown that $\sum_{i=1}^b x_i \log^2 x_i \leq \max\{\log^2(3), \log^2(b)\}$, when $\sum_{i=1}^b x_i = 1$. Hence

$$\text{Var}_{w(\cdot|u)} \log \frac{w(\cdot|u)}{Q^n(u)} \leq n \cdot c, \text{ where } c \triangleq \max\{\log^2 3, \log^2 |\mathcal{Y}|\}.$$

Now choose for all $i = 1, \dots, N$ $\Theta_i = \Theta \triangleq nI(P^*, W) + \sqrt{\frac{2n}{1-\lambda}}c$ and $Q \triangleq Q^n$. Then by Chebyshev's Inequality

$$\begin{aligned}
w(B_u(\Theta, Q)|u) &= w\left(\frac{w(\cdot|u)}{Q^n(\cdot)} \geq 2^\Theta | u\right) \\
&= w\left(\log \frac{w(\cdot|u)}{Q^n(\cdot)} - \mathbb{E}_{w(\cdot|u)} \log \frac{w(\cdot|u)}{Q^n(\cdot)} \geq \sqrt{\frac{2nc}{1-\lambda}} | u\right) \\
&\leq \frac{nc}{\left(\sqrt{\frac{2nc}{1-\lambda}}\right)^2} = \frac{1-\lambda}{2}.
\end{aligned}$$

If we use the preceding lemma with $\gamma \triangleq \frac{1-\lambda}{2}$ we obtain

$$N \leq \left(1 - \lambda - \frac{1-\lambda}{2}\right)^{-1} 2^{nI(P^*, W) + \sqrt{\frac{2nc}{1-\lambda}}} = \frac{2}{1-\lambda} \exp \left\{ nI(P^*, W) + \sqrt{\frac{2nc}{1-\lambda}} \right\}.$$

From this we can conclude that

$$N \leq \exp \left\{ Cn + \sqrt{\frac{2nc}{1-\lambda}} + a \log(n+1) + \log \frac{2}{1-\lambda} \right\}.$$

This proves the strong converse. \square

6.3.2 On the “Capacity” of the Product of Channels

1. Wyner’s Approach

As in Lecture on Chap. 4 we define a parallel channel $W_1 \times \cdots \times W_k$. Each component is abstract, with or without memory, continuous or discrete, but with time structure.

Definition 32 A real number R is said to be a permissible rate of transmission for W if for every $\delta > 0$ and for n sufficiently large, there exists a code with parameter n with $M = \exp Rn$ codewords and error probability $\bar{\lambda} \leq \delta$. Since $R = 0$ is a permissible rate, the set of permissible rates is not empty. We define the channel capacity C as the supremum of permissible rates (see [CK])!

Definition 33 The strong converse holds if for any fixed $R > C \lim_{n \rightarrow \infty} \lambda(n, e^{Rn}) = 1$. Equivalently, for any $R > C$ and $\delta < 1$ and n sufficiently large, any (n, e^{Rn}) code has $\lambda \geq \delta$.

Remark

1. Augustin [14] has proved that these channels have a coding theorem with a weak converse with capacity C^* . However, they do not always have a strong converse. So there may be a lack of uniformity for $0 \leq \varepsilon < 1$, however we know now that there is always this uniformity for $0 \leq \varepsilon \leq \varepsilon_0$.

Actually this occurred under different circumstances for averaged channels and for compound channels for average errors.

2. Interesting is also a quotation from [50], where Wyner writes

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...

Wyner states that

Theorem 40 (i) *The capacity of the product channel (or parallel channel, or special nonstationary channel) is the sum of the capacities of the component channels.*

(ii) *The “strong converse” holds for the product channel, if it holds for each of the component channels.*

Proof Inductively it suffices to consider two factors

(i) To show that any $R < C_1 + C_2$ is permissible choose $R_1 + R_2 = R$, $R_i < C_i$ ($i = 1, 2$) and concatenate existing $(n, e^{R_i n})$ codes

$$\{(u_j^i, D_j^i) : 1 \leq j \leq M_i\}, \quad M_i = e^{R_i n}, \quad i = 1, 2,$$

with error probabilities λ_i and get

$$\{(u_j^1, u_k^2, D_j^1 \times D_k^2) : 1 \leq j \leq M_1, 1 \leq k \leq M_2\}$$

with error probability $\lambda \leq \lambda_1 + \lambda_2$.

Thus $C^2 \geq C_1 + C_2$. For the “converses” use the fact that from a listcode one can select at random a subcode (as in 5.3.3) such that the following relations hold. \square

Lemma 57 *Let W^n be a channel with input space $\mathcal{X}^n = \prod_1^n \mathcal{X}$ and output space $\mathcal{Y}^n = \prod_1^n \mathcal{Y}$, then for $M = e^{Rn}$, $L = e^{R_L n}$ and $R_1 < R - R_L$, $M_1 = e^{R_1 n}$*

$$\lambda(1, M_1, n) \leq \lambda(L, M, n) + \varepsilon(n), \quad (6.41)$$

where $\lim_{n \rightarrow \infty} \varepsilon(n) = 0$. (In particular, if the strong converse holds, then also $\lambda(e^{R_2 n}, e^{Rn}, n) \rightarrow 1$ as $n \rightarrow \infty$, if $R - R_L > C$.)

Additionally we use Lemma 60 in Lecture 7.1 and get in particular for the product channel

$$\lambda(1, M, n) \geq \lambda^{(1)}(L, M, n) \lambda^{(2)}(1, L, n). \quad (6.42)$$

Let now for the product channel be a code with rate $R > C_1 + C_2$.

Set $\eta = \frac{R - (C_1 + C_2)}{2} > 0$, $R_L = C_2 + \eta$ and $L = e^{R_L n}$. From (6.42)

$$\lambda \geq \lambda(1, e^{R_L n}, n) \geq \lambda^{(1)}(e^{R_L n}, e^{R_L n}, n) \lambda^{(2)}(1, e^{R_L n}, n). \quad (6.43)$$

Since $R_L > C_2$, the capacity of the second channel, necessarily $\lambda^{(2)}(1, e^{R_L n}, n)$ is “bounded away from 0”. Also by Lemma 57

$$\lambda^{(1)}(e^{R_L n}, e^{R_L n}, n) \geq \lambda^{(1)}(1, e^{R_L n}, n) + \varepsilon(n) \quad (6.44)$$

with $\varepsilon(n) \rightarrow 0$ as $n \rightarrow \infty$ and $R_2 < R - R_L$.

If we set $R_2 = C_1 + \frac{\eta}{2} = C_1 + \frac{R - C_1 - C_2}{4} = R - R_L - \eta/2 < R - R_L$ (6.44) is applicable.

Since $R_2 > C_1$, the capacity of channel 1, $\lambda^{(1)}(1, e^{R_2 n}, n)$ is “also bounded away from zero”. Hence by (6.44) $\lambda^{(1)}(e^{R_L n}, e^{R_L n}, n)$ is also “bounded away from zero”, so that by (6.43) λ is bounded away from zero and the weak converse is established.

(ii) In this reasoning replacing “bounded away from zero” by “bounded below by a constant $\gamma < 1$ ”. This holds for every $\gamma < 1$, the strong converse holds again.

Remarks

1. We actually proved (ii), but where is the flaw in the proof of (i)? (The product of two functions, each taking values 0 and 1 can be identical 0!)
2. “Usual” definitions of weak converse $\inf_{\lambda > 0} \lim_{n \rightarrow \infty} \frac{1}{n} \log M(n, \lambda) = C$ “cannot stay away from 0” again and again!

$$\text{Example } C^2 > C_1 + C_2 \quad W = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} W' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{array}{ccc} \frac{W \dots W}{n_1} & \frac{W' \dots W'}{n_2} & \text{Channel 1} \\ \frac{W' \dots W'}{n_1} & \frac{W \dots W}{n_2} & \text{Channel 2} \\ & C^2 = 1 & n_{t+1} = n_t^2 \\ & C_1 = 0, C_2 = 0 & \end{array}$$

Shows error in Wyner’s chapter (i). No strong converses rate $\frac{1}{2}$ has again and again error probability smaller than ε . Also no “weak converses” (Wyner’s sense) $\frac{1}{2} > R > 0 \exists \lambda(R) > 0$ but $\lambda_n(R)$ fluctuates between $0^+ 1^-$.

The capacity theorem is wrong!

More important: The “folklore” definition is paradoxical: Two channels have capacity 0, but combined they have capacity 1.

3. In my definition (for nonstationary channels) capacity functions of parallel channels *are additive* (modulo $o(n)$)!

2. Augustin's Approach

In [14] the formulation is this:

For an abstract channel \mathcal{K} let $M(\mathcal{K}, \varepsilon)$ be the maximal size of an $(1 - \varepsilon)$ -maximal probability of error code. Actually $M(\mathcal{K}, \varepsilon) = \sup\{M : M \text{ is length of an } (1 - \varepsilon)\text{-code for } \mathcal{K}\}$. Clearly, $1 \leq M(\mathcal{K}, \varepsilon) = \infty$. Let now $\mathcal{K}_1, \mathcal{K}_2$ be abstract channels and $\mathcal{K}^2 = \mathcal{K}_1 \times \mathcal{K}_2$.

Then obviously by concatenation

$$M(\mathcal{K}^2, \varepsilon_1 \cdot \varepsilon_2) \geq M(\mathcal{K}_1, \varepsilon_1)M(\mathcal{K}_2, \varepsilon_2). \quad (6.45)$$

Next we apply Augustin's version of using Fano*-sources for the packing lemma as for nonstationary channels. Formally, let $\{(u_i, D_i) : 1 \leq i \leq M\}$ be an $(1 - \varepsilon)$ -code for $\mathcal{K}_1 \times \mathcal{K}_2$ and let for $u_i = a_i b_i$

$$q \triangleq q_1 \times q_2 \triangleq \left(\frac{1}{M} \sum_{i=1}^M W_1(\cdot | a_i) \right) \times \left(\frac{1}{M} \sum_{i=1}^M W_2(\cdot | b_i) \right). \quad (6.46)$$

Then for arbitrary $\theta_1, \theta_2 > 0$

$$\begin{aligned} \theta_1 \theta_2 q(D_i) &\geq W \left(\left\{ \frac{dW(\cdot | u_i)}{dq} \leq \theta_1 \theta_2 \right\} \cap D_i | u_i \right) \\ &\geq q(D_i) - W \left(\left\{ \frac{dW(\cdot | u_i)}{dq} > \theta_1 \theta_2 \right\} | u_i \right) \\ &> \varepsilon - W \left(\left\{ \frac{dW(\cdot | u_i)}{dq} > \theta_1 \theta_2 \right\} | u_i \right) \\ &> \varepsilon - W_1 \left(\left\{ \frac{dW_1(\cdot | a_i)}{dq_1} > \theta_1 \right\} | a_i \right) - W_2 \left(\left\{ \frac{dW_2(\cdot | b_i)}{dq_2} > \theta_2 \right\} | b_i \right). \end{aligned}$$

Hence

$$\begin{aligned} \theta_1 \theta_2 \frac{1}{M} &\geq \theta_1 \theta_2 \frac{1}{M} \sum_{i=1}^M q(D_i) > \varepsilon - \frac{1}{M} \sum_{i=1}^M W_1 \left(\left\{ \frac{dw_1(\cdot | a_i)}{dq_1} > \theta_1 \right\} \right) \\ &\quad - \frac{1}{M} \sum_{i=1}^M W_2 \left(\left\{ \frac{dw_2(\cdot | b_i)}{dq_2} > \theta_2 \right\} \right). \end{aligned} \quad (6.47)$$

The maximal code estimate (see Chap. 5) gives

$$M(\mathcal{K}_j, \varepsilon_j^*) > \theta_j \left[\frac{1}{M_j} \sum_{i=1}^M W_j \left(\left\{ \frac{dW(\cdot | a_i)}{dq_j} > \theta_j \right\} \right) - \varepsilon_j^* \right] \quad (6.48)$$

$$0 < \varepsilon_j^* < 1.$$

Combining (6.47) and (6.48) gives

$$\theta_1 \theta_2 \frac{1}{M} > \varepsilon - (\varepsilon_1^* + \varepsilon_2^*) - \frac{M(\mathcal{K}_1, \varepsilon_1^*)}{\theta_1} - \frac{M(\mathcal{K}_2, \varepsilon_2^*)}{\theta_2}. \quad (6.49)$$

Assume now $\varepsilon_1^* + \varepsilon_2^* = (1 - c)\varepsilon$ for $0 < c < 1$ and $M(\mathcal{K}_1, \varepsilon_1^*) \cdot M(\mathcal{K}_2, \varepsilon_2^*) < \infty$, because otherwise (3) below is trivially true, and set

$$\theta_j = \frac{3M(\mathcal{K}_j, \varepsilon_j^*)}{c\varepsilon} \quad (j = 1, 2). \quad (6.50)$$

Then (6.49) yields

$$\frac{q}{c^2 \varepsilon^2} M(\mathcal{K}_1, \varepsilon_1^*) M(\mathcal{K}_2, \varepsilon_2^*) > \frac{c\varepsilon}{3} M(\mathcal{K}^2, \varepsilon) \quad (6.51)$$

and on the other hand for $\varepsilon \leq \varepsilon_1 \varepsilon_2$ by (6.45)

$$M(\mathcal{K}_1, \varepsilon_1) M(\mathcal{K}_2, \varepsilon_2) \leq M(\mathcal{K}^2, \varepsilon). \quad (6.52)$$

The two inequalities together give the

Theorem 41 *For the product channel $\mathcal{K}_1 \times \mathcal{K}_2$*

$$M(\mathcal{K}_1, \varepsilon_1) M(\mathcal{K}_2, \varepsilon_2) \leq M(\mathcal{K}^2, \varepsilon) \leq \left(\frac{3}{c\varepsilon} \right)^3 M(\mathcal{K}_1, \varepsilon_1^*) M(\mathcal{K}_2, \varepsilon_2^*) \quad (6.53)$$

where $\varepsilon \leq \varepsilon_1 \varepsilon_2$ and $\varepsilon_1^* + \varepsilon_2^* < (1 - c)\varepsilon$ ($0 < \varepsilon_1, \varepsilon_2, \varepsilon, \varepsilon_1^*, \varepsilon_2^* < 1$).

Introducing time structure we see that for instance the assumptions for $j = 1, 2$ imply

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M(\mathcal{K}_j, \lambda, n) = C_j \quad (0 < \lambda < 1)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M(\mathcal{K}^2, \lambda, n) = C_1 + C_2 \quad (0 < \lambda < 1).$$

Strong Converse and Coding Theorem (in this sense) are preserved under the product.

Remarks

1. In (6.45) we have a trivial lower bound on $M(\mathcal{K}^2, \varepsilon)$. The upper bound comes, as usual, from the Lagerungs Lemma, however, in terms of unknown information quantities, which we relate to $M(\mathcal{K}_j, \varepsilon)$ ($j = 1, 2$) by the conjunction with the maximal code lemma involving the same unknown quantities, to be eliminated.
2. We discuss this theorem in connection with our general discussion in Lectures 6.1 and 6.2 about rates, error exponents and capacity concepts.

6.4 Lecture on Every Channel with Time Structure has a Capacity Sequence

6.4.1 The Concept

On August 25th 2008 we lectured at the workshop in Budapest which honored I. Csiszár in the year of his 70th birthday, on the result that for every AVC under maximal error probability pessimistic capacity and optimistic capacity are equal. This strongly motivated us to think again about performance criteria and we came back to what we called already a long time ago [2] a (weak) capacity function (now sequence!). But this time we were bold enough to conjecture the theorem below. Its proof was done in hours. In the light of this shining observation we omit now the word “weak” which came from the connection with the weak converse and make the following

Definition 34 For a channel with time structure $\mathcal{K} = (W^n)_{n=1}^\infty$ $C : \mathbb{N} \rightarrow \mathbb{R}_+$ is a capacity sequence, if for maximal code size $M(n, \lambda)$, where n is the block length or time and λ is the permitted error probability, and the corresponding rate $R(n, \lambda) = \frac{1}{n} \log M(n, \lambda)$

$$\inf_{\lambda > 0} \underline{\lim}_{n \rightarrow \infty} (R(n, \lambda) - C(n)) \geq 0 \quad (6.54)$$

$$\inf_{\lambda > 0} \overline{\lim}_{n \rightarrow \infty} (R(n, \lambda) - C(n)) \leq 0. \quad (6.55)$$

Recall that the pessimistic capacity is $\underline{C} = \inf_{\lambda > 0} \underline{\lim}_{n \rightarrow \infty} R(n, \lambda)$ and the optimistic capacity is $\overline{C} = \inf_{\lambda > 0} \overline{\lim}_{n \rightarrow \infty} R(n, \lambda)$.

6.4.2 The Existence Result and Its Proof

Theorem 42 Every channel with time structure has a capacity sequence, if $\overline{C} > \infty$. Moreover, if (C, C, C, \dots) is a capacity sequence, then $C = \underline{C} = \overline{C}$.

Proof We use only that $R(n, \lambda)$ is not decreasing in λ . \square

Let $(\delta_l)_{l=1}^\infty$ be a null-sequence of positive numbers and let $(\lambda_l)_{l=1}^\infty$ be such that $\lambda_l \in (0, 1)$ and

$$\underline{C} + \delta_l \geq \lim_{n \rightarrow \infty} \frac{1}{n} R(n, \lambda) \geq \underline{C} \quad (6.56)$$

$$\overline{C} + \delta_l \geq \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} R(n, \lambda) \geq \overline{C}. \quad (6.57)$$

Moreover, let $(n_l)_{l=1}^\infty$ be a monotone increasing sequence of natural numbers such that for all $n \geq n_l$

$$\overline{C} + \delta_l \geq R(n, \lambda_l) \geq \underline{C} - \delta_l. \quad (6.58)$$

For $d_l = \left\lceil \frac{\overline{C} - \underline{C}}{\delta_l} \right\rceil$ define

$$A_l(i) = \{n : n_l \leq n, \overline{C} - (i-1)\delta_l \geq R(n, \lambda_l) \geq \overline{C} - i\delta_l\}, \text{ if } 2 \leq i \leq d_l - 1,$$

$$A_l(1) = \{n : n_l \leq n, \overline{C} + \delta_l \geq R(n, \lambda_l) \geq \overline{C} - \delta_l\},$$

$$A_l(d_l) = \{n : n_l \leq n, \overline{C} - (d_l - 1)\delta_l \geq R(n, \lambda_l) \geq \underline{C} - \delta_l\}.$$

Define by using lower end points

$$C(n) = \begin{cases} \overline{C} - i\delta_l & \text{for } n \in A_l(i), n < n_{l+1} \text{ and } 1 \leq i \leq d_l - 1 \\ \underline{C} - \delta_l & \text{for } n \in A_l(d_l), n < n_{l+1}. \end{cases}$$

Now for any $\lambda \in (0, 1)$ and $\lambda_l < \lambda$

$$R(n, \lambda) - C(n) \geq R(n, \lambda_{i+j}) - C(n) \geq 0$$

for $n_{l+j} \leq n < n_{l+j+1}$ and $j = 0, 1, 2, \dots$ and thus

$$\lim_{n \rightarrow \infty} R(n, \lambda) - C(n) \geq 0$$

and (6.54) follows.

Finally, for any $\lambda < \lambda_l$ by monotonicity of $R(n, \lambda)$

$$\overline{\lim}_{n \rightarrow \infty} R(n, \lambda) - C(n) \leq \overline{\lim}_{n \rightarrow \infty} R(n, \lambda_l) - C(n) \leq 2\delta_l$$

and

$$\inf_{\lambda \in (0, 1)} \overline{\lim}_{n \rightarrow \infty} R(n, \lambda) - C(n) \leq \lim_{l \rightarrow \infty} 2\delta_l = 0$$

and (6.55) holds. \square

Example The channel with

$$\frac{1}{n} \log M(n, \lambda) = \begin{cases} \lambda n & \text{for even } n \\ 0 & \text{for odd } n \end{cases} \quad \text{has } \underline{C} = 0, \overline{C} = \infty$$

and no capacity sequence.

6.4.3 Parallel Channels or the Product of Channels

Theorem 43 *For two channels with time structure $\mathcal{K}_1 = (W^n)_{n=1}^\infty$ and $\mathcal{K}_2 = (V^n)_{n=1}^\infty$, which have capacity sequences C_1 and C_2 , the product channel $\mathcal{K}_1 \times \mathcal{K}_2$ has capacity sequence*

$$C_{12} = C_1 + C_2.$$

This theorem was not easy to discover, because Wyner proved additivity of Wolfowitz's strong capacity and so did Augustin (subsequently) with an elegant proof based on the following lemma. Wyner's result of additivity of (pessimistic) capacities is false. We could save this part under the assumption $\underline{C} = \overline{C}$ for both channels, that is, if the weak converse holds in the usual sense with a constant. However, often it does not hold like for the channel $\mathcal{K} = AA \dots AB \dots$, where $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $B = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, and $\overline{C}(\mathcal{K}) = 1$, $\underline{C}(\mathcal{K}) = 0$, if alternating strings of A 's and B 's are sufficiently long (see [6]). We call it AB-channel.

Actually, switching the A 's and the B 's gives a nonstationary DMC (see [6]) \mathcal{K}' again with $\overline{C}(\mathcal{K}') = 1$, $\underline{C}(\mathcal{K}') = 0$. However, $C(\mathcal{K} \times \mathcal{K}') = 1 > 0 + 0!$

On the other hand additivity is a desirable property as can be seen from the fact that Shannon and later also Lóvasz conjectured it for the zero-error capacity (that means, I think, wanted it to hold). But it was disproved by Haemers.

Recently Hastings [34] disproved additivity for the capacity in the HSW-Theorem for quantum channels ([35, 45])

Lemma 58 (Augustin [14]) *For the product channel $\mathcal{K}_1 \times \mathcal{K}_2$ of abstract (time free, discrete) channels \mathcal{K}_1 and \mathcal{K}_2*

$$N(\mathcal{K}_1, \epsilon_1)N(\mathcal{K}_2, \epsilon_2) \leq N(\mathcal{K}_1 \times \mathcal{K}_2, \epsilon) \leq \left(\frac{3}{c\epsilon}\right)^3 N(\mathcal{K}_1, \epsilon_1^*)N(\mathcal{K}_2, \epsilon_2^*) \quad (6.59)$$

where $\epsilon \leq \epsilon_1\epsilon_2$ and $\epsilon_1^* + \epsilon_2^* < (1 - c)\epsilon$ ($0 < \epsilon_1, \epsilon_2, \epsilon, \epsilon_1^*, \epsilon_2^* < 1$) and $N(\cdot, \lambda) = M(\cdot, 1 - \lambda)$.

Remarks

1. In [6] we also wanted additivity to hold and found a complicated way by assigning *classes of functions* to a single channel as performance criterion. The present discovery is oriented at the “weak converse” and shows that a capacity sequence always exists under the mild assumption $\overline{C} < \infty$. So now the class of functions can be reduced to one modulo $o(W)$.
2. In [2] the strong converse was defined for the capacity sequence $(C_n)_{n=1}^\infty$ by

$$\frac{1}{n} \log M(n, \lambda) \leq C_n + \frac{o(n)}{n}. \quad (6.60)$$

3. Also in [2] the weak converse was defined by

$$\frac{1}{n} \log M(n, \lambda) \leq f(\lambda) C_n$$

for all large n and $\inf_\lambda f(\lambda) = 1$.

Proof of Theorem 43 Apply Lemma 58 for $W^n, V^n, \epsilon_1 = \epsilon_2 = 1 - \lambda, c = \frac{1}{2}, \epsilon_1^* = \epsilon_2^* = \frac{1-c}{2}(1 - \lambda)^2$ and conclude that

$$\begin{aligned} R(W^n, \lambda) + R(V^n, \lambda) &\leq R(W^n \times V^n, 2\lambda - \lambda^2) \\ &\leq \frac{1}{n} \log \frac{3}{c(1 - \lambda)^2} + R(W^n, \lambda') + R(V^n, \lambda') \end{aligned}$$

where $\lambda' = 1 - \frac{1-c}{2}(1 - \lambda)^2$.

$$\begin{aligned} &\underline{\lim}_{n \rightarrow \infty} (R(W^n \times V^n, \lambda) - C_1(n) - C_2(n)) \\ &\geq \underline{\lim}_{n \rightarrow \infty} (R(W^n, \lambda) - C_1(n)) + \underline{\lim}_{n \rightarrow \infty} (R(V^n, \lambda) - C_2(n)) \geq 0 \end{aligned}$$

also

$$\begin{aligned} &\inf_\lambda (R(W^n \times V^n, \lambda) - C_1(n) - C_2(n)) \geq 0 \\ &\inf_\lambda \overline{\lim}_{n \rightarrow \infty} (R(W^n \times V^n, \lambda) - C_1(n) - C_2(n)) \\ &\leq \inf_\lambda \left(\overline{\lim}_{n \rightarrow \infty} (R(W^n, \lambda) - C_1(n)) + \overline{\lim}_{n \rightarrow \infty} (R(V^n, \lambda) - C_2(n)) \right) \leq \lim_{l \rightarrow \infty} 2\delta_l = 0. \end{aligned}$$

□

6.4.4 Zero-error Capacity

Replacing \inf and $\underline{\lim}$ (resp. $\overline{\lim}$) we get a definition of $C_0 : \mathbb{N} \rightarrow \mathbb{R}_+$ by the conditions

$$\overline{\lim}_{n \rightarrow \infty} R(n, 0) - C_0(n) \geq 0$$

and

$$\overline{\lim}_{n \rightarrow \infty} R(n, 0) - C_0(n) \leq 0.$$

Since

$$M(n_1 + n_2, 0) \geq M(n_1, 0)M(n_2, 0) \quad (6.61)$$

for the DMC we get superadditivity for parallel channels. We do not have (6.61) for general channels, however, we have

$$M_{12}(n, 0) \geq M_1(n, 0)M_2(n, 0)$$

and therefore superadditivity for capacity sequences. Why has every channel with this structure a zero-error capacity sequence? Because we can choose $C_0(n) = R(n, 0)$.

6.5 Lecture on the Capacity Functions $C(\lambda)$, $C(\bar{\lambda})$

6.5.1 Explanations Via Compound Channels

With some abuse of notations we use as in [13] for capacity functions for maximal error probability λ and average error probability $\bar{\lambda}$ the same symbol C (for the different functions).

Recall that $N(N, \lambda)$ (resp. $N(n, \bar{\lambda})$) is the maximal size of n -block codes for a channel in question with maximal error probabilities $\leq \lambda$ (resp. average error probability $\leq \bar{\lambda}$).

Since $0 \leq \log N(n, \lambda) \leq n \log a$,

$$C^+(\lambda) = \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log N(n, \bar{\lambda}) \quad (6.62)$$

and

$$C^-(\lambda) = \underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log N(n, \bar{\lambda}) \quad (6.63)$$

are well-defined for all $\lambda \in (0, 1)$. If for a λ $C^*(\lambda) = C^-(\lambda)$, then $C(\lambda)$ exists. Analogous definitions are given for $\bar{\lambda}$.

It was shown in [13] that for CC with finitely many states $C(\bar{\lambda})$ is a non-decreasing step functions with finitely many jumps at specified arguments $\bar{\lambda}_1, \dots, \bar{\lambda}_{k(|S|)}$.

Remark 1 It is essential that we required $\leq \bar{\lambda}$ rather than $< \bar{\lambda}$, which would make the capacity functions continuous from the left (or lower semicontinuous) and thus it would be defined and known in the jump points.

The same is the case for instance for averaged channels with finitely many constituents.

In the definition zero-error codes play a role. The principal observation is easiest to explain for the case $|S| = 2$, where we have two DMC, $W(\cdot | \cdot | 1)$ and $W(\cdot | \cdot | 2)$, with corresponding capacities C_1 and C_2 with $C_1 \geq C_2$ w.l.o.g.

A further quantity is C_{12} , the compound capacity. Here

$$C(\bar{\lambda}) = \begin{cases} C_{12} & \text{for } \bar{\lambda} < \frac{1}{2} \\ C_{12} \vee (C_{01} \wedge C_{02}) & \text{for } \bar{\lambda} = \frac{1}{2} \\ C_2 = C_1 \wedge C_2 & \text{for } \bar{\lambda} > \frac{1}{2}, \end{cases}$$

where C_{0i} is the zero-error capacity of the i th channel.

Obviously the new definition with $< \bar{\lambda}$ would have the second line replaced by “ C_{12} for $\bar{\lambda} = \frac{1}{2}$ ”.

$$(C_{12} \vee C_{13} \vee C_{23}) \wedge C_1 \wedge C_2 \wedge C_3 = (C_{12} \wedge C_3) \vee (C_{13} \wedge C_2) \vee (C_{23} \wedge C_1)$$

Suppose $C_{12} \wedge C_3 > C_{13} \wedge C_2$, $C_{23} \wedge C_1$ then $C(\bar{\lambda}) = C_{12} \wedge C_3$.

Permuting the indices we would get $C(\bar{\lambda}) \neq C_{12} \wedge C_3$.

Given $\mathcal{W} = \{W(\cdot | \cdot | s) : 1 \leq s \leq k\}$ define $C_{lr} \cdots = \max_P \inf_{s=l,r} I(P, W(\cdot | \cdot | s))$.

From these basic formulas we can build $C(\bar{\lambda})$.

Theorem 44 *Except perhaps for finitely many points $\lambda_1, \dots, \lambda_{K^*(k)}$ for every $\bar{\lambda} \in (0, 1)$*

$$C(\bar{\lambda}) = C_S = \max_P \inf_{s \in S} I(P, W(\cdot | \cdot | s))$$

for some $S \subset \mathcal{S} = \{1, 2, \dots, k\}$.

Example $k = 3$

$$C(\bar{\lambda}) = \begin{cases} C_{123} & \text{for } 0 < \bar{\lambda} < \frac{1}{3} \\ C_{12} \wedge C_{13} \wedge C_{23} & \text{for } \frac{1}{3} < \bar{\lambda} < \frac{1}{2} \\ (C_{12} \vee C_{13} \vee C_{23}) \wedge C_1 \wedge C_2 \wedge C_3 & \text{for } \frac{1}{2} < \bar{\lambda} < \frac{2}{3} \\ C_1 \wedge C_2 \wedge C_3 & \text{for } \frac{2}{3} < \bar{\lambda} < 1 \end{cases}$$

Remark 2 This example shows that $C(\bar{\lambda})$ cannot be expressed by $C(\lambda)$.

6.5.2 Explanations Via Abstract Channels with Time Structure

For a channel with time structure $\mathcal{K} = (W^n)_{n=1}^\infty$ the function $R : \mathbb{N} \times (0, 1) \rightarrow \mathbb{R}_+$ defined by

$$R(n, \lambda) = \frac{1}{n} \log M(n, \lambda) \quad (6.64)$$

is the rate function-sequence, $C^+(\bar{\lambda})$ is the optimistic, $C^-(\bar{\lambda})$ is the pessimistic, and if $C^-(\bar{\lambda}) = C^+(\bar{\lambda})$ then $C(\bar{\lambda}) = C^-(\bar{\lambda}) = C^+(\bar{\lambda})$ is the error capacity function. For fixed $\bar{\lambda}$ $C(\bar{\lambda})$ is the $\bar{\lambda}$ -capacity.

Remarks

3. For the AB-channel we have constant optimistic and pessimistic capacity functions. $\overline{C}(\lambda) = 1$, $\underline{C}(\lambda) = 0$ for all $\lambda \in (0, 1)$.
4. Notabene: In the terminology of *capacity sequences* there is no problem in defining “error capacity sequences”. The rate function-sequence R takes care of it by (6.64).

6.5.3 Further Clarifications for Error Capacity Functions

We have introduced and analyzed time-capacity sequences in Lecture 6.4 and error-capacity functions $C(\bar{\lambda})$ above. They were first introduced by Ahlswede/Wolfowitz [13]. It must be emphasized that here $C(\bar{\lambda})$ exists, if $\overline{C}(\bar{\lambda}) = \underline{C}(\bar{\lambda})$. Much earlier in 1963 Parthasarathy introduced for maximal error probability ε , $\varepsilon \in (0, 1)$, what he called ε -capacity

$$C(\varepsilon) = \lim_{n \rightarrow \infty} n^{-1} \log N(n, \varepsilon)$$

(see also Kieffer [43]: the “optimum asymptotic rate”), we write this quantity as above in the form $\underline{C}(\varepsilon)$. It is monotone non-decreasing, and therefore it has at most countably many discontinuities, which are all jump discontinuities.

The right hand and the left hand limits of the pessimistic ε -capacity functions are

$$\begin{aligned} \underline{C}(\varepsilon^+) &= \lim_{\varepsilon' \downarrow \varepsilon} \underline{C}(\varepsilon'), \quad 0 \leq \varepsilon < 1. \\ \underline{C}(\varepsilon^-) &= \lim_{\varepsilon' \uparrow \varepsilon} \underline{C}(\varepsilon'), \quad 0 < \varepsilon < 1. \end{aligned}$$

Clearly $\underline{C}(\varepsilon^-) \leq \underline{C}(\varepsilon) \leq \underline{C}(\varepsilon^+)$, $0 < \varepsilon < 1$. Note that $\underline{C}(0^+)$ is the *pessimistic capacity*.

Kieffer determines $\underline{C}(\varepsilon)$ for any $\varepsilon \in (0, 1)$ for which the gap $\underline{C}(\varepsilon^+) - \underline{C}(\varepsilon^-)$ is sufficiently small for a BSAC. The problem of computing $\underline{C}(\varepsilon)$ for general DC is still open and is generally much harder than computing $\underline{C}(\varepsilon^+)$.

There has been work on “information quantities” by Kieffer [40], Gray/Ornstein [25], Kieffer [41], and Verdú/Han [48]. This is discussed in [43], where the belief is expressed that the techniques used extend also to some other channels.

We are convinced that it is easier to follow the definition under Remark 1 and cannot see any *practical reason* for a very, very small loosening of the error constraint (from $< \lambda$ to $\leq \lambda$).

Unfortunately, we have to point out some errors in terminology of some of our experts.

Remarks

1. Verdú/Han defined in [48], page 1149, the pessimistic $\underline{C}(\varepsilon)$ as “ ε -capacity” and the (pessimistic) capacity $C = \lim_{\varepsilon \downarrow 0} \underline{C}(\varepsilon)$. Kieffer erroneously wrote on page 1268 of [42] that Verdú/Han called $\limsup_{n \rightarrow \infty} n^{-1} \log N(n, \varepsilon)$ the ε -capacity! This is our $\bar{C}(\varepsilon)$.
2. On page 1153 of [48] the authors write that Wolfowitz referred to the conventional capacity (the pessimistic in the previous remark) as weak capacity. Actually, this term was introduced for a constant for which coding theorem and weak converse hold, however, more importantly it was claimed that the weak capacity equals the (pessimistic) capacity above. Of course the latter exists by definition, but the former need not exist. If it does, then the quantities are equal and in fact they also equal the (optimistic) \bar{C} .

6.5.4 Analogous Concepts for Sources

In addition to input alphabet \mathcal{X} and output alphabet \mathcal{Y} there is now a source alphabet \mathcal{Z} . To channels with time structure correspond sources with time structure $Q = (P^n)_{n=1}^{\infty}$ where P^n is a PD on Z^n .

Definition 35 For a source with time structure $Q = (P^n)_{n=1}^{\infty}$ the function $T : \mathbb{N} \rightarrow \mathbb{R}_+$ is a compression sequence if for minimal code size $M(n, \varepsilon)$, where n is the block length or time and ε is the permitted error probability, and the corresponding rate $R(n, \varepsilon) = \frac{1}{n} \log M(n, \varepsilon)$

$$\sup_{\varepsilon > 0} \overline{\lim}_{n \rightarrow \infty} (R(n, \varepsilon) - T(n)) \leq 0 \quad (6.65)$$

and

$$\sup_{\varepsilon > 0} \underline{\lim}_{n \rightarrow \infty} (R(n, \varepsilon) - T(n)) \geq 0. \quad (6.66)$$

We call

$$\bar{T} = \sup_{\varepsilon > 0} \overline{\lim}_{n \rightarrow \infty} R(n, \varepsilon) \quad (6.67)$$

the pessimistic compression and so

$$\underline{T} = \sup_{\varepsilon > 0} \lim_{n \rightarrow \infty} R(n, \varepsilon) \quad (6.68)$$

the optimistic compression.

6.5.5 The Analogous Result and Its Proof

Theorem 45 *Every source with time structure has a compression function if $\underline{T} < \infty$. Moreover, if (T, T, T, \dots) is a compression sequence, then $T = \bar{T} = \underline{T}$.*

Proof We use only that $R(n, \varepsilon)$ is not increasing in ε . □

Let $(\delta_l)_{l=1}^\infty$ be a null-sequence of positive numbers and let $(\varepsilon_l)_{l=1}^\infty$ be such that $\varepsilon_l \in (0, 1)$ and

$$\bar{T} + \delta_l \geq \overline{\lim}_{n \rightarrow \infty} R(n, \varepsilon_n) \geq \bar{T} \quad (6.69)$$

$$\underline{T} + \delta_l \geq \underline{\lim}_{n \rightarrow \infty} R(n, \varepsilon_n) \geq \underline{T} \quad (6.70)$$

Moreover, let $(n_l)_{l=1}^\infty$ be a monotone increasing sequence of natural numbers such that for all $n \geq n_l$

$$\bar{T} + \delta_l \geq R(n, \varepsilon_l) \geq \underline{T} - \delta_l. \quad (6.71)$$

Replacing $\underline{C}, \bar{C}, \lambda_l$ by $\underline{T}, \bar{T}, \varepsilon_l$ in the definition we get again sets $A_l(i)$ for $1 \leq i \leq d_l$ and a sequence $(T(n))_{n=1}^\infty$ instead of $(C(n))_{n=1}^\infty$.

We get now for any $\varepsilon \in (0, 1)$ and $\varepsilon_l < \varepsilon$

$$R(n, \varepsilon) - T(n) \leq K(n, \varepsilon_{l+j}) - T(n) \leq s \quad (6.72)$$

for $n_{l+j} \leq n < n_{l+j+1}$ and $j = 0, 1, 2, \dots$ and thus

$$\overline{\lim}_{n \rightarrow \infty} R(n, \varepsilon) - T(n) \leq 0 \quad (6.73)$$

and thus (6.65).

Finally, for any $\varepsilon < \varepsilon_l$ by $R(n, \varepsilon)$ monotonically decreasing in ε

$$\underline{\lim}_{n \rightarrow \infty} R(n, \varepsilon) - T(n) \geq \underline{\lim}_{n \rightarrow \infty} R(n, \varepsilon_l) - T(n) \geq 2\delta_l \quad (6.74)$$

and

$$\sup_{\varepsilon \in (0,1)} \lim_{n \rightarrow \infty} R(n, \varepsilon) - T(n) \geq \lim_{l \rightarrow \infty} 2\delta_l = 0 \quad (6.75)$$

and (6.63) holds.

Example No compression function.

6.5.6 Parallel Sources or the Product of Sources

Theorem 46 For two sources with time structure $Q_1 = (P_1^n)_{n=1}^\infty$ and $Q_2 = (P_2^n)_{n=1}^\infty$, which have compression sequences T_1 and T_2 , the product source $Q_1 \times Q_2$ has compression sequence $T_{12} = T_1 + T_2$.

There may be a proof like the one for channels. However, we suggest another approach. Obviously for $\varepsilon = \varepsilon_1, \dots, \varepsilon_n$

$$N(Q_1, \varepsilon_1)N(Q_2, \varepsilon_2) \geq N(Q_1 \times Q_2, \varepsilon_1, \dots)$$

Establishing an appropriate lower bound leads to a nice combinatorial (or analytic) problem. For PD's $P = (p_1, p_2, \dots, p_a)$ and $Q = (q_1, \dots, q_b)$ and any integer $s \in [a, b]$ consider

$$F(a, b, s) = \max \left\{ \sum_{(i,j) \in S} p_i q_j \text{ for } S \in [a] \times [b] \text{ with } |S| = s \right\}.$$

Conjecture

$$F(a, b, s) \geq \frac{1}{2} \max \left\{ \sum_{i \in A} p_i \sum_{j \in B} q_j : A \subset [a], B \subset [b], |A||B| \leq s \right\}.$$

It may relate to Chebyshev's inequality. Put for $c \leq \min(a, b)$

$$r_k = \max\{p_i q_k, \alpha_k q_i : 1 \leq i \leq k\}, \quad 1 \leq k \leq c,$$

then

$$\sum_{i=1}^c p_i \sum_{j=1}^c q_j \leq c \sum_{k=1}^c r_k.$$

In particular, if $p_1 \leq p_2 \leq \dots \leq p_c$ and

$$\sum_{i=1}^c c p_i \sum_{j=1}^c q_j \leq c \sum_{i=1}^c p_i q_i.$$

Further Reading

1. R. Ahlswede, On two-way communication channels and a problem by Zarankiewicz. in Sixth Prague Conference on Information Theory, Statistical Decision Function's and Random Process (Publishing House of the Czech Academy of Sciences, 1973), pp. 23–37
2. R. Ahlswede, Channel capacities for list codes. *J. Appl. Probab.* **10**, 824–836 (1973)
3. R. Ahlswede, Elimination of correlation in random codes for arbitrarily varying channels. *Z. Wahrscheinlichkeitstheorie und verw. Geb.* **44**, 159–175 (1978)
4. R. Ahlswede, Coloring hypergraphs: a new approach to multi-user source coding I. *J. Comb. Inf. Syst. Sci.* **4**(1), 76–115 (1979)
5. R. Ahlswede, Coloring hypergraphs: a new approach to multi-user source coding II. *J. Comb. Inf. Syst. Sci.* **5**(3), 220–268 (1980)
6. R. Ahlswede, A method of coding and its application to arbitrarily varying channels. *J. Comb. Inf. Syst. Sci.* **5**(1), 10–35 (1980)
7. R. Ahlswede, An elementary proof of the strong converse theorem for the multiple-access channel. *J. Comb. Inf. Syst. Sci.* **7**(3), 216–230 (1982)
8. R. Ahlswede, V. Balakirsky, Identification under random processes, problemy peredachii informatsii (special issue devoted to M.S. Pinsker). *Probl. Inf. Transm.* **32**(1), 123–138
9. R. Ahlswede, V. Balakirsky, Identification under random processes, problemy peredachii informatsii (special issue devoted to M.S. Pinsker). *Probl. Inf. Transm.* **32**(1), 144–160 (1996)
10. R. Ahlswede, I. Csiszár, Common randomness in information theory and cryptography, part I: secret sharing. *IEEE Trans. Inform. Theor.* **39**(4), 1121–1132 (1993)
11. R. Ahlswede, 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. Inform. Theor.* **44**(1), 55–62 (1998)
12. R. Ahlswede, 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)
13. R. Ahlswede, G. Dueck, Good codes can be produced by a few permutations. *IEEE Trans. Inform. Theor.* **IT-28**(3), 430–443 (1982)
14. R. Ahlswede, G. Dueck, Identification in the presence of feedback—a discovery of new capacity formulas. *IEEE Trans. Inform. Theor.* **35**(1), 30–39 (1989)
15. R. Ahlswede, P. Gács, J. Körner, Bounds on conditional probabilities with applications in multiuser communication. *Z. Wahrscheinlichkeitstheorie und verw. Geb.* **34**, 157–177 (1976)
16. R. Ahlswede, T.S. Han, On source coding with side information via a multiple-access channel and related problems in multi-user information theory. *IEEE Trans. Inform. Theor.* **IT-29**(3), 396–412 (1983)
17. R. Ahlswede, B. Verboven, On identification via multi-way channels with feedback. *IEEE Trans. Inform. Theor.* **37**(5), 1519–1526 (1991)
18. R. Ahlswede, 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. Inform. Theor.* **41**(4), 1040–1050 (1995)

19. S. Arimoto, On the converse to the coding theorem for the discrete memoryless channels. *IEEE Trans. Inform. Theor.* IT-19, 357–359 (1973)
20. R. Ash, *Information Theory, Interscience Tracts in Pure and Applied Mathematics*, vol. 19 (Wiley, New York, 1965)
21. G. Aubrun, S. Szarek, E. Werner, Hastings' additivity counterexample via Dvoretzky's theorem. *Commun. Math. Phys.* **305**, 85–97 (2011)
22. R.E. Blahut, Hypothesis testing and information theory, *IEEE Trans. Inform. Theor.* IT-20, 405–417 (1974)
23. R.E. Blahut, Composition bounds for channel block codes. *IEEE Trans. Inform. Theor.* IT-23, 656–674 (1977)
24. I. Csiszár, J. Körner, Graph decomposition: a new key to coding theorems. *IEEE Trans. Inform. Theor.* IT-27, 5–12 (1981)
25. I. Csiszár, J. Körner, K. Marton, *A new look at the error exponent of a discrete memoryless channel (preprint)* (IEEE Intern. Symp. Inform. Theory, Ithaca, NY, 1977)
26. R.L. Dobrushin, S.Z. Stambler, Coding theorems for classes of arbitrarily varying discrete memoryless channels. *Prob. Peredachi Inform.* **11**, 3–22 (1975)
27. G. Dueck, *Omnisophie: über richtige, wahre und natürliche Menschen* (Springer, Berlin Heidelberg, 2003)
28. G. Dueck, J. Körner, Reliability function of a discrete memoryless channel at rates above capacity. *IEEE Trans. Inform. Theor.* IT-25, 82–85 (1979)
29. H. Dudley, The vocoder. *Bell. Lab. Rec.* **18**, 122–126 (1939)
30. A. Feinstein, A new basic theorem of information theory. *IRE Trans. Inform. Theor.* **4**, 2–22 (1954)
31. R.G. Gallager, A simple derivation of the coding theorem and some applications. *IEEE Trans. Inform. Theor.* IT-11, 3–18 (1965)
32. R.G. Gallager, Source coding with side information and universal coding (preprint). in *IEEE International Symposium Information Theory* (Ronneby, Sweden, 1976)
33. V.D. Goppa, Nonprobabilistic mutual information without memory. *Prob. Contr. Inform. Theor.* **4**, 97–102 (1975)
34. A. Haroutunian, Estimates of the error exponent for the semi-continuous memoryless channel. *Prob. Peredachi Inform.* **4**, 37–48 (1968)
35. H. Kesten, Some remarks on the capacity of compound channels in the semicontinuous case. *Inform. Contr.* **4**, 169–184 (1961)
36. V.N. Koselev, On a problem of separate coding of two dependent sources. *Prob. Peredachi Inform.* **13**, 26–32 (1977)
37. J.K. Omura, A lower bounding method for channel and source coding probabilities. *Inform. Contr.* **27**, 148–177 (1975)
38. C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(379–423), 632–656 (1948)
39. C.E. Shannon, R.G. Gallager, E.R. Berlekamp, Lower bounds to error probability for coding on discrete memoryless channels I-II. *Inform. Contr.* **10**(65–103), 522–552 (1967)
40. D. Slepian, J.K. Wolf, Noiseless coding of correlated information sources, *IEEE Trans. Inform. Theor.* IT-19, 471–480 (1973)
41. J. Wolfowitz, The coding of messages subject to chance errors. III. *J. Math.* **1**, 591–606 (1957)

References

1. R. Ahlswede, Certain results in coding theory for compound channels. in Proceedings of the Colloquium Information Theory (Debrecen, Hungary, 1967), pp. 35–60
2. R. Ahlswede, Beiträge zur Shannonschen Informationstheorie im Fall nichtstationärer Kanäle, Z. Wahrscheinlichkeitstheorie und verw. Geb. 10, 1–42 (1968) (Diploma 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, Multi-way communication channels, in Proceedings of 2nd International Symposium on Information Theory, Thakadsor, Armenian SSR, September 1971 (Akademiai Kiado, Budapest, 1973), pp. 23–52
5. R. Ahlswede, The capacity region of a channel with two senders and two receivers. Ann. Probab. **2**(5), 805–814 (1974)
6. R. Ahlswede, in textitOn Concepts of Performance Parameters for Channels, General Theory of Information Transfer and Combinatorics. Lecture Notes in Computer Science, vol. 4123 (Springer, 2006), pp. 639–663
7. 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 **156**(9), 1348–1388 (2008)
8. R. Ahlswede, M.V. Burnashev, On minimax estimation in the presence of side information about remote data. Ann. Statist. **18**(1), 141–171 (1990)
9. R. Ahlswede, N. Cai, 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. Inform. Theor. **42**(1), 55–62 (1996)
10. R. Ahlswede, N. Cai, Z. Zhang, in Secrecy Systems for Identification Via Channels with Additive-Like Instantaneous Block Encipherers, General Theory of Information Transfer and Combinatorics. Lecture Notes in Computer Science, vol. 4123 (Springer, 2006), pp. 285–292
11. R. Ahlswede, I. Csiszár, Hypothesis testing under communication constraints. IEEE Trans. Inform. Theor. **IT-32**(4), 533–543 (1986)
12. R. Ahlswede, G. Dueck, Identification via channels. IEEE Trans. Inform. Theor. **35**(1), 15–29 (1989)
13. R. Ahlswede, J. Wolfowitz, The structure of capacity functions for compound channels. in Proceedings of the International Symposium on Probability and Information Theory, April 1968 (McMaster University, Canada, 1969), pp. 12–54
14. U. Augustin, Gedächtnisfreie Kanäle für diskrete Zeit. Z. Wahrscheinlichkeitstheorie u. verw. Geb. **6**, 10–61 (1966)
15. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression* (Prentice-Hall, Englewood Cliffs, 1971)
16. T.M. Cover, J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991)
17. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic, New York, 1981)
18. R.L. Dobrushin, General formulation of Shannon’s main theorem of information theory. Usp. Math. Nauk. **14**, 3–104 (1959)
19. R.L. Dobrushin, General formulation of Shannon’s main theorem of information theory. Am. Math. Soc. Trans. **33**, 323–438 (1962)
20. S.M. Dodunekov, Optimization problems in coding theory. in Workshop on Combinatorial Search, Budapest, 23–26 April 2005
21. J.L. Doob, Review of “A mathematical theory of communication”. Math. Rev. **10**, 133 (1949)
22. R.M. Fano, *Transmission of Information, A Statistical Theory of Communication* (Wiley, New York, 1961)
23. A. Feinstein, *Foundations of Information Theory* (McGraw-Hill, New York, 1958)

24. R.G. Gallager, *Information Theory and Reliable Communication* (Wiley, New York, 1968)
25. R.M. Gray, D.S. Ornstein, Block coding for discrete stationary \bar{d} -continuous noisy channels. *IEEE Trans. Inf. Theor.* **IT-25**(3), 292–306 (1979)
26. S. Györi, Coding for a multiple access OR channel: a survey. *General Theory of Information Transfer and Combinatorics, Special Issue of, Discrete Applied Mathematics*, 156(9), 2008
27. W. Haemers, On some problems of Lovasz concerning the Shannon capacity of a graph. *IEEE Trans. Inform. Theor.* **25**(2), 231–232 (1979)
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, S.I. Amari, Statistical inference under multiterminal data compression, information theory: 1948–1998. *IEEE Trans. Inform. Theor.* **44**(6), 2300–2324 (1998)
31. T.S. Han, S. Verdú, Approximation theory of output statistics. *IEEE Trans. Inf. Theor.* **IT-39**(3), 752–772 (1993)
32. T.S. Han, S. Verdú, New results in the theory of identification via channels. *IEEE Trans. Inform. Theor.* **39**(3), 752–772 (1993)
33. E.A. Haroutunian, Upper estimate of transmission rate for memoryless channel with countable number of output signals under given error probability exponent, (in Russian). in 3rd All-Union Conference on Theory of Information Transmission and Coding (Publication house of Uzbek Academy of Sciences, Uzhgorod, Tashkent, 1967), pp. 83–86
34. M.B. Hastings, Superadditivity of communication capacity using entangled inputs. *Letters* (2009)
35. A.S. Holevo, The capacity of quantum channel with general signal states. *IEEE Trans. Inf. Theor.* **44**, 269–273 (1998)
36. M. Horodecki, Is the Classical Broadcast Channel Additive? Oral Communication (Cambridge England, 2004)
37. K. Jacobs, Almost periodic channels. in Colloquium on Combinatorial Methods in Probability Theory, Matematisk Institute, Aarhus University, 1–10 Aug 1962, pp. 118–126
38. F. Jelinek, *Probabilistic Information Theory* (McGraw-Hill, New York, 1968)
39. W. Kautz, R. Singleton, Nonrandom binary superimposed codes. *IEEE Trans. Inform. Theor.* **10**, 363–377 (1964)
40. J.C. Kieffer, A general formula for the capacity of stationary nonanticipatory channels. *Inf. Contr.* **26**, 381–391 (1974)
41. J.C. Kieffer, Block coding for weakly continuous channels. *IEEE Trans. Inf. Theor.* **IT-27**(6), 721–727 (1981)
42. J.C. Kieffer, Epsilon-capacity of a class of nonergodic channels. in *Proceedings of IEEE International Symposium Information Theory* (Seattle, WA, 2006), pp. 1268–1271.
43. J.C. Kieffer, ε -capacity of binary symmetric averaged channels. *IEEE Trans. Inf. Theor.* **53**(1), 288–303 (2007)
44. M.S. Pinsker, *Information and Stability of Random Variables and Processes* (Izd-vo Akademii Nauk, Moscow, 1960)
45. B. Schumacher, M.D. Westmoreland, Sending classical information via noisy quantum channels. *Phys. Rev. A* **56**(1), 131–138 (1997)
46. C.E. Shannon, The zero error capacity of a noisy channel. *IRE Trans. Inform. Theor.* **2**, 8–19 (1956)
47. C.E. Shannon, Certain results in coding theory for noisy channels. *Inform. Contr.* **1**, 6–25 (1957)
48. S. Verdú, T.S. Han, A general formula for channel capacity. *IEEE Trans. Inform. Theor.* **40**(4), 1147–1157 (1994)
49. J. Wolfowitz, Coding theorems of information theory. in *Ergebnisse der Mathematik und ihrer Grenzgebiete*, vol. 31, 3rd edn., (*Springer* (Englewood Cliffs, Berlin-Göttingen-Heidelberg, Prentice-Hall, 1978), p. 1961
50. A.D. Wyner, The capacity of the product channel. *Inform. Contr.* **9**, 423–430 (1966)

51. R.W. Yeung, *A First Course in Information Theory, Information Technology: Transmission, Processing and Storage* (Kluwer Academic/Plenum Publishers, New York, 2002)
52. J. Ziv, Back from Infinity: a constrained resources approach to information theory. *IEEE Inform. Theor. Soc. Newslett.* **48**(1), 30–33 (1998)

Storing and Transmitting Data

Rudolf Ahlswede's Lectures on Information Theory 1

Ahlswede, R. - Ahlswede, A.; Althöfer, I.; Deppe, C.;

Tamm, U. (Eds.)

2014, X, 302 p. 6 illus., Hardcover

ISBN: 978-3-319-05478-0